Journal of Machine Learning Research 6 (2005) 1099–1125 Submitted 10/04; Revised 2/05; Published 7/05

Learning the Kernel Function via Regularization

Charles A. Micchelli CAM@MATH.ALBANY.EDU
Department of Mathematics and Statistics
State University of New York
The University at Albany
1400 Washington Avenue
Albany, NY, 12222, USA

Massimiliano Pontil M.PONTIL@CS.UCL.AC.UK
Department of Computer Science
University College London
Gower Street, London WC1E, UK

Editor: Peter Bartlett

Abstract

We study the problem of finding an optimal kernel from a prescribed convex set of kernels $\mathcal{K}$ for learning a real-valued function by regularization. We establish for a wide variety of regularization functionals that this leads to a convex optimization problem and, for square loss regularization, we characterize the solution of this problem. We show that, although $\mathcal{K}$ may be an uncountable set, the optimal kernel is always obtained as a convex combination of at most $m+2$ basic kernels, where m is the number of data examples. In particular, our results apply to learning the optimal radial kernel or the optimal dot product kernel.

1. Introduction

A widely used approach to estimate a function from empirical data consists in minimizing a regularization functional in a Hilbert space $\mathcal{H}$ of real–valued functions $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a set. Specifically, regularization estimates f as a *minimizer* of the functional

$$Q(I_{\mathbf{x}}(f)) + \mu\Omega(f)$$

where μ is a positive parameter, $I_{\mathbf{x}}(f) = (f(x_j) : j \in \mathbb{N}_m)$ is the *vector* of values of f on the *set* $\mathbf{x} = \{x_j : j \in \mathbb{N}_m\}$ and $\mathbb{N}_m = \{1, \ldots, m\}$. This functional trades off *empirical error*, measured by the function $Q : \mathbb{R}^m \to \mathbb{R}_+$, with *smoothness* of the solution, measured by the functional $\Omega : \mathcal{H} \to \mathbb{R}_+$. The empirical error depends upon a finite set $\{(x_j, y_j) : j \in \mathbb{N}_m\} \subset \mathcal{X} \times \mathbb{R}$ of *input-output examples* and the function Q depends on y but we suppress it in our notation since it is fixed throughout our discussion. In particular, one often considers the case that Q is defined, for $v = (v_j : j \in \mathbb{N}_m) \in \mathbb{R}^m$, as $Q(v) = \sum_{j \in \mathbb{N}_m} V(v_j, y_j)$ where $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a prescribed *loss function*.

In this paper we assume that $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_K$ with kernel K and choose $\Omega(f) = \langle f, f \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}_K$, although some of the ideas we develop may be relevant in other circumstances. This leads us to study the variational problem

$$Q_\mu(K) := \inf \left\{ Q(I_{\mathbf{x}}(f)) + \mu \langle f, f \rangle : f \in \mathcal{H}_K \right\}. \tag{1}$$

We recall that an RKHS is a Hilbert space of real-valued functions everywhere defined on $\mathcal{X}$ such that, for every $x \in \mathcal{X}$, the point evaluation functional defined, for $f \in \mathcal{H}$, by $L_x(f) := f(x)$ is continuous on $\mathcal{H}$ (Aronszajn, 1950). This implies that $\mathcal{H}$ admits a reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that, for every $x \in \mathcal{X}$, $K(x,\cdot) \in \mathcal{H}$ and $f(x) = \langle f, K(x,\cdot)\rangle$. In particular, for $x,t \in \mathcal{X}$, $K(x,t) = \langle K(x,\cdot), K(t,\cdot)\rangle$ implying that the $m \times m$ matrix $K_{\mathbf{x}} := (K(x_i,x_j) : i,j \in \mathbb{N}_m)$ is symmetric and positive semi-definite for *any* set of inputs $\mathbf{x} \subseteq \mathcal{X}$.

Often RKHS's are introduced through the notion of *feature map* $\Phi : \mathcal{X} \to \mathcal{W}$, where $\mathcal{W}$ is a Hilbert space with inner product denoted by $(\cdot,\cdot)$. A feature map gives rise to the linear space of all functions $f : \mathcal{X} \to \mathbb{R}$ which are a linear combination of features whose norm is taken to be the norm of its coefficients. That is, for $w \in \mathcal{W}$, $f = (w, \Phi)$ and $\langle f, f\rangle = (w,w)$. This space is an RKHS with kernel K defined, for $x,t \in \mathcal{X}$, as $K(x,t) = (\Phi(x), \Phi(t))$. Using these equations, the regularization functional in (1) can be rewritten as a functional of w.

Regularization in an RKHS has a number of attractive features, including the availability of effective error bounds and stability analysis relative to perturbations of the data (see, for example, Bousquet and Elisseeff, 2002; Cucker and Smale, 2002; Mukherjee et al., in press; Scovel and Steinwart, 2004; Smale and Zhou, 2003; Vapnik, 1998; Ying and Zhou, 2004; Zhang, 2004; Zhou, 2002). Moreover, one can show that if f is a minimizer of the above functional it has the form

$$f(x) = \sum_{j \in \mathbb{N}_m} c_j K(x_j, x), \quad x \in \mathcal{X} \tag{2}$$

for some real vector $c = (c_j : j \in \mathbb{N}_m)$ of coefficients (see, for example, De Vito et al., 2005; Girosi, 1998; Kimeldorf and Wahba, 1971; Micchelli and Pontil, 2005; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). This result is known in Machine Learning as the *representer theorem*. Although it is simple to prove, this result is remarkable as it makes the variational problem (1) amenable for computations.

If Q is convex, the unique minimizer of problem (1) can be found by replacing f by the right hand side of equation (2) in equation (1) and then optimizing with respect to the vector c. For example, when Q is the square error defined for $v = (v_j : j \in \mathbb{N}_m) \in \mathbb{R}^m$ as $Q(v) = \sum_{j \in \mathbb{N}_m} (v_j - y_j)^2$ the functional in the right hand side of (1) is a quadratic in the vector c and its minimizer is obtained by solving a linear system of equations.

Because of their simplicity and generality, kernels and associated RKHS's play an increasingly important role in Machine Learning, Pattern Recognition and their applications. This was initiated with the introduction of support vector machines (see, for example, Vapnik, 1998), and continued with the development of many other kernel-based learning algorithms (see, for example, Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004, and references therein). As kernels can be defined on any input space, kernel-based methods have been successfully applied to learning functions defined on complex data structures, ranging from images, text data, speech data, biological data, among others.

Despite this great success, there still remain important problems to be addressed concerning kernel methods in Machine Learning. When the kernel is fixed, an immediate concern with problem (1) is the *choice of the regularization parameter* μ. This is typically solved by means of cross validation or generalized cross validation (see, for example, Hastie, Tibshirani and Friedman, 2002; Wahba, 1990) or by means of regularization path methods (see, for example, Bach, Thibaux and Jordan, 2004; Hastie et al., 2004; Pontil and Verri, 1998). But, how is the kernel chosen? Indeed, a challenging and central problem is the *choice of the kernel* itself. As we said before, when $\mathcal{H}$

is constructed as linear combinations of features associated to the kernel K, they can provide some guideline for the choice of the kernel. Thus, the choice of the kernel is tied to the problem of choosing a representation of the input. This choice can make a significant difference in practice. For example, techniques such as radial basis functions can perform poorly if the parameter of the radial kernel is not tuned to the given data. A similar circumstance occurs for translation invariant kernels modeled by Gaussian mixtures. When the number of parameters is large cross validation encounters severe computational limitations. To overcome this problem, easily computable approximations to the leave-one-out error have been derived (Chapelle et al., 2002; Wahba, 1990). Nonetheless, these methods are usually non-convex and may lead to undesirable local minima.

In this paper, we propose a method for finding a kernel function which belongs to a *compact* and *convex* set $\mathcal{K}$. Our method is based on the minimization of the functional in equation (1), that is, we consider the variational problem

$$\inf\{Q_\mu(K) : K \in \mathcal{K}\}. \tag{3}$$

This problem shares some similarities with recent progress in the context of kernel–based methods (Bach, Lanckriet and Jordan, 2004; Bousquet and Herrmann, 2003; Cristianini et al., 2002; Graepel, 2002; Lanckriet et al., 2002, 2004; Lee et al., 2004; Lin and Zhang, 2003; Herbster, 2001; Ong, Smola and Williamson, 2003; Wu, Ying and Zhou, 2004; Zhang, Yeung and Kwok, 2004). In particular, the third and fifth papers motivated our work. In contrast to the point of view of these papers, our setting applies to convex combinations of kernels parameterized by a compact set, a circumstance which is relevant for applications. We also wish to emphasize that although we focus on learning methods based on the minimization of the functional (1), the ideas which we present here may prove useful for learning kernels or feature representations using different forms of regularization such as entropy regularization (Jaakkola, Meila and Jebara, 1999), kernel density estimation (see, for example, Vapnik, 1998), or one-class SVM (Tax and Duin, 1999) as well as in other Machine Learning frameworks such as those arising in Bayesian learning where a kernel is seen as the covariance of a Gaussian process, (see, for example, Wahba, 1990; Williams and Rasmussen, 1996) or in online learning, (see, for example, Herbster, 2001).

In Section 2 we establish the existence of a solution to problem (3), show that the functional Q_μ is *convex* in K, and observe that, although $\mathcal{K}$ may be an uncountable set, the optimal kernel is always obtained as a convex combination of at most $m+2$ basic kernels (see below), where m is the number of training data. The simplest case of our setup is a set of convex combinations of finitely many kernels $\{K_j : j \in \mathbb{N}_n\}$. For example each K_j could be a Gaussian, a polynomial kernel, or simply a kernel consisting of only one feature. In all of these cases our method will seek the optimal convex combination of these kernels. Another example included in our framework is learning the optimal radial kernel or the optimal polynomial kernel in which case the space $\mathcal{K}$ is the convex hull of a prescribed set of kernels parameterized by a *locally compact* set. In Section 3 we study square loss regularization and provide improvements and simplifications of the results in Section 2. In particular, we discuss the connection to minimal norm interpolation and establish necessary and sufficient conditions for a kernel to be optimal. Finally, in Section 4 we comment on previous work, present some numerical simulations based on our analysis and discuss some extensions of our framework.

2. Optimal Convex Combination of Kernels

Let X be a set. By a *kernel* we mean a symmetric function $K : X \times X \to \mathbb{R}$ such that for every finite set of inputs $\mathbf{x} = \{x_j : j \in \mathbb{N}_m\} \subseteq X$ 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. 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. Also, we use $\mathcal{A}(X)$ for the set of all kernels on the set X and $\mathcal{A}_+(X)$ for the subset of kernels K such that, for each input $\mathbf{x}$, $K_{\mathbf{x}} \in \mathcal{L}_+(\mathbb{R}^m)$. We also occasionally refer to the set of *all* symmetric $m \times m$ matrices and use $\mathcal{S}(\mathbb{R}^m)$ to denote them.

According to Aronszajn and Moore (see Aronszajn, 1950), every kernel has associated to it an (essentially) *unique* Hilbert space $\mathcal{H}_K$ with inner product $\langle \cdot, \cdot \rangle$ such that K is its reproducing kernel. This means that for every $f \in \mathcal{H}_K$ and $x \in X$, $\langle f, K_x \rangle = f(x)$, where K_x is the function $K(x, \cdot)$.

Let $D := \{(x_j, y_j) : j \in \mathbb{N}_m\} \subset X \times \mathbb{R}$ be prescribed data and y the vector $(y_j : j \in \mathbb{N}_m)$. For each $f \in \mathcal{H}_K$, we introduce the *information operator* $I_{\mathbf{x}}(f) := (f(x_j) : j \in \mathbb{N}_m)$ of values of f on the set $\mathbf{x} := \{x_j : j \in \mathbb{N}_m\}$. We prescribe a nonnegative function $Q : \mathbb{R}^m \to \mathbb{R}_+$ and introduce the *regularization functional*

$$Q_\mu(f, K) := Q(I_{\mathbf{x}}(f)) + \mu \| f \|_K^2 \tag{4}$$

where $\| f \|_K^2 := \langle f, f \rangle$, μ is a positive constant and Q depends on y but we suppress it in our notation as it is fixed throughout our discussion. A noteworthy special case of Q_μ is the *square loss regularization* functional given by

$$S_\mu(f, K) := \| y - I_{\mathbf{x}}(f) \|^2 + \mu \| f \|_K^2 \tag{5}$$

where $\| \cdot \|$ is the standard Euclidean norm on $\mathbb{R}^m$. There are many other choices of the functional Q_μ which are important for applications, see the work of Vapnik (1998) for a discussion.

Associated with the functional Q_μ and the kernel K is the variational problem

$$Q_\mu(K) := \inf\{Q_\mu(f, K) : f \in \mathcal{H}_K\} \tag{6}$$

which defines a function $Q_\mu : \mathcal{A}(X) \to \mathbb{R}_+$. We remark, in passing, that all of what we say about problem (6) applies to functions Q which are bounded from below on $\mathbb{R}^m$ as we can merely adjust the expression (4) by a constant independent of f and K. Let us first point out that the infimum in (6) is achieved, at least when Q is continuous.

Lemma 1 *If $Q : \mathbb{R}^m \to \mathbb{R}_+$ is continuous and μ is a positive number then the infimum in (6) is achieved for a function in $\mathcal{H}_K$. Moreover, when Q is convex this function is unique.*

PROOF. The proof of this fact is straightforward and uses *weak compactness* of the unit ball in $\mathcal{H}_K$. The uniqueness of the solution relies on the fact that when Q is convex Q_μ is strictly convex because μ is positive. □

The point of view of this paper is that the functional (6) can be used as a *design criterion to select the kernel K*. To this end, we specify an arbitrary convex subset $\mathcal{K}$ of $\mathcal{A}(X)$ and focus on the problem

$$Q_\mu(\mathcal{K}) = \inf\{Q_\mu(K) : K \in \mathcal{K}\}. \tag{7}$$

Recall that the solution of (6) is given in the form $f = \sum_{j \in \mathbb{N}_m} c_j K_{x_j}$ for some vector $c := (c_j : j \in \mathbb{N}_m)$, (see, for example, De Vito et al., 2005; Girosi, 1998; Kimeldorf and Wahba, 1971; Micchelli and Pontil, 2005; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). Such a function

representation for learning the function f is central for many diverse applications of kernel-based algorithms in Machine Learning. Moreover, the coefficient vector c is found as the solution of the *finite dimensional* variational problem

$$Q_\mu(K) := \min\{Q(K_{\mathbf{x}}c) + \mu(c, K_{\mathbf{x}}c) : c \in \mathbb{R}^m\}$$

where $(\cdot,\cdot)$ is the standard inner product on $\mathbb{R}^m$.

Before we address basic questions concerning the variational problem (7) we describe some terminology that allows for a precise description of our observations. Every input set $\mathbf{x}$ and set of *basic kernels* $\mathcal{G}$ on $\mathcal{X}\times\mathcal{X}$ determines a set of *matrices* in $\mathcal{L}(\mathbb{R}^m)$, namely

$$\mathcal{G}(\mathbf{x}) := \{G_{\mathbf{x}} : G \in \mathcal{G}\}.$$

Obviously, it is the set of matrices $\mathcal{K}(\mathbf{x})$ that affects the variational problem (7). Note that $\mathcal{G}(\mathbf{x})$ being a subset of $\mathcal{S}(\mathbb{R}^m)$ is identifiable as a set of vectors in $\mathbb{R}^N$, where $N := \frac{m(m+1)}{2}$. As such $\mathcal{G}(\mathbf{x})$ inherits the standard topology from $\mathbb{R}^N$. That is, convergence of a sequence of matrices in $\mathcal{G}(\mathbf{x})$ means that the respective elements of the matrices converge. For this reason, we use $\overline{\mathcal{G}}$ (the closure of $\mathcal{G}$) to mean the set of all kernels K on $\mathcal{X}\times\mathcal{X}$ with the property that for each $\mathbf{x}\subseteq\mathcal{X}$, the matrix $K_{\mathbf{x}} \in \overline{\mathcal{G}(\mathbf{x})}$, the closure of $\mathcal{G}(\mathbf{x})$ relative to $\mathbb{R}^N$. We say a set of kernels $\mathcal{G}$ is closed provided that $\overline{\mathcal{G}} = \mathcal{G}$. Also, we say $\mathcal{G}$ is a compact convex set of kernels whenever for each $\mathbf{x}\subseteq\mathcal{X}$, $\mathcal{G}(\mathbf{x})$ is a compact convex set of matrices in $\mathcal{S}(\mathbb{R}^m)$. Our next result establishes the existence of the solution to problem (7).

Lemma 2 *If $\mathcal{K}$ is a compact and convex subset of $\mathcal{A}_+(\mathcal{X})$ and $Q:\mathbb{R}^m\to\mathbb{R}$ is continuous then the minimum of (7) exists.*

PROOF. Fix $\mathbf{x}\subseteq\mathcal{X}$, choose a minimizing sequence of kernels $\{K^n : n\in\mathbb{N}\}$, that is, $\lim_{n\to\infty} Q_\mu(K^n) = Q_\mu(\mathcal{K})$ and a sequence of vectors $\{c^n : n\in\mathbb{N}\}$ such that

$$Q_\mu(K^n) = Q(K^n_{\mathbf{x}}c^n) + \mu(c^n, K^n_{\mathbf{x}}c^n).$$

Since $\mathcal{K}$ is compact there is a subsequence $\{K^{n(\ell)} : \ell\in\mathbb{N}\}$ such that $\lim_{\ell\to\infty} K^{n(\ell)}_{\mathbf{x}} = \tilde{K}_{\mathbf{x}}$, for some kernel $\tilde{K}\in\mathcal{K}$. We claim that $\{c^n : n\in\mathbb{N}\}$ is bounded. Indeed, there is a positive constant ρ such that $(c^n, K^n_{\mathbf{x}}c^n) \le \rho$. Set $a^n = \frac{c^n}{\|c^n\|}$ so that $(a^n, K^n_{\mathbf{x}}a^n) \le \frac{\rho}{\|c^n\|^2}$ and choose a convergent subsequence $\{a^{n(\ell(q))} : q\in\mathbb{N}\}$ such that $\lim_{q\to\infty} a^{n(\ell(q))} = a$ and $\|a\| = 1$ for some vector $a\in\mathbb{R}^m$. If the sequence $\{c^n : n\in\mathbb{N}\}$ is not bounded we conclude that $(a, \tilde{K}_{\mathbf{x}}a) = 0$ contradicting our hypothesis that $\tilde{K}\in\mathcal{A}_+(\mathcal{X})$. Hence there is a subsequence $\{c^{n(\ell(q))} : q\in\mathbb{N}\}$ such that $\lim_{q\to\infty} c^{n(\ell(q))} = c$, for some $c\in\mathbb{R}^m$. Therefore, we conclude that

$$Q_\mu(\mathcal{K}) = Q(\tilde{K}_{\mathbf{x}}c) + \mu(c, \tilde{K}_{\mathbf{x}}c) \ge Q_\mu(\tilde{K})$$

from which it follows that $Q_\mu(\mathcal{K}) = Q_\mu(\tilde{K})$. □

The proof of this lemma requires that all kernels in $\mathcal{K}$ are in $\mathcal{A}_+(\mathcal{X})$. If we wish to use kernels K only in $\mathcal{A}(\mathcal{X})$ we may always modify them by adding *any* positive multiple of the *delta function kernel* Δ defined, for $x,t\in\mathcal{X}$, as

$$\Delta(x,t) = \begin{cases} 1, & x = t \\ 0, & x \ne t \end{cases} \tag{8}$$

that is, replace K by $K + a\Delta$ where a is a positive constant.

There are two useful cases of a set $\mathcal{K}$ of kernels which are compact and convex. The first is formed by the convex hull of a *finite* number of kernels in $\mathcal{A}_+(\mathcal{X})$. The second example extends this to a compact Hausdorff space Ω, (see, for example, Royden, 1988), and a mapping $G : \Omega \to \mathcal{A}_+(\mathcal{X})$. For each $\omega \in \Omega$, the value of the kernel $G(\omega)$ at $x,t \in \mathcal{X}$ is denoted by $G(\omega)(x,t)$ and we assume that the function of $\omega \mapsto G(\omega)(x,t)$ is continuous on Ω for each $x,t \in \mathcal{X}$. When this is the case we say G is *continuous*. We let $\mathcal{M}(\Omega)$ be the set of all *probability measures* on Ω and observe that

$$\mathcal{K}(G) := \left\{ \int_\Omega G(\omega) dp(\omega) : p \in \mathcal{M}(\Omega) \right\} \tag{9}$$

is a compact and convex set of kernels in $\mathcal{A}_+(\mathcal{X})$. The compactness of the set $\mathcal{K}(G)$ is a consequence of weak*-compactness of the unit ball of the dual space of $C(\Omega)$, the set of all continuous real-valued functions g on Ω with norm $\|g\|_\Omega := \max\{|g(\omega)| : \omega \in \Omega\}$ (Royden, 1988). For example, we choose $\Omega = [a,b]$, where $a > 0$ and $G(\omega)(x,t) = e^{-\omega\|x-t\|^2}$, $x,t \in \mathbb{R}^d$, $\omega \in \Omega$, to obtain *radial kernels*, or $G(\omega)(x,t) = e^{\omega(x,t)}$, $x,t \in \mathbb{R}^d$ to obtain *dot product kernels*. Note that the choice $\Omega = \mathbb{N}_n$ corresponds to our first example.

In preparation for the next theorem we need to express the set $\mathcal{K}(G)$ in an alternate form. We have in mind the following basic fact.

Lemma 3 *If Ω is a compact Hausdorff space, $G : \Omega \to \mathcal{A}_+(\mathcal{X})$ a continuous map as defined above and $\mathcal{G} := \{G(\omega) : \omega \in \Omega\}$ then $\mathcal{K}(G) = \overline{co\mathcal{G}}$.*

PROOF. First, we shall show that $\overline{co\mathcal{G}} \subseteq \mathcal{K}(\mathcal{G})$. To this end, we let $K \in \overline{co\mathcal{G}}$ and $\mathbf{x} \subseteq \mathcal{X}$. By the definition of convex hull, we obtain, for some sequence of probability measures $\{p_\ell : \ell \in \mathbb{N}\}$, that $K_\mathbf{x} = \lim_{\ell\to\infty} \int_\Omega G_\mathbf{x}(\omega) dp_\ell(\omega)$ where each p_ℓ is a *finite* sum of point measures. Since for each $\ell \in \mathbb{N}$, $\int_\Omega G_\mathbf{x}(\omega) dp_\ell(\omega) \in \mathcal{K}(G)$ and $\mathcal{K}(\mathcal{G})$ is closed it follows that $K_\mathbf{x} \subseteq \mathcal{K}(G)$, that is, we have established that $\overline{co\mathcal{G}} \subseteq \mathcal{K}(\mathcal{G})$.

On the other hand, if there is a kernel $K \in \mathcal{K}(G)$ which does not belong to $\overline{co\mathcal{G}}$ then there is an input set $\mathbf{x}$ such that $K_\mathbf{x} \notin \overline{co\mathcal{G}(\mathbf{x})}$ while $K_\mathbf{x} = \int_\Omega G_\mathbf{x}(\omega) dp(\omega)$ for some $p \in \mathcal{M}(\Omega)$. Hence, there exists a hyperplane which separates the matrix $K_\mathbf{x}$ from the set of matrices $\overline{co\mathcal{G}(\mathbf{x})}$ (Royden, 1988). This means that there is a linear functional L on $\mathcal{S}(\mathbb{R}^m)$ and $c \in \mathbb{R}$ such that $L(K_\mathbf{x}) > c$ but $L(G_\mathbf{x}(\omega)) < c$ for all $\omega \in \Omega$. We integrate the last inequality over $\omega \in \Omega$ relative to the measure dp and conclude by the linearity of L that $L(K_\mathbf{x}) < c$, a contradiction. This concludes the proof. □

Observe that the set $\mathcal{G} = \{G(\omega) : \omega \in \Omega\}$ in the above lemma is compact since G is continuous and Ω compact. In general, we wish to point out a useful fact about the kernels in $\overline{co\mathcal{G}}$ *whenever* $\mathcal{G}$ is a *compact* set of kernels. To this end, we recall a theorem of Caratheodory (see, for example, Rockafellar, 1970, Ch. 17).

Theorem 4 *If A is a subset of $\mathbb{R}^n$ then every $a \in coA$ is a convex combination of at most $n+1$ elements of A.*

An immediate consequence of the above theorem is the following fact which we shall use later.

Lemma 5 *If A is a compact subset of $\mathbb{R}^n$ then $\overline{coA}$ is compact and every element in it is a convex combination of at most $n+1$ elements of A.*

In particular, we have the following corollary.

Corollary 6 *If $\mathcal{G}$ is a compact set of kernels on $\mathcal{X}\times\mathcal{X}$ then $\overline{co\mathcal{G}}$ is a compact set of kernels. Moreover, for each input set* $\mathbf{x}$ *a matrix $C\in\overline{co\mathcal{G}(\mathbf{x})}$ if and only if there exists a kernel T which is a convex combination of* at most $\frac{m(m+1)}{2}+1$ *kernels in $\mathcal{G}$ and $T_{\mathbf{x}}=C$.*

Our next result shows whenever $\mathcal{K}$ is the closed convex hull of a compact set of kernels $\mathcal{G}$ that the optimal kernel lies in a the convex hull of some *finite* subset of $\mathcal{G}$.

Theorem 7 *If $\mathcal{G}\subseteq\mathcal{A}_+(\mathcal{X})$ is a compact set of basic kernels, $\mathcal{K}=\overline{co\mathcal{G}}$, $Q:\mathbb{R}^m\to\mathbb{R}_+$ is continuous and μ is a positive number then there exists $\mathcal{T}\subseteq\mathcal{G}$ containing at most $m+2$ basic kernels such that Q_μ admit a minimizer $\tilde{K}\in co\mathcal{T}$ and $Q_\mu(\mathcal{T})=Q_\mu(\mathcal{K})$.*

PROOF. Let $(\hat{c},\hat{K})\in\mathbb{R}^m\times\mathcal{K}$ be a minimizer of Q_μ, that is, we have that

$$Q_\mu(\mathcal{K})=\min\{Q(\hat{K}_{\mathbf{x}}c)+\mu(c,\hat{K}_{\mathbf{x}}c):c\in\mathbb{R}^m\}=Q(\hat{K}_{\mathbf{x}}\hat{c})+\mu(\hat{c},\hat{K}_{\mathbf{x}}\hat{c}).$$

We define the set of vectors $\mathcal{U}:=\{(K_{\mathbf{x}}\hat{c},(\hat{c},K_{\mathbf{x}}\hat{c})):K\in\mathcal{K}\}\subset\mathbb{R}^{m+1}$. Note that $\mathcal{U}=\overline{co\mathcal{V}}$ where $\mathcal{V}=\{(G_{\mathbf{x}}\hat{c},(\hat{c},G_{\mathbf{x}}\hat{c})):G\in\mathcal{G}\}$ and $\mathcal{V}$ is compact since $\mathcal{G}$ is compact. By Lemma 5 the vector $(\hat{K}_{\mathbf{x}}\hat{c},(\hat{c},K_{\mathbf{x}}\hat{c}))$ can be written as a convex combination of at most $m+2$ vectors in $\mathcal{V}$, that is

$$(\hat{K}_{\mathbf{x}}\hat{c},(\hat{c},\hat{K}_{\mathbf{x}}\hat{c}))=(\tilde{K}_{\mathbf{x}}\hat{c},(\hat{c},\tilde{K}_{\mathbf{x}}\hat{c}))$$

where $\tilde{K}$ is the convex combination of at most $m+2$ kernels in $\mathcal{G}$. Consequently, we have that

$$\begin{aligned}Q_\mu(\mathcal{K}) &= Q(\tilde{K}_{\mathbf{x}}\hat{c})+\mu(\hat{c},\tilde{K}_{\mathbf{x}},\hat{c})\\ &\geq \min\{Q(\tilde{K}_{\mathbf{x}}c)+\mu(c,\tilde{K}_{\mathbf{x}}c):c\in\mathbb{R}^m\}\\ &= Q_\mu(\tilde{K})\geq Q_\mu(\mathcal{K})\end{aligned}$$

implying that $Q_\mu(\hat{K})=Q_\mu(\tilde{K})$. □

Note that Theorem 7 asserts the *existence* of a q which is *at most* $m+2$, that is, an optimal kernel is expressed by a convex combination of at most $m+2$ kernels.

Note that in the definition of $Q_\mu(\mathcal{K})$ we minimize first over $f\in\mathcal{H}_K$ and then over $K\in\mathcal{K}$. There arises the question of what would happen if we interchange these minima. We address this issue in the case that $\mathcal{K}$ is the convex hull of a finite set of kernels. To this end, we use the notation $\bigoplus_{j\in\mathbb{N}_n}\mathcal{H}_{K_j}$ for the direct sum of the Hilbert spaces $\{\mathcal{H}_{K_j}:j\in\mathbb{N}_n\}$.

Lemma 8 *If $\mathcal{K}_n=\{K_j:j\in\mathbb{N}_n\}$ is a family of kernels on $\mathcal{X}\times\mathcal{X}$ and $f\in\bigoplus_{j\in\mathbb{N}_n}\mathcal{H}_{K_j}$ then*

$$\inf\{\|f\|_K:K\in co\mathcal{K}_n\}=\min\left\{\sum_{j\in\mathbb{N}_n}\|f_j\|_{K_j}:f=\sum_{j\in\mathbb{N}_n}f_j,\ f_\ell\in\mathcal{H}_{K_\ell},\ell\in\mathbb{N}_n\right\}. \tag{10}$$

As the result is not needed in our subsequent analysis we postpone its proof to the appendix (for related results, see also, Herbster, 2004; Lin and Zhang, 2003). We note that the expression on the right hand side of equation (10) is an *intermediate* norm for $\bigoplus_{j\in\mathbb{N}_n} \mathcal{H}_{K_j}$ (see Bennett and Sharpley, 1988, p. 97) for a discussion. This lemma suggests a reformulation of our extremal problem (7) for kernels of the form (9) where G is expressed in terms of a feature map. Although this fact is interesting, it is not central to our point of view in this paper and, so, we describe it in the appendix.

Next, we establish that the variational problem (7) is a *convex optimization problem*. Specifically, we shall show that if the function mapping $Q : \mathbb{R}^m \to \mathbb{R}$ is convex then the functional $Q_\mu : \mathcal{A}_+(\mathcal{K}) \to \mathbb{R}_+$ is a convex as well. It is curious that this does not seem to follow directly from the *definition* of Q_μ. We take a sojourn through the notion of *conjugate function*. Recall that the conjugate function of Q denoted by $Q^* : \mathbb{R}^m \to \mathbb{R}$ is defined, for every $v \in \mathbb{R}^m$, as

$$Q^*(v) = \sup\{(c,v) - Q(c) : c \in \mathbb{R}^m\}$$

and it follows, for every $c \in \mathbb{R}^m$, that

$$Q(c) = \sup\{(c,v) - Q^*(v) : v \in \mathbb{R}^m\}$$

(see, for example, Rockafellar, 1970; Borwein and Lewis, 2000). A nice recent application of the conjugate function to linear statistical models appears in (Zhang, 2002).

The proof we present below for the convexity of $Q_\mu : \mathcal{A}_+(\mathcal{X}) \to \mathbb{R}_+$ is based upon the von Neumann minimax theorem which we record in the appendix. We begin by introducing for each $r > 0$ a function $\phi_r : \mathbb{R}_+ \to \mathbb{R}_+$ defined, for $t \in \mathbb{R}_+$, as

$$\phi_r(t) := \mu(\frac{1}{2\mu}\sqrt{t} - r)_+^2 - \frac{1}{4\mu}t$$

where $(z)_+ := \max(0,z)$. Note that

$$\lim_{r\to\infty} \phi_r(t) = -\frac{1}{4\mu}t$$

pointwise for $t > 0$. Also, for each fixed $t > 0$, $\phi_r(t)$ is a non-increasing function of r and, for each $r > 0$, ϕ_r is continuously differentiable, decreasing and convex on $\mathbb{R}_+$.

Lemma 9 *If $K \in \mathcal{A}(\mathcal{X})$, $\mathbf{x}$ a set of m distinct points of $\mathcal{X}$ such that $K_\mathbf{x} \in \mathcal{L}_+(\mathbb{R}^m)$ and $Q : \mathbb{R}^m \to \mathbb{R}$ a convex function, then there exists $r_0 > 0$ such that for all $r > r_0$ there holds the formula*

$$Q_\mu(K) = \sup\{\phi_r((v, K_\mathbf{x} v)) - Q^*(v) : v \in \mathbb{R}^m\}. \tag{11}$$

PROOF. By the definition of Q_μ we have that

$$Q_\mu(K) = \min\{\sup\{(K_\mathbf{x}c, v) - Q^*(v) + \mu(c, K_\mathbf{x}c) : v \in \mathbb{R}^m\} : c \in \mathbb{R}^m\}.$$

According to Lemma 2 the minimum above exists. Therefore, there is a $r_0 > 0$ such that for all $r > r_0$ we have that

$$Q_\mu(K) = \min\{\sup\{(K_\mathbf{x}c, v) - Q^*(v) + \mu(c, K_\mathbf{x}c) : v \in \mathbb{R}^m\} : c \in \mathbb{R}^m, (c, K_\mathbf{x}c) \le r^2\}.$$

By the minimax theorem, see Theorem 22 in the appendix, we conclude that

$$Q_\mu(K) = \sup\{\min\{(K_{\mathbf{x}}c, v) - Q^*(v) + \mu(c, K_{\mathbf{x}}c) : c \in \mathbb{R}^m, (c, K_{\mathbf{x}}c) \le r^2\} : v \in \mathbb{R}^m\}.$$

For each $v \in \mathbb{R}^m$, we shall now explicitly compute the minimum of the above expression. To this end, we let $K_{\mathbf{x}} := B^2$ where B is a $m \times m$ positive definite matrix, that is, B is the square root of $K_{\mathbf{x}}$, and observe that

$$\min\{(c, K_{\mathbf{x}}v) + \mu(c, K_{\mathbf{x}}c) : (c, K_{\mathbf{x}}c) \le r^2\} = \min\{\mu\|Bc + \frac{1}{2\mu}Bv\|^2 - \frac{1}{4\mu}\|Bv\|^2 : \|Bc\| \le r\}.$$

If the vector $c_0 := -\frac{1}{2\mu}v$ has the property that $\|Bc_0\| \le r$, that is, $\|Bv\| \le 2\mu r$ then the minimum above is $-\frac{1}{4\mu}\|Bv\|^2$, otherwise $\|Bv\| > 2\mu r$ and the triangle inequality says that

$$\|Bc + \frac{1}{2\mu}Bv\| \ge \frac{1}{2\mu}\|Bv\| - \|Bc\| \ge \frac{1}{2\mu}\|Bv\| - r.$$

Since, for the vector $\hat{c} := -\frac{vr}{\|Bv\|}$, we have that

$$\|B\hat{c} + \frac{1}{2\mu}Bv\| = \frac{1}{2\mu}\|Bv\| - r$$

this inequality is sharp. Therefore, we get that

$$Q_\mu(K) = \sup\left\{\mu(\frac{1}{2\mu}\|Bv\| - r)_+^2 - \frac{1}{4\mu}\|Bv\|^2 - Q^*(v) : v \in \mathbb{R}^m\right\}$$

and the result follows by the definition of ϕ_r. □

Let us specialize this lemma to the example of the square loss S defined, for $w \in \mathbb{R}^m$, as $S(w) = \|y - w\|^2$. In this case, the conjugate function is given explicitly for $v \in \mathbb{R}^m$ as

$$S^*(v) = \max\{(w, v) - \|w - y\|^2 : w \in \mathbb{R}^m\} = \frac{1}{4}\|v\|^2 + (y, v).$$

We shall show later in Lemma 14 by a *direct* computation *without* the use of the conjugate function that $S_\mu = \mu(y, (K_{\mathbf{x}} + \mu I)^{-1}y)$. Alternatively, if we formally let $r = \infty$ in the right hand side of equation (11) we get

$$\sup\left\{-\frac{1}{4\mu}(v, (K_{\mathbf{x}} + \mu I)v) - (y, v) : v \in \mathbb{R}^m\right\}$$

which by a direct computation equals $\mu(y, (K_{\mathbf{x}} + \mu I)^{-1}y)$. This suggests that Lemma 9 may even hold when $r = \infty$ and without the hypothesis that $K_{\mathbf{x}} \in \mathcal{L}_+(\mathbb{R}^m)$. We shall confirm this with another version of the von Neumann minimax theorem.

Lemma 10 *If $K \in \mathcal{A}(\mathcal{X})$, $\mathbf{x}$ a set of m distinct points of $\mathcal{X}$ such that $K_{\mathbf{x}} \in \mathcal{L}_+(\mathbb{R}^m)$ and $Q : \mathbb{R}^m \to \mathbb{R}$ a convex function, then there holds the formula*

$$Q_\mu(K) = \sup\left\{-\frac{1}{4\mu}(v, K_{\mathbf{x}}v) - Q^*(v) : v \in \mathbb{R}^m\right\}.$$

PROOF. Theorem 23 applies since $K_{\mathbf{x}} \in \mathcal{L}_+(\mathbb{R}^m)$. Indeed, we let $f(c,v) = (K_x c, v) - Q^*(v) + \mu(c, K_{\mathbf{x}} c)$, $\mathcal{A} = \mathcal{B} = \mathbb{R}^m$ and $v_0 = 0$ then the set $\{c : c \in \mathbb{R}^m, f(c,v_0) \leq \lambda\}$ is compact and all the hypotheses of Theorem 23 hold. Hence, we may proceed as in the proof of Lemma 9 with $r = \infty$. □

To interpret Lemma 9, we say that $A \preceq B$ whenever $A, B \in \mathcal{L}(\mathbb{R}^m)$, if $B - A$ is positive semi-definite. We also say, for $K, J \in \mathcal{A}(X)$, that $K \preceq J$ if $K_{\mathbf{x}} \preceq J_{\mathbf{x}}$ for every $\mathbf{x} \subseteq X$.

Definition 11 *A function $\phi : \mathcal{B} \to \mathbb{R}$ is said non–decreasing on $\mathcal{B} \subseteq \mathcal{A}(X)$ if, for every $A, B \in \mathcal{B}$ with $A \preceq B$ it follows that $\phi(A) \leq \phi(B)$. If the reverse inequality holds we say ϕ is non–increasing.*

Definition 12 *A function $\phi : \mathcal{B} \to \mathbb{R}$ is said convex on $\mathcal{B} \subseteq \mathcal{A}(X)$ if, for every $A, B \in \mathcal{B}$ and $\lambda \in [0,1]$ there holds the inequality*

$$\phi(\lambda A + (1-\lambda)B) \leq \lambda \phi(A) + (1-\lambda)\phi(B). \tag{12}$$

If the reverse of inequality (12) holds we say that the ϕ is concave.

Proposition 13 *If $Q : \mathbb{R}^m \to \mathbb{R}_+$ is convex then for every $\mu > 0$ the function $Q_\mu : \mathcal{A}_+(X) \to \mathbb{R}_+$ is convex and non-increasing.*

PROOF. The proof of the proposition follows from Lemma 9. Specifically, equation (11) expresses Q_μ as the supremum of a family of functions which are convex and non-increasing on $\mathcal{A}(X)$. □

We note that the convexity of the function Q_μ was already proven by Lanckriet et al. (2004) for the hinge loss and stated in (Ong, Smola and Williamson, 2003) for differentiable convex loss functions.

3. Square Regularization

In this section we exclusively study the case of the square loss regularization functional S_μ in equation (5) and provide improvements and simplifications of our previous results. We begin by determining the *explicit* expression for this functional which we briefly mentioned earlier after the proof of Lemma 9.

Lemma 14 *For any kernel K, inputs $\mathbf{x} := \{x_j : j \in \mathbb{N}_m\}$, samples $y = (y_j : j \in \mathbb{N}_m)$ and positive constant μ we have that*

$$S_\mu(K) = \mu(y, (\mu I + K_{\mathbf{x}})^{-1} y) \tag{13}$$

where I is the $m \times m$ identity matrix.

PROOF. We have that $S_\mu(K) = \min\{R(c) : c \in \mathbb{R}^m\}$ where for each $c \in \mathbb{R}^m$ we set $R(c) := \|y - K_{\mathbf{x}} c\|^2 + \mu(c, K_{\mathbf{x}} c)$. We define the vector $w := (\mu I + K_{\mathbf{x}})^{-1} y$, observe that $R(w) = (y, \mu(\mu I + K_{\mathbf{x}})^{-1} y)$ and for every vector $c \in \mathbb{R}_m$ we have that

$$R(c) = R(w) + \|K_{\mathbf{x}}(w - c)\|^2 + \mu(c - w, K_{\mathbf{x}}(c - w)).$$

With this formula the result follows. □

From this lemma we conclude, when the matrix $K_{\mathbf{x}}$ is in $\mathcal{L}_+(\mathbb{R}^m)$ then $\lim_{\mu\to 0}\mu^{-1}S_\mu(K)=\gamma(K_{\mathbf{x}})$, where for every $A\in\mathcal{L}_+(\mathbb{R}^m)$ we set $\gamma(A):=(y,A^{-1}y)$. The function $\gamma:\mathcal{L}_+(\mathbb{R}^m)\to\mathbb{R}_+$ has the alternate form

$$\frac{1}{\gamma(A)}:=\min\{(c,Ac):c\in\mathbb{R}^m,(c,y)=1\},\ \ A\in\mathcal{L}_+(\mathbb{R}^m) \tag{14}$$

and the unique vector which achieves this minimum is given by

$$c(A):=\frac{A^{-1}y}{(y,A^{-1}y)}. \tag{15}$$

A proof of these facts follow directly from the Cauchy-Schwarz inequality for the inner product (u,Av), $u,v\in\mathbb{R}^m$. Moreover, this alternate form for $\gamma(A)$ connects the function γ to the *minimal norm interpolant* in $\mathcal{H}_K$ to the data D. Let us explain this connection next.

Recall, for every kernel K on $X\times X$, that the minimal norm interpolation to the data D is the solution to the variational problem

$$\rho(K):=\min\{\|f\|_K^2:f\in\mathcal{H}_K,f(x_j)=y_j,j\in\mathbb{N}_m\}. \tag{16}$$

The following result is well-known (for a proof see, for example, Micchelli and Pontil, 2005).

Proposition 15 *If $K\in\mathcal{A}(X)$ and $\mathbf{x}$ is an input set in X such that the matrix $K_{\mathbf{x}}$ is in $\mathcal{L}_+(\mathbb{R}^m)$ then the solution of the minimal norm interpolation problem (16) is unique and is given by*

$$f=\sum_{j\in\mathbb{N}_m}c_jK(x_j,\cdot)$$

where the coefficient vector $c=(c_j:j\in\mathbb{N}_m)$ solves the linear system of equations $K_{\mathbf{x}}c=y$ and we have that

$$\rho(K)=\gamma(K_{\mathbf{x}})=(y,K_{\mathbf{x}}^{-1}y). \tag{17}$$

The function $\gamma:\mathcal{L}_+(\mathbb{R}^m)\to\mathbb{R}_+$ is continuous. We record additional facts about this function in the next two lemmas.

Lemma 16 *The function γ is non–increasing and whenever $A,B\in\mathcal{L}_+(\mathbb{R}^m)$, $\gamma(A)=\gamma(B)$ if and only if $A^{-1}y=B^{-1}y$.*

PROOF. If $A\preceq B$ then for every $c\in\mathbb{R}^m$, $(c,Ac)\le(c,Bc)$ and it follows that $\frac{1}{\gamma(A)}\le\frac{1}{\gamma(B)}$. Clearly $A^{-1}y=B^{-1}y$ implies that $\gamma(A)=\gamma(B)$. On the other hand if $\gamma(A)=\gamma(B)$, the inequalities $\frac{1}{\gamma(A)}\le(c(B),Ac(B))\le(c(B),Bc(B))=\frac{1}{\gamma(B)}$ imply that $c(A)=c(B)$ and the result follows. □

Lemma 17 *The function γ is convex and the function γ^{-1} concave. Moreover, for every $A,B\in\mathcal{L}_+(\mathbb{R}^m)$, $\lambda\in[0,1]$, we have that*

$$\frac{1}{\gamma(\lambda A+(1-\lambda)B)}=\lambda\frac{1}{\gamma(A)}+(1-\lambda)\frac{1}{\gamma(B)} \tag{18}$$

if and only if $c(A)=c(B)=c(\lambda A+(1-\lambda)B)$.

PROOF. For every $\lambda \in [0,1]$ we define the matrix $D_\lambda = \lambda A + (1-\lambda)B$ and for all $c \in \mathbb{R}^m$ for which $(c,y) = 1$ note that

$$(c, D_\lambda c) = \lambda(c, Ac) + (1-\lambda)(c, Bc) \geq \lambda \frac{1}{\gamma(A)} + (1-\lambda)\frac{1}{\gamma(B)}. \tag{19}$$

Consequently, we have that $\frac{1}{\gamma(D_\lambda)} \geq \lambda \frac{1}{\gamma(A)} + (1-\lambda)\frac{1}{\gamma(B)}$, showing that γ^{-1} is concave. Alternatively, equation (14) expresses $\gamma^{-1}(A)$ as the minimum of a family of functions which are linear in the matrix A and hence γ^{-1} is concave. Similarly, using this equation we have that

$$\gamma(A) = \max\left\{(c, Ac)^{-1} : c \in \mathbb{R}^m, (c,y) = 1\right\}$$

thereby expressing γ as a maximum of a family of convex functions.

If (18) holds, we choose $c = c_\lambda := c(D_\lambda)$ in (19) and conclude by the uniqueness of the vector $c(A)$ in equation (15) that $c_\lambda = c(A) = c(B)$. Conversely, when this conclusion holds we have that

$$\begin{aligned} \frac{1}{\gamma(D_\lambda)} &= \lambda(c_\lambda, Ac_\lambda) + (1-\lambda)(c_\lambda, Bc_\lambda) \\ &= \lambda(c(A), Ac(A)) + (1-\lambda)(c(B), Bc(B)) \\ &= \lambda \frac{1}{\gamma(A)} + (1-\lambda)\frac{1}{\gamma(B)} \end{aligned}$$

which concludes the proof. □

Lemma 16 and 17 established that the function $\phi : \mathcal{L}_+(\mathbb{R}^m) \to \mathbb{R}$ defined, for some $d \in \mathbb{R}^m$ and all $A \in \mathcal{L}_+(\mathbb{R}^m)$, as $\phi(A) = (d, A^{-1}d)$ is non-increasing and convex (see also the work of Marshall and Olkin, 1979).

Proposition 15 and Lemma 14 connects minimal norm interpolation to square loss regularization. This connection allows us in this section to turn our attention to the function $\rho : \mathcal{A}(\mathcal{X}) \to \mathbb{R}_+$ and consider the variational problem

$$\rho(\mathcal{K}) := \inf\{\rho(K) : K \in \mathcal{K}\} \tag{20}$$

where $\mathcal{K}$ is a prescribed set of kernels. The approach of Lemma 2 applies directly to establish the following lemma.

Lemma 18 *If $\mathcal{K}$ is a compact and convex set of kernels in $\mathcal{A}_+(\mathcal{X})$ then the minimum of (20) exists.*

Our next result describes the solution of the problem of determining $\rho(\mathcal{K})$ for the case that $\mathcal{K} = co\mathcal{K}_a$ where $\mathcal{K}_a = \{K_\ell : \ell \in \mathbb{N}_n\}$ is a prescribed finite subset of $\mathcal{A}_+(\mathcal{X})$. In its presentation we use the notion $K_{\mathbf{x},\ell}$ for the matrix $(K_\ell)_{\mathbf{x}}$.

Theorem 19 *If $\mathcal{K}_a = \{K_j : j \in \mathbb{N}_n\} \subset \mathcal{A}_+(\mathcal{X})$ there exists a kernel $\hat{K} = \sum_{j\in J} \lambda_j K_j \in co\mathcal{K}_a$, where $J \subseteq \mathbb{N}_n$, $card(J) \leq min(m+1, n)$ with $\sum_{j\in J} \lambda_j = 1$ such that, for every $j \in J$, $\lambda_j > 0$,*

$$(\hat{c}, K_{\mathbf{x},j}\hat{c}) = \max\{(\hat{c}, K_{\mathbf{x},\ell}\hat{c}) : \ell \in \mathbb{N}_n\}, \;\; \hat{c} = c(\hat{K}_{\mathbf{x}}),$$

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

and for every $c \in \mathbb{R}^m$ *with* $(c,y) = 1$ *and every* $K \in co\mathcal{K}_a$

$$(\hat{c}, K_{\mathbf{x}}\hat{c}) \leq (\hat{c}, \hat{K}_{\mathbf{x}}\hat{c}) \leq (c, \hat{K}_{\mathbf{x}}c). \tag{21}$$

Inequality (21) expresses the fact that the pair $(\hat{c}, \hat{K})$ is a *saddle point* for the minimax problem

$$\tilde{\rho}^{-1} = \min\{\max\{(c, K_{\mathbf{x}}c) : K \in co\mathcal{K}_a\} : c \in \mathbb{R}^m, (c,y) = 1\}.$$

The existence of $(\hat{c}, \hat{K})$ above implies that the minimum and maximum can be interchanged, that is,

$$\max\{\min\{(c, K_{\mathbf{x}}c) : c \in \mathbb{R}^m, (c,y) = 1\} : K \in co\mathcal{K}_a\} \tag{22}$$

$$= \quad \min\{\max\{(c, K_{\mathbf{x}}c) : K \in co\mathcal{K}_a\} : c \in \mathbb{R}^m, (c,y) = 1\}. \tag{23}$$

Moreover, *any* $\hat{c}$ and $\hat{K}$ with the properties described in Theorem 19 is a saddle point of this minimax problem. Indeed, the upper bound in (21) follows from the definition of the vector $\hat{c}$ and the function γ defined earlier, see equations (14) and (15). The lower bound follows from the fact that for any $K \in co\mathcal{K}_a$ we have that $(\hat{c}, K_{\mathbf{x}}\hat{c}) \leq \max\{(\hat{c}, K_{\mathbf{x},\ell}\hat{c}) : \ell \in \mathbb{N}_n\}$.

Let us now turn to the existence of $\hat{K}$. Note that by equation (14) and Proposition 15 the expression in (22) is $1/\rho(\mathcal{K})$, the reciprocal of the quantity of interest to us. It is the quantity in equation (23) which we examine in the proof of Theorem 19 and it has been denoted by $\tilde{\rho}^{-1}$. A consequence of Theorem 19 is that $\tilde{\rho} = \rho(\mathcal{K})$. Certainly, by their definitions it is clear that $\tilde{\rho} \leq \rho(\mathcal{K})$.

We now present the proof of Theorem 19.

PROOF. Let $\tilde{c}$ be a solution to problem (23). We define the set

$$J^* \equiv J(\tilde{c}) := \left\{ j : j \in \mathbb{N}_n, (\tilde{c}, K_{\mathbf{x},j}\tilde{c}) = \max\{(\tilde{c}, K_{\mathbf{x},i}\tilde{c}) : i \in \mathbb{N}_n\} \right\}$$

the convex function $\varphi : \mathbb{R}^m \to \mathbb{R}$ by setting for each $c \in \mathbb{R}^m$, $\varphi(c) := \max\{(c, K_{\mathbf{x},j}c) : j \in \mathbb{N}_n\}$ and note that by Lemma 24 the directional derivative of φ along the "direction" $d \in \mathbb{R}^m$, denoted by $\varphi'_+(c;d)$, is given by

$$\varphi'_+(c;d) = 2\max\{(d, K_{\mathbf{x},j}c) : j \in J(c)\}.$$

Since $\tilde{c}$ is a minimum for (14) we have that

$$\max\{(d, K_{\mathbf{x},j}\tilde{c}) : j \in J^*\} \geq 0$$

for every $d \in \mathbb{R}^m$ such that $(d,y) = 0$. Let $\mathcal{M}$ be the convex hull of the set of vectors $\mathcal{N} := \{K_{\mathbf{x},j}\tilde{c} : j \in J^*\} \subset \mathbb{R}^m$. Since $\mathcal{M} \subseteq \mathbb{R}^m$, by the Caratheodory theorem (see, for example, Rockafellar, 1970, Ch. 17) every vector in $\mathcal{M}$ can be expressed as a convex combination of at most $q := \min(m+1, |J^*|) \leq \min(m+1, n)$ elements of $\mathcal{N}$. We will show that $\mathcal{M}$ intersects the line spanned by the vector y. Indeed, if these two sets did not intersect then there exists a hyperplane $\{c : c \in \mathbb{R}^m, (w,c) + \alpha = 0\}$, where $\alpha \in \mathbb{R}$, $w \in \mathbb{R}^m$, which strictly separates them, that is,

$$(w, ty) + \alpha > 0, \quad t \in \mathbb{R}$$

and

$$(w, K_{\mathbf{x},j}\tilde{c}) + \alpha < 0, \quad j \in J^*, \tag{24}$$

(see, for example, Royden, 1988).

The first condition, for $t = 0$, implies that $\alpha > 0$ and since t can take any real value we also have that $(w,y) = 0$. Consequently, from equation (24) we get that

$$\max\{(w, K_{\mathbf{x},j}\tilde{c}) : j \in J^*\} < 0$$

which contradicts our hypothesis that $\tilde{c}$ is a minimum. Thus, it must be the case that $t_0 y \in \mathcal{M}$ for some $t_0 \in \mathbb{R}$, that is,

$$t_0 y = \sum_{j\in J} \gamma_j K_{\mathbf{x},j}\tilde{c} \tag{25}$$

for some subset J of J^* of cardinality at most q and positive constants γ_j with $\sum_{j\in J}\gamma_j = 1$. Taking the inner product of both sides of equation (25) with $\tilde{c}$, and recalling the fact that $(\tilde{c},y) = 1$ we obtain that $t_0 = \tilde{\rho}^{-1}$. Setting

$$\hat{K} := \sum_{j\in J} \gamma_j K_j$$

we have from (25) that $\tilde{c} = \tilde{\rho}^{-1}\hat{K}_{\mathbf{x}}^{-1} y$, and $\tilde{\rho} = (y, \hat{K}_{\mathbf{x}}^{-1} y)$. Therefore, by Proposition 15 we conclude that $\tilde{\rho} = \rho(\hat{K})$ and $\tilde{c} = \hat{c}$ where $\hat{c}$ is defined in the theorem. In particular, we obtain $\tilde{\rho} \geq \rho(\mathcal{K})$ and so by our previous remarks just before the beginning of the proof, we conclude that $\tilde{\rho} = \rho(\mathcal{K})$. □

Recall, that earlier we introduced the class $\mathcal{K}(G)$ induced by a continuous mapping $G : \Omega \to \mathcal{A}_+(X)$ where Ω is a compact Hausdorff space. Theorem 15 extends to this generality. No essential difference occur in the proof. However, the conclusion is striking. Not only do we characterize the optimal kernel $\hat{K} \in \mathcal{K}(G)$ but we show that it comes from a *discrete* probability measure $\hat{p} \in \mathcal{M}(\Omega)$ with *at most* $m+1$ *atoms*, that is, $\hat{K} = \int_\Omega G(\omega)d\hat{p}(\omega)$.

Theorem 20 *If Ω is a compact Hausdorff topological space and $G : \Omega \to \mathcal{A}_+(X)$ is continuous then there exists a kernel $\hat{K} = \int_\Omega G(\omega)d\hat{p}(\omega) \in \mathcal{K}(G)$ such that $\hat{p}$ is a discrete probability measure in $\mathcal{M}(\Omega)$ with at most $m+1$ atoms. Moreover, for any atom $\hat{\omega} \in \Omega$ of $\hat{p}$, we have that*

$$(\hat{c}, G_{\mathbf{x}}(\hat{\omega})\hat{c}) = \max\{(\hat{c}, G_{\mathbf{x}}(\omega)\hat{c}) : \omega \in \Omega\}, \;\; \hat{c} = c(\hat{K}_{\mathbf{x}}),$$

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

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

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

PROOF. Let $\tilde{c}$ be a solution to problem (23) where $co\mathcal{K}_a$ is replaced by $\mathcal{K}(G)$ and define the set

$$\Omega^* \equiv \Omega(\tilde{c}) := \{\tau : \tau \in \Omega, (\tilde{c}, G_{\mathbf{x}}(\tau)\tilde{c}) = \max\{(\tilde{c}, G_{\mathbf{x}}(\omega)\tilde{c}) : \omega \in \Omega\}\}.$$

where we denoted the matrix $(G(\omega))_{\mathbf{x}}$ by $G_{\mathbf{x}}(\omega)$. We define the convex function $\varphi : \mathbb{R}^m \to \mathbb{R}$ by setting for each $c \in \mathbb{R}^m$, $\varphi(c) := \max\{(c, G_{\mathbf{x}}(\omega)c) : \omega \in \Omega\}$ and note that by Lemma 24 the directional derivative of φ along the "direction" $d \in \mathbb{R}^m$, denoted by $\varphi'_+(c;d)$, is given by

$$\varphi'_+(c;d) = 2\max\{(d, G_{\mathbf{x}}(\omega)c) : \omega \in \Omega^*\}.$$

Since $\tilde{c}$ is a minimum for (14) we have that

$$\max\{(d, G_{\mathbf{x}}(\omega)\tilde{c}) : \omega \in \Omega(c)\} \geq 0$$

for every $d \in \mathbb{R}^m$ such that $(d,y) = 0$. Let $\mathcal{M}$ be the convex hull of the set of vectors $\mathcal{N} := \{G_{\mathbf{x}}(\omega)\tilde{c} : \omega \in \Omega^*\} \subset \mathbb{R}^m$. Since $\mathcal{M} \subseteq \mathbb{R}^m$, by the Caratheodory theorem every vector in $\mathcal{M}$ can be expressed as a convex combination of at most $m+1$ elements of $\mathcal{N}$. We will show that $\mathcal{M}$ intersects the line spanned by the vector y. Indeed, if these two sets did not intersect then there exist a hyperplane $(w,c) + \alpha = 0$, $\alpha \in \mathbb{R}$, $w \in \mathbb{R}^m$, which strictly separates them, that is,

$$(w,ty) + \alpha > 0, \;\; t \in \mathbb{R}$$

and

$$(w, G_{\mathbf{x}}(\omega)\tilde{c}) + \alpha < 0, \;\; \omega \in \Omega^*, \tag{26}$$

(see Royden, 1988).

The first condition, for $t = 0$, implies that $\alpha > 0$ and since t can take any real value we also have that $(w,y) = 0$. Consequently, from equation (26) we get that

$$\max\{(w, G_{\mathbf{x}}(\omega)\tilde{c}) : \omega \in \Omega^*\} < 0.$$

which contradicts our hypothesis that $\tilde{c}$ is a minimum. Thus, it must be the case that $t_0 y \in \mathcal{M}$ for some $t_0 \in \mathbb{R}$, that is,

$$t_0 y = \int_{\Omega} G_{\mathbf{x}}(\omega)\tilde{c}\, d\hat{p}(\omega) \tag{27}$$

where $\hat{p} \in \mathcal{M}(\Omega)$ is a discrete probability measure with at most $m+1$ atoms. Taking the inner product of both sides of equation (27) with $\tilde{c}$, and recalling the fact that $(\tilde{c},y) = 1$ we obtain that $t_0 = \tilde{\rho}^{-1}$. Setting

$$\hat{K} := \int_{\Omega} G_{\mathbf{x}}(\omega)\, d\hat{p}(\omega)$$

we have from (27) that $\tilde{c} = \tilde{\rho}^{-1}\hat{K}^{-1}y$, and $\tilde{\rho} = (y, \hat{K}^{-1}y)$. Therefore, by Proposition 15 we conclude that $\tilde{\rho} = \rho(\hat{K})$ and $\tilde{c} = \hat{c}$ where $\hat{c}$ is defined in the theorem. In particular, we obtain $\tilde{\rho} \geq \rho(\mathcal{K})$ and so by our previous remarks we conclude that $\tilde{\rho}_0 = \rho(\mathcal{K})$. □

This theorem applies to the Gaussian kernel.

Corollary 21 *If $a > 0$ and $N : [a,b] \to \mathcal{A}_+(\mathcal{X})$ is defined as*

$$N(\omega)(x,t) = e^{-\omega\|x-t\|^2}, \;\; x,t \in \mathbb{R}^d, \; \omega \in \mathbb{R}_+$$

then there exists a kernel $\hat{K} = \int_{\Omega} N(\omega)\, d\hat{p}(\omega) \in \mathcal{K}(N)$ such that $\hat{p}$ is a discrete probability measure in $\mathcal{M}(\Omega)$ with at most $m+1$ atoms. Moreover, for any atom $\hat{\omega} \in \Omega$ of $\hat{p}$, we have that

$$(\hat{c}, N_{\mathbf{x}}(\hat{\omega})\hat{c}) = \max\{(\hat{c}, N_{\mathbf{x}}(\omega)\hat{c}) : \omega \in \Omega\}, \;\; \hat{c} = c(\hat{K}_{\mathbf{x}}),$$

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

and for every $c \in \mathbb{R}^m$ and $K \in \mathcal{K}(N)$ we have that

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

We note that, in view of equations (13) and (17), Theorem 19 and Theorem 20 apply directly, up to an unimportant constant μ, to the square loss functional by merely adding the kernel $\mu\Delta$ to the class of kernels considered in these theorems. That is, we minimize the quantity in equation (17) over the compact convex set of kernels

$$\tilde{\mathcal{K}} = \{\tilde{K} : \tilde{K} = K + \mu\Delta,\ K \in \mathcal{K}\}$$

where the kernel Δ is defined in equation (8).

An important example of the above construction is to choose K_j to be polynomials on $\mathbb{R}^d$, namely $K_j(x,t) = (x,t)^j$, $x,t \in \mathbb{R}^d$. From a practical point of view we should limit the range of the index j and therefore Theorem 19 adequately covers this case. On the contrary if we decide to use, as it is done often, Gaussians, there arises not only how many Gaussians to choose but also which ones to choose. This raises the question of looking at the *whole class of radial basis functions* and trying to choose the best kernel amongst this class. To this end, we recall a beautiful result of Schoenberg (1938). Let φ be a real–valued function defined on $\mathbb{R}_+$ which we normalize so that $\varphi(0) = 1$. We form a kernel K on $\mathbb{R}^d$ by setting for each $x,t \in \mathbb{R}^d$ $K(x,t) = \varphi(\|x-t\|^2)$. Schoenberg showed that K is positive definite for *any* d if and only if there is a probability measure p on $\mathbb{R}_+$ such that

$$K(x,t) = \int_{\mathbb{R}_+} e^{-\sigma\|x-t\|^2} dp(\sigma),\ \ x,t \in \mathbb{R}^d.$$

Note that the set $\mathbb{R}_+$ is *not* compact and the kernel $N(0)$ is not in $\mathcal{A}_+(\mathbb{R}^d)$. Therefore, on both accounts Theorem 20 does not apply in this circumstance unless, of course, we impose a positive lower bound and a finite upper bound on the variance of the Gaussian kernels $N(\omega)$. We may overcome this difficulty by a limiting process which can handle kernel maps on *locally compact* Hausdorff spaces. This will lead us to an extension of Theorem 20 where Ω is locally compact. However, we only describe our approach in detail for the Gaussian case and $\Omega = \mathbb{R}_+$. An important ingredient in this discussion presented below is that $N(\infty) = \Delta$.

For every $\ell \in \mathbb{N}$ we consider the Gaussian kernel map on the interval $\Omega_\ell := [\ell^{-1}, \ell]$ and appeal to Theorem 20 to produce a sequence of kernels $\hat{K}_\ell = \int_{\Omega_\ell} N(\omega) d p_\ell(\omega)$ with the properties described there. In particular, p_ℓ is a discrete probability measure with at most $m+1$ atoms, a number *independent* of ℓ. Let us examine that may happen as ℓ tends towards infinity. Each of the atoms of p_ℓ as well their corresponding weights have subsequences which converge. Some atoms may converge to zero while others to infinity. In either case, the Gaussian kernel map *approaches a limit*. Therefore, we can extract a convergent subsequence $\{p_{n_\ell} : \ell \in \mathbb{N}\}$ of probability measures and kernels $\{K_{n_\ell} : \ell \in \mathbb{N}\}$ such that $\lim_{\ell\to\infty} p_{n_\ell} = \hat{p}$, $\lim_{\ell\to\infty} K_{n_\ell} = \hat{K}$, and $\hat{K} = \int_{\mathbb{R}_+} N(\omega)\hat{p}(\omega)$ with the provision that $\hat{p}$ may have atoms at either zero of infinity. In either case, we replace the Gaussian by its limit, namely $N(0)$, the identically one kernel, or $N(\infty)$, the delta kernel, in the integral which defines $\hat{K}$. *All* of the properties described in Theorem 20 and remarks following it hold for $\hat{K}$ because of the simplicity of the objective function for the minimax problem studied there. Hence $\hat{K}$ is the *best radial kernel*.

4. Discussion

In this final section we comment on two recent papers related to ours, present some numerical simulations and outline possible extensions of the ideas presented above.

4.1 Related Works

Lanckriet et al. (2004) address learning kernels in the context of transductive learning, that is, learning the value of a function at a finite set of test points. In this case the kernel is computed only on the training and test sets and, so, it is regarded as a matrix. The authors propose different criteria to find a positive semi-definite kernel matrix and discuss how these can be casted as positive semi-definite programming problems. For example, they maximize the *margin* of a binary support vector machine (SVM) trained with the kernel K, which is the square root of the reciprocal of the quantity defined by the equation

$$\rho_{hard}(K) = \min\left\{\|f\|_K^2 : y_j f(x_j) \geq 1, j \in \mathbb{N}_m\right\}. \tag{28}$$

where $y_j \in \{-1,1\}$ are class labels, (see, for example, Vapnik, 1998). The margin is the maximum distance of the closed point, relative to a set of labeled points, amongst all separating functions in the RKHS. These functions are hyperplanes in the space spanned by the features associated to a Mercer expansion of the kernel K. When the optimal separating hyperplane does not exist, the standard approach is to relax the separation constraints in problem (28) to obtain the so-called soft margin SVM,

$$\rho_{soft}(K) := \min\left\{\sum_{j\in\mathbb{N}_m} \xi_j + \mu\|f\|_K^2 : y_j f(x_j) \geq 1-\xi_j,\ \xi_j \geq 0, j \in \mathbb{N}_m,\ f \in \mathcal{H}_K\right\}. \tag{29}$$

These two problems are related. Indeed, if problem (28) admits a solution, that is, the constraints are feasible, problem (29) gives the same solution provided the parameter μ is small enough.

Lanckriet et al. (2004) consider the minimization problem (29) when $\mathcal{K}$ is a set of positive semi-definite matrices which are linear combinations of some prescribed matrices $K_j, j \in \mathbb{N}_n$. In particular, if K_j are positive semi-definite $\mathcal{K}$ could be the set of convex combination of such matrices. They show that $\rho_{soft}(K)$ is convex in $K \in \mathcal{K}$. Our observations in Section 2 confirm that the margin and the soft margin are convex functions of the kernel. Indeed, problem (29) is equivalent to the variational problem (1) when Q is the *hinge error function* defined on $\mathbb{R}^m$ by

$$Q(w) := \sum_{j\in\mathbb{N}_n} (1-y_j w_j)_+,\ \ w := (w_j : j \in \mathbb{N}_m)$$

where $t_+ := \max(0,t), t \in \mathbb{R}$, (see, for example, Evgeniou, Pontil and Poggio, 2000).

Ong, Smola and Williamson (2003) consider learning a kernel function rather than a kernel matrix. They choose a set $\mathcal{K}$ in the space of kernels which are in a Hilbert space of functions generated by a so-called hyper-kernel. This is a kernel $H : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$, where $\mathcal{X}^2 = \mathcal{X} \times \mathcal{X}$, with the property that, for every $(x,t) \in \mathcal{X}^2$, $H((x,t),(\cdot,\cdot))$ is a kernel on $\mathcal{X} \times \mathcal{X}$. This construction includes convex combinations of a possibly infinite number or kernels provided they are *pointwise nonnegative*. For example Gaussian kernels or polynomial kernels with even degree satisfy this assumption although those with odd degree, such as linear kernels or other radial kernels do not.

4.2 Numerical Simulations

In this section we discuss two numerical simulations we carried out to compute a convex combination of a finite set of kernels $\{K_\ell : \ell \in \mathbb{N}_n\}$ which minimizes the square loss regularization functional

μ	10^{-4}	10^{-3}	10^{-2}	0.1	1	10
Method 1	2.41 (1.04)	1.69 (0.68)	0.60 (0.11)	0.27 (0.08)	**0.26 (0.05)**	3.20 (0.48)
Method 2	1.54 (0.58)	0.91 (0.22)	0.47 (0.08)	**0.40 (0.07)**	0.61 (0.11)	3.80 (0.58)
Method 3	4.65 (7.81)	0.95 (1.24)	0.21 (0.06)	**0.10 (0.05)**	0.12 (0.08)	2.40 (0.60)

Table 1: *Experiment 1: Average mean square error with its standard deviation (in parenthesis) for methods 1 to 3 for different values of the regularization parameter μ (see text for the description). The unit measure for the errors is 10^{-3}.*

S_μ in equation (5). For this purpose, we use an interior point method, that is, we define, for every $\lambda = (\lambda_\ell : \ell \in \mathbb{N}_n) \in \mathbb{R}^n$, the penalized function

$$F_\nu(\lambda) := S_\mu\left(\sum_{\ell\in\mathbb{N}_n} \lambda_\ell K_\ell\right) - \nu \sum_{\ell\in\mathbb{N}_n} \ln\lambda_\ell \tag{30}$$

where ν is a positive parameter and solve the variational problem

$$\min\left\{F_\nu(\lambda) : \lambda \in \mathbb{R}^n,\ \sum_{\ell\in\mathbb{N}_n} \lambda_\ell = 1\right\}. \tag{31}$$

Clearly, when ν is small the solution to this problem is close to a minimizer of S_μ, although the penalty term in (30) forces this solution to be interior to the set $\{\lambda : \sum_{\ell\in\mathbb{N}_n} \lambda_\ell = 1,\ \lambda_\ell \geq 0,\ \ell \in \mathbb{N}_n\}$. In order to reach such a minimizer we choose an iteration number $R \in \mathbb{N}$ and iteratively compute the solution to problem (31) for a decreasing sequence of values of the parameter ν. Specifically we set, for $r \in \mathbb{N}_R$, $\nu_r = \overline{\nu}A^{r-1}$ where $\overline{\nu}$ is the initial value of ν and $A \in (0,1)$ is some prescribed parameter. The optimality conditions for problem (31) (see, for example, Rockafellar, 1970; Borwein and Lewis, 2000) are given by the system of *non-linear* equations

$$\begin{aligned} \nabla F_\nu - \eta e &= 0 \\ -(e,\lambda) + 1 &= 0 \end{aligned}$$

where e is the vector in $\mathbb{R}^n$ all of whose components are one and $\eta \in \mathbb{R}$ is the Lagrange multiplier associated to the equality constraint in that problem. We solve these equations by a Newton method (see, for example, Mangasarian, 1994) which consists in iteratively solving the system of *linear* equations

$$\begin{aligned} \nabla^2 F_\nu(\hat{\lambda})\Delta_\lambda - \Delta_\eta e &= \hat{\eta} e - \nabla F_\nu(\hat{\lambda}) \\ -(e,\Delta_\lambda) &= 0 \end{aligned}$$

to obtain the vector $\Delta_\lambda \in \mathbb{R}^m$ and $\Delta_\eta \in \mathbb{R}$, where $\hat{\lambda}$ and $\hat{\eta}$ are the previous values of λ and η. We then update the parameters as $\lambda = \hat{\lambda} + \alpha\Delta_\lambda$ and $\eta = \hat{\eta} + \alpha\Delta_\eta$, where, in order to insure that $\lambda \in [0,1]^n$, we have set $\alpha := \min(1, 0.5\max\{\alpha > 0 : \hat{\lambda} + \alpha\Delta_\lambda \in [0,1]\})$. In our experiments below we choose $R = 5$, $\overline{\nu} = 10$ and $A = 0.5$.

In both experiments we tried to learn a target function $f : [0, 2\pi] \to \mathbb{R}$ from a set of its samples. In the first experiment we fixed $f(x) = \frac{1}{10}(x + 2(e^{-8(\frac{4}{3}\pi - x)^2} - e^{-8(\frac{\pi}{2} - x)^2} - e^{-8(\frac{3}{2}\pi - x)^2}))$, $x \in [0, 2\pi]$, and,

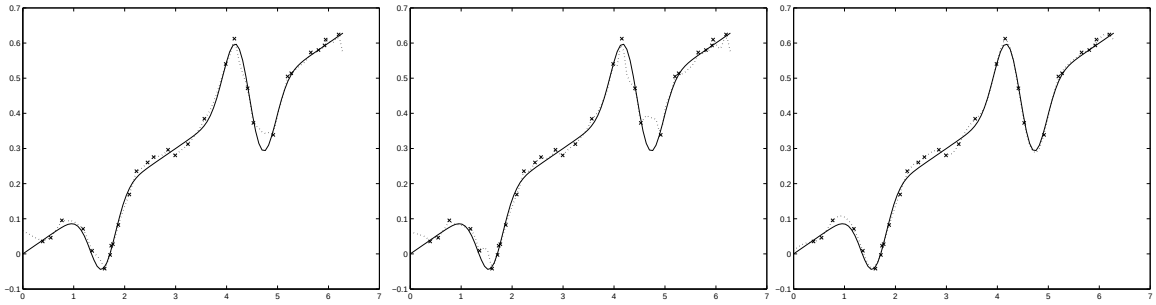

Figure 1: *Experiment 1: function learned by method 1 (left), method 2 (center) and method 3 (right). Regularization parameter is* $\mu = 0.1$*, the number of training points is* 50*. Solid line is the target function, crosses are the sampled points and the dotted line is the method used. The vertical scale has been reduce*

μ	10^{-4}	10^{-3}	10^{-2}	0.1	1	10
Method 1	3.46 (1.39)	3.46(1.39)	3.45 (1.38)	3.35 (1.35)	**2.64 (1.10)**	14.1 (10.3)
Method 2	4.46 (1.82)	4.46 (1.79)	3.85 (1.18)	**3.78 (1.03)**	4.00 (1.02)	62.6 (5.11)
Method 3	0.52 (0.56)	0.51 (0.56)	**0.51 (0.55)**	0.51 (0.57)	0.53 (0.63)	3.51 (1.47)

Table 2: *Experiment 2: Average mean square error with its standard deviation (in parenthesis) for methods 1 to 3 for different values of the regularization parameter* μ *(see text for the description). The unit measure for the errors is* 10^{-3}*.*

for every $x,t \in [0,2\pi]$, we set $K_\ell(x,t) = (xt)^{\ell-1}$ if $\ell \in \{1,2,3\}$ and $K_\ell(x,t) = e^{-\omega_\ell (x-t)^2}$ if $\ell \in \{4,5,6\}$ where $\omega_\ell = 2^{8-5(\ell-4)}$. We generated a training set of fifty points $\{(x_j,y_j) : j \in \mathbb{N}_{50}\} \subset [0,2\pi] \times \mathbb{R}$ obtained by sampling f with noise. Specifically, we choose x_j uniformly distributed in the interval $[0,2\pi]$ and $y_j = f(x_j) + \varepsilon$ with ε also uniformly sampled in the interval $[-0.02, 0.02]$. We then computed on a test set of 100 samples the mean square error between the target function f and the function learned from the training set for different values of the parameter μ. We compare three methods. *Method 1* is our proposed approach, *method 2* is the average of the kernels, that is we use the kernel $K = \frac{1}{n}\sum K_\ell$ and *method 3* is the kernel $K = K_2 + K_5$, the "ideal" kernel, that is, the kernel used to generate the target function. The results are shown in Table 1. Figure 1 shows the function learned by each method.

In our second experiment we fixed $f(x) = sin(x) + \frac{1}{2} sin(3x)$, $x \in [0,2\pi]$ and $K_\ell(x,t) = \sin(\ell x)\sin(\ell t)$, $x,t \in [0,2\pi]$, $\ell \in \mathbb{N}_n$. The set up is similar to that in Experiment 1. *Method 1* is our proposed approach, *method 2* is the average of the kernels and *method 3* is the ideal kernel given by $K(x,t) = \frac{2}{3}\sin(x)\sin(t) + \frac{1}{3}\sin(3x)\sin(3t)$. The noise ε is now uniformly sampled in the interval $[-0.2, 0.2]$. The results are reported in Table 2. Figure 2 shows the function learned by each method.

4.3 Extensions

We discuss some extensions of the problems studied in this paper. The first one that comes to mind is obtained by taking the expectation of the functional (4) with respect to a probability measure P on $\mathbb{R}^m$, that is,

$$Q_\mu^{av}(K) := \int_{\mathbb{R}^m} Q_\mu(K,y) P(y) dy, \quad K \in \mathcal{K} \tag{32}$$

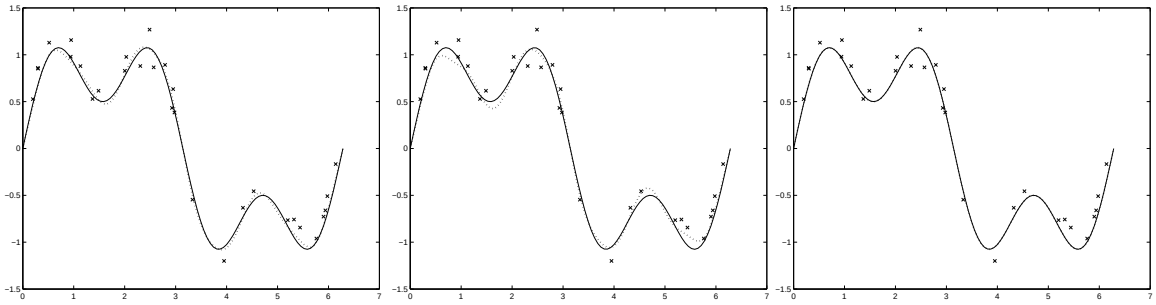

Figure 2: *Experiment 2: function learned by method 1 (left), method 2 (center) and method 3 (right). Regularization parameter is* $\mu = 0.1$*, the number of training points is* 50*. Solid line is the target function, crosses are the sampled points and the dotted line is the method used.*

where we indicated the dependency of $Q_\mu(K)$ on y by writing $Q_\mu(K,y)$. Since $Q_\mu(K,y)$ is convex in K for each $y \in \mathbb{R}^m$ so is $Q_\mu^{av}(K)$. We then minimize $Q_\mu^{av}(K)$ over $K \in \mathcal{K}$. For the square loss regularization we obtain that

$$S_\mu^{av}(K) = \mu \, trace((K_{\mathbf{x}} + \mu I)^{-1}\Sigma) \tag{33}$$

where Σ is the correlation matrix of P. Minimizing the quantity (33) over a convex class $\mathcal{K}$ may be valuable in image reconstruction and compression where we are provided with a collection of images and we wish to find a good average representation for them. In this case the input $\mathbf{x} = \{x_i : i \in \mathbb{N}_m\}$ represents the locations of the image pixels. For gray level images we can assume that $y \in [0,1]^m$ and therefore we should choose P to have support on $[0,1]^m$. Thus, if $\{y^\ell : \ell \in \mathbb{N}_n\}$ is a sample of such images with $n < m$ and Σ is the rank n empirical correlation matrix our goal is to find a kernel which well-represents this collection on the average.

Another approach is provided by replacing the average in equation (32) with the maximum over all y with bounded norm, that is, we minimize the functional

$$Q_\mu^{max}(K) := \max\{Q_\mu(K,y) : \|y\| \le 1\}, \quad K \in \mathcal{K}.$$

Again, this function is convex in K. In particular, for square loss regularization and the Euclidean norm on $\mathbb{R}^m$ we obtain

$$\max\{S_\mu(K,y) : \|y\| \le 1\} = \max\{\mu(y,(K_{\mathbf{x}} + \mu I)^{-1}y) : \|y\| \le 1\} = \frac{\mu}{\lambda_{min}(K_{\mathbf{x}}) + \mu}$$

where $\lambda_{min}(K_{\mathbf{x}})$ is the smallest eigenvalue of the matrix $K_{\mathbf{x}}$. Consequently, we have that

$$\min\{\max\{S_\mu(K,y) : \|y| \le 1\} : K \in \mathcal{K}\} = \frac{\mu}{\max\{\lambda_{min}(K_{\mathbf{x}}) : K \in \mathcal{K}\} + \mu}.$$

It is well-known that $\lambda_{min}(K_{\mathbf{x}})$ is a concave function of $K_{\mathbf{x}}$, (see, for example, Marshall and Olkin, 1979, p. 475). Therefore, our results provide an alternate proof of this fact.

We also remark that instead of learning a function f from function values the information operator I can be of the form $I(f) = ((g_j, f) : j \in \mathbb{N}_m)$, $f \in \mathcal{H}$, where $\{g_j : j \in \mathbb{N}_m\}$ is a set of prescribed functions in a Hilbert space, see the work of Micchelli and Pontil (2004) for a discussion. In this

case, the matrix $K_{\mathbf{x}}$ becomes the Gram matrix of these functions. The previous sections considered the choice $g_j = K(x_j, \cdot)$ and the Gram matrix is $K_{\mathbf{x}}$. This extension has wide applications in inverse problems, for example for computing the solution of first order integral equations.

Lemma 17 indicates that $Q_\mu : \mathcal{A}_+(\mathcal{X}) \to \mathbb{R}_+$ is, generally, not strictly convex. We may modify the functional Q_μ with a penalty term which depends on the kernel matrix $K_{\mathbf{x}}$ to enforce uniqueness of the optimal kernel in $\mathcal{K}$. Therefore, we consider the variational problem

$$\min\{Q_\mu(K) + R(K_{\mathbf{x}})\} \tag{34}$$

where R is a strictly convex function on $\mathcal{L}(\mathbb{R}^m)$. In this case, the method of proof of Theorem 7 shows that the optimal kernel can be found as a convex combination of at most $\frac{1}{2}m(m+1)$ kernels. For example, we may choose $R(A) = trace(A^2)$, $A \in \mathcal{L}(\mathbb{R}^m)$.

The variational problem (34) may be a preferred approach for choosing an optimal kernel. Indeed, if Q vanishes at some point in $\mathbb{R}^m$ and there is a kernel $K \in \mathcal{K}$ such that for all $t > 0$, $tK \in \mathcal{K}$ then it follows that $Q_\mu(\mathcal{K}) = 0$. This fact follows since $\lim_{t\to\infty} Q_\mu(tK) = 0$, by elementary properties of the norm in $\mathcal{H}_{tK}$. However, if the kernels in $\mathcal{K}$ have the property that $\sup_{K\in\mathcal{G}} \sup_{x\in\mathcal{X}} K(x,x) < \infty$, that is, they are uniformly bounded, the above circumstance cannot occur. This observation suggests that our criterion may be free from overfitting. Preliminary experiments with Gaussian kernels confirm that overfitting does not occur (Argyriou, Micchelli and Pontil, 2005). We leave for a future occasion a detailed investigation of this important issue.

As a final comment, let us point out that a kernel map can also be parameterized by matrices. For example, to each $A \in \mathcal{L}(\mathbb{R}^d)$ we define the linear kernel $K_A(x,t) = (x, At)$, $x, t \in \mathbb{R}^d$ and so our results apply to any convex compact subset of $\mathcal{L}(\mathbb{R}^d)$ for this kernel map. Another example are Gaussians parameterized by covariances $\Sigma \in \mathcal{L}(\mathbb{R}^d)$, that is,

$$N(\Sigma)(x,t) = \frac{1}{\sqrt{det(\Sigma)(2\pi)^d}} e^{-(x-t,\Sigma^{-1}(x-t))}, \quad x,t \in \mathbb{R}^d.$$

For compact convex sets of covariances our results say that Gaussian mixture models give optimal kernels.

5. Conclusion

The intent of this paper is to enlarge the theoretical understanding of the study of optimal kernels via the minimization of a regularization functional. Our analysis of this problem builds upon and extends the work of Lanckriet et al. (2004) and Lin and Zhang (2003). In contrast to the point of view of these papers, our setting applies to convex combinations of kernels parameterized by a compact set. Our analysis establishes that the regularization functional Q_μ is convex in K and that any optimizing kernel can be expressed as the convex combination of at most $m+2$ basic kernels. We have also provided a detailed characterization of the resulting minimax problem for square loss regularization. We have only marginally addressed at this stage implementation and algorithms for the search of optimal kernels. Since the proofs provided in Theorems 19 and 20 are constructive it should be possible to make use of them to derive practical algorithms for learning an optimal kernel such as a mixture of Gaussians, see (Argyriou, Micchelli and Pontil, 2005) for some recent results in this direction. Finally, an important direction which has not been explored in this paper is that of deriving error bounds, see (Micchelli et al., 2005) for some very recent progress on this.

Acknowledgments

We are grateful to Mark Herbster of University College London (UCL) for a remark which lead to Lemma 25, Raphael Hauser of Oxford University for suggesting a method to minimize the square loss regularization functional and many useful observations and to Andreas Argyriou of UCL for many useful comments. We also wish to thank Cheng Soon Ong of the Australian National University for discussions on his work which relates to ours, and are grateful to the referees for helping us clarify our presentation.

This work was partially supported by NSF Grant Number ITR-0312113, EPSRC Grant Number GR/T18707/01 and by the IST Programme of the European Community, under the PASCAL Network of Excellence IST-2002-506778.

Appendix A

The first result we record here is a useful version of the classical von Neumann minimax theorem.

Theorem 22 *Let $f : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ where $\mathcal{A}$ is a compact convex subset of a Hausdorff topological vector space $\mathcal{X}$ and $\mathcal{B}$ is a convex subset of a vector space $\mathcal{Y}$. If the function $x \mapsto f(x,y)$ is convex and lower semi-continuous for every $y \in \mathcal{B}$ and $y \mapsto f(x,y)$ is concave for every $x \in \mathcal{A}$ then we have that*

$$\min\{\sup\{f(x,y) : y \in \mathcal{B}\} : x \in \mathcal{A}\} = \sup\{\inf\{f(x,y) : x \in \mathcal{A}\} : y \in \mathcal{B}\} \tag{35}$$

Theorem 23 *Let $f : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ where $\mathcal{A}$ is a closed convex subset of a Hausdorff topological vector space $\mathcal{X}$ and $\mathcal{B}$ is a convex subset of a vector space $\mathcal{Y}$. If the function $x \mapsto f(x,y)$ is convex and lower semi-continuous for every $y \in \mathcal{B}$, $y \mapsto f(x,y)$ is concave for every $x \in \mathcal{A}$ and there exists a $y_0 \in \mathcal{B}$ such that for all $\lambda \in \mathbb{R}$ the set*

$$\{x : x \in \mathcal{A}, f(x,y_0) \leq \lambda\}$$

is compact then there is an $x_0 \in \mathcal{A}$ such that

$$\sup\{f(x_0,y) : y \in \mathcal{B}\} = \sup\{\inf\{f(x,y) : x \in \mathcal{A}\} : y \in \mathcal{B}\}$$

in particular, (35) holds

Theorem 22 is subsumed by Theorem 23 whose proof can be found in (Aubin, 1982, Ch. 7). The hypothesis of lower semi-continuity means, for all $\lambda \in \mathbb{R}$ and $y \in \mathcal{B}$, that the set $\{x : x \in \mathcal{A}, f(x,y) \leq \lambda\}$ is a closed subset of $\mathcal{A}$.

The next result concerns differentiation of a "max" function. The version we use comes from (Micchelli, 1969). Let $\mathcal{X}$ be a topological vector space. If g is a continuous real-valued function on $\mathcal{X}$, we define its right derivative at $x \in \mathcal{X}$ in the direction $y \in \mathcal{X}$ by the formula

$$g'_+(x,y) = \lim_{\varepsilon \to 0^+} \frac{g(x+\varepsilon y) - g(x)}{\varepsilon}$$

whenever it exists.

Lemma 24 *Let $\mathcal{T}$ a compact set and $G(t,x)$ a real-valued function on $\mathcal{T}\times\mathcal{X}$ such that, for every $x\in\mathcal{X}$ $G(\cdot,x)$ is continuous on $\mathcal{T}$ and, for every $t\in\mathcal{T}$, $G(t,\cdot)$ is convex on $\mathcal{X}$. We define the real-valued convex function g on $\mathcal{X}$ by the formula*

$$g(x) := \max\{G(t,x) : t\in\mathcal{T}\},\ \ x\in\mathcal{X}$$

and the set

$$M(x) := \{t : t\in\mathcal{T}, G(t,x) = g(x)\}.$$

Then the right derivative of g in the direction $y\in\mathcal{X}$ is given by

$$g'_+(x,y) = \max\{G'_+(t,x,y) : t\in M(x)\}$$

where $G'_+(t,x,y)$ is the right derivative of G with respect to its second argument in the direction y.

PROOF. We first observe, for every $t\in M(x)$ and $\lambda>0$, that

$$\frac{g(x+\lambda y)-g(x)}{\lambda} \geq \frac{G(t,x+\lambda y)-G(t,x)}{\lambda}$$

which, letting $\lambda\to 0^+$, implies that $g'_+(x,y)\geq G'_+(t,x,y)$ and, so,

$$g'_+(x,y)\geq \sup\{G'_+(t,x,y) : t\in M(x)\}.$$

To prove the reverse inequality we use the fact that if f is convex on $[0,\infty)$ and $f(0)=0$ then $f(\lambda)/\lambda$ is a nondecreasing function of $\lambda>0$. In particular, this is true for the function of λ defined, for every $x,y\in\mathcal{X}$, as

$$\frac{g(x+\lambda y)-g(x)}{\lambda}.$$

Consequently, we obtain, for every $\lambda>0$ that

$$\frac{g(x+\lambda y)-g(x)}{\lambda}\geq g_+(x,y).$$

Now, we define

$$h(\lambda,t) := \frac{G(t,x+\lambda y)-g(x)}{\lambda},\ \ \lambda>0$$

and observe that, for each $t\in\mathcal{T}$, it is a nondecreasing function of λ because

$$h(\lambda,t) = \frac{G(t,x+\lambda y)-G(t,x)}{\lambda} - \frac{g(x)-G(t,x)}{\lambda}.$$

Therefore, the sets $A_\lambda := \{t\in\mathcal{T} : h(\lambda,t)\geq g'_+(x,y)\}$ are nonempty, closed and nested for $\lambda>0$ and, so, the compactness of $\mathcal{T}$ implies that there exists a $t_0\in\bigcap_{\lambda>0}A_\lambda$, that is,

$$G(t_0,x+\lambda y)\geq \lambda g'_+(x,y)+g(x),\ \ \lambda>0.$$

Thus, $t_0\in M(x)$ and $g'_+(x,y)\leq G'_+(t_0,x,y)$. □

We now present the proof of Lemma 8 in an extended form. To this end, we let r be *any* positive number and let

$$co_r\mathcal{K}_n := \left\{K : K=\sum_{j\in\mathbb{N}_n}\lambda_j K_j,\ \lambda_\ell\geq 0, \ell\in\mathbb{N}_n,\ \sum_{j\in\mathbb{N}_n}\lambda_j^r = 1\right\}.$$

Note that $co_1\mathcal{K}_n = co\mathcal{K}_n$ where $\mathcal{K}_n = \{K_j : j\in\mathbb{N}_n\}$.

Lemma 25 *If $\mathcal{K}_n = \{K_j : j \in \mathbb{N}_n\}$ is a family of kernels on $X \times X$ and $f \in \bigoplus_{j\in\mathbb{N}_n} \mathcal{H}_{K_j}$, and $s := \frac{2r}{r+1}$ then*

$$\inf\{\|f\|_K : K \in co_r \mathcal{K}_n\} = \min\left\{\left(\sum_{j\in\mathbb{N}_n} \|f_j\|^s\right)^{\frac{1}{s}} : f = \sum_{j\in\mathbb{N}_n} f_j, f_\ell \in \mathcal{H}_{K_\ell}, \ell \in \mathbb{N}_n\right\}.$$

PROOF. The first step is to appeal to a result of Aronszajn, (see Aronszajn, 1950, p. 352-3), which states that for any $f \in \bigoplus_{j\in\mathbb{N}_n} \mathcal{H}_{K_j}$ we have for $K = \sum_{j\in\mathbb{N}_n} \lambda_j K_j$, with $\lambda_\ell > 0$, $\ell \in \mathbb{N}_n$ that

$$\|f\|_K^2 = \min\left\{\sum_{j\in\mathbb{N}_n} \frac{\|f_j\|^2}{\lambda_j} : f = \sum_{j\in\mathbb{N}_n} f_j, f_\ell \in \mathcal{H}_{K_\ell}, \ell \in \mathbb{N}_n\right\}.$$

Thus, the lemma follows from the following fact.

Lemma 26 *If $r > 0$, $p := 1 + \frac{1}{r}$, and $\{a_j : j \in \mathbb{N}_n\} \subset \mathbb{R}$ then*

$$\min\left\{\left(\sum_{j\in\mathbb{N}_n} \frac{a_j^2}{\lambda_j}\right)^{\frac{1}{2}} : \lambda_\ell \geq 0, \ell \in \mathbb{N}_n, \sum_{j\in\mathbb{N}_n} \lambda_j^r \leq 1\right\} = \left(\sum_{j\in\mathbb{N}_n} |a_j|^{\frac{2}{p}}\right)^{\frac{p}{2}}$$

and the equality occurs for $\sum_{j\in\mathbb{N}_n} |a_j| > 0$ at

$$\tilde{\lambda}_j := \frac{|a_j|^{\frac{2}{r+1}}}{\left(\sum_{j\in\mathbb{N}_n} |a_j|^{\frac{2r}{r+1}}\right)^{\frac{1}{r}}}. \tag{36}$$

PROOF. This fact follows from Hölder inequality. To this end, we let $q = r+1$ so that $\frac{1}{p} + \frac{1}{q} = 1$ and, so, we have that

$$\begin{aligned}
\sum_{j\in\mathbb{N}_n} |a_j|^{\frac{2r}{q}} &= \sum_{j\in\mathbb{N}_n} \frac{|a_j|^{\frac{2r}{q}}}{\lambda_j^{\frac{r}{q}}} \lambda_j^{\frac{r}{q}} \\
&\leq \left(\sum_{j\in\mathbb{N}_n} \frac{|a_j|^{\frac{2rp}{q}}}{\lambda_j^{\frac{rp}{q}}}\right)^{\frac{1}{p}} \left(\sum_{j\in\mathbb{N}_n} \lambda_j^r\right)^{\frac{1}{q}} \\
&= \left(\sum_{j\in\mathbb{N}_n} \frac{a_j^2}{\lambda_j}\right)^{\frac{1}{p}} \left(\sum_{j\in\mathbb{N}_n} \lambda_j^r\right)^{\frac{1}{q}} \leq \left(\sum_{j\in\mathbb{N}_n} \frac{a_j^2}{\lambda_j}\right)^{\frac{1}{p}}.
\end{aligned}$$

For the choice (36) equality holds above, thereby completing the proof. □

The proof of Lemma 25 is completed. □

References

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.

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

J. P. Aubin. *Mathematical methods of game and economic theory*. Studies in Mathematics and its applications, Vol. 7, North-Holland, 1982.

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.

F. R. Bach, R. Thibaux and M. I. Jordan. Computing regularization paths for learning multiple kernels. *Advances in Neural Information Processing Systems*, 17, 2004.

C. Bennett and R. Sharpley. *Interpolation of Operators*. Vol. 129, Pure and Appl. Math, Academic Press, Boston, 1988.

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

O. Bousquet and A. Elisseeff. Stability and generalization. *J. of Machine Learning Research*, 2: 499–526, 2002.

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

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1): 131–159, 2002.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39 (1): 1–49, 2002.

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

E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, A. Verri. Some properties of regularized kernel methods. *J. of Machine Learning Research*, 5(Oct):1363–1390, 2004.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13: 1–50, 2000.

F. Girosi. An Equivalence Between Sparse Approximation and Support Vector Machines. *Neural Computation*, 10 (6): 1455–1480, 1998.

T. Graepel. Kernel matrix completion by semi-definite programming. Proc. of ICANN, pages 694–699, 2002.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2002.

T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for support vector machines, *J. of Machine Learning Research*, 5, 1391–1415, 2004.

M. Herbster. Learning Additive Models Online with Fast Evaluating Kernels. *Proc. of the The 14-th Annual Conference on Computational Learning Theory (COLT)*, pages 444–460, 2001.

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.

T. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. MIT AI-Lab Technical Report, 1999.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, 33: 82–95, 1971.

G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, M. I. Jordan. Learning the kernel matrix with semi-definite programming. In C. Sammut and A.Hoffmann (Eds.), Proc. of the 19-th Int. Conf. on Machine Learning, Sydney, Australia, Morgann Kaufmann, 2002.

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.

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.

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.

O. L. Mangasarian. *Nonlinear Programming*. Classics in Applied Mathematics, SIAM, 1994.

A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and its Applications*. Academic Press, San Diego, 1979.

C. A. Micchelli. *Saturation Classes and Iterates of Operators.* PhD Thesis, Stanford University, 1969.

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

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

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.

S. Mukherjee, P. Niyogi, T. Poggio, R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for empirical risk minimization. *Advances in Computational Mathematics*, to appear, 2004.

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.

M. Pontil and A. Verri. Properties of support vector machines. *Neural Computation*, 10: 955–974, 1998.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.

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

I. J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39(4): 811–841, 1938.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, USA, 2002.

C. Scovel and I. Steinwart. Fast rates for support vector machines. Preprint, 2004.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

S. Smale, and D. X. Zhou. Estimating the approximation error in learning theory, *Anal. Appl.*, 1: 1–25, 2003.

D. M. J. Tax and R. P. W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20: 1191–1199, 1999.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

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

C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. Advances in Neural Processing Systems 8: 598–604, D. S. Touretzky, M. C. Mozer, M. E. Hasselmo (eds.), MIT Press, Cambridge, MA, 1996.

Q. Wu, Y. Ying and D. X. Zhou. Multi-kernel regularization classifiers. *Preprint*, City University of Hong Kong, 2004.

Y. M. Ying and D. X. Zhou. Learnability of Gaussians with flexible variances. Preprint, City University of Hong Kong, 2004.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statis.*, 32: 56–85, 2004.

T. Zhang. On the dual formulation of regularized linear systems with convex risks. *Machine Learning*, 46: 91–129, 2002.

Z. Zhang, D.-Y. Yeung and J. T. Kwok. Bayesian inference for transductive learning of kernel matrix using the Tanner-Wong data augmentation algorithm. *Proc. 21-st Int. Conf. Machine Learning (ICML-2004)*, pages 935-942, Banff, Alberta, Canada, July 2004.

D. X. Zhou. The covering number in learning theory. *J. of Complexity*, 18: 739–767, 2002.