Learning Convex Combinations of Continuously Parameterized Basic Kernels*

Andreas Argyriou[1], Charles A. Micchelli[2], and Massimiliano Pontil[1]

[1] Department of Computer Science
University College London
Gower Street, London WC1E 6BT, England, UK
{a.argyriou,m.pontil}@cs.ucl.ac.uk
[2] Department of Mathematics and Statistics
State University of New York, The University at Albany
1400 Washington Avenue, Albany, NY, 12222, USA

Abstract. We study the problem of learning a kernel which minimizes a regularization error functional such as that used in regularization networks or support vector machines. We consider this problem when the kernel is in the convex hull of basic kernels, for example, Gaussian kernels which are continuously parameterized by a compact set. We show that there always exists an optimal kernel which is the convex combination of at most $m+1$ basic kernels, where m is the sample size, and provide a necessary and sufficient condition for a kernel to be optimal. The proof of our results is constructive and leads to a greedy algorithm for learning the kernel. We discuss the properties of this algorithm and present some preliminary numerical simulations.

1 Introduction

A common theme in machine learning is that a function can be learned from a finite set of input/output examples by minimizing a regularization functional which models a trade-off between an error term, measuring the fit to the data, and a smoothness term, measuring the function complexity. In this paper we focus on learning methods which, given examples $\{(x_j, y_j) : j \in \mathbb{N}_m\} \subseteq \mathcal{X} \times \mathbb{R}$, estimate a real-valued function by minimizing the regularization functional

$$Q_\mu(f, K) = \sum_{j \in \mathbb{N}_m} q(y_j, f(x_j)) + \mu \|f\|_K^2 \tag{1}$$

where $q : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a prescribed *loss function*, μ is a positive parameter and $\mathbb{N}_m := \{1, \ldots, m\}$. The minimum is taken over $f \in \mathcal{H}_K$, a *reproducing kernel Hilbert space* (RKHS) with kernel K, see [1].

This approach has a long history. It has been studied, from different perspectives, in statistics [16], in optimal recovery [10], and more recently, has been

* This work was supported by EPSRC Grant GR/T18707/01, NSF Grant ITR-0312113 and the PASCAL European Network of Excellence.

a focus of attention in machine learning theory, see, for example [14, 15] and references therein. The choice of the loss function q leads to different learning methods among which the most prominent have been square loss regularization and support vector machines.

As new parametric families of kernels are being proposed to model functions defined on possibly complex/structured input domains (see, for example, [14] for a review) it is increasingly important to develop optimization methods for tuning kernel-based learning algorithms over a possibly large number of kernel parameters. This motivates us to study the problem of minimizing functional (1) not only over f but also over K, that is, we consider the variational problem

$$Q_\mu(\mathcal{K}) = \inf\{Q_\mu(f, K) : f \in \mathcal{H}_K, K \in \mathcal{K}\} \tag{2}$$

where $\mathcal{K}$ is a prescribed *convex* set of kernels. This point of view was proposed in [3, 8] where the problem (2) was mainly studied in the case of support vector machines and when $\mathcal{K}$ is formed by combinations of a *finite* number of basic kernels. Other related work on this topic appears in the papers [4, 9, 11, 18].

In this paper, we present a framework which allows us to model richer families of kernels parameterized by a compact set Ω, that is, we consider kernels of the type

$$\mathcal{K} = \left\{ \int_\Omega G(\omega) dp(\omega) : p \in \mathcal{P}(\Omega) \right\} \tag{3}$$

where $\mathcal{P}(\Omega)$ is the set of all probability measures on Ω. For example, when $\Omega \subseteq \mathbb{R}_+$ and the function $G(\omega)$ is a multivariate Gaussian kernel with variance ω then $\mathcal{K}$ is a subset of the class of radial kernels. The set-up for the family of kernels in (3) is discussed in Section 2, where we also review some earlier results from [9]. In particular, we establish that if q is convex then problem (2) is equivalent to solving a saddle-point problem. In Section 3, we derive optimality conditions for problem (2). We present a necessary and sufficient condition which characterizes a solution to this problem (see Theorem 2) and show that there always exists an optimal kernel $\hat{K}$ with a *finite representation*. Specifically, for this kernel the probability measure p in (3) is an atomic measure with *at most* $m+1$ atoms (see Theorem 1). As we shall see, this implies, for example, that the optimal radial kernel is a finite mixture of Gaussian kernels when the variance is bounded above and away from zero. We mention, in passing, that a version of our characterization also holds when Ω is locally compact (see Theorem 4). The proof of our results is constructive and can be used to derive algorithms for learning the kernel. In Section 4, we propose a greedy algorithm for learning the kernel and present some preliminary experiments on optical character recognition.

2 Background and Notation

In this section we review our notation and present some background results from [9] concerning problem (2).

We begin by recalling the notion of a kernel and RKHS $\mathcal{H}_K$. Let $\mathcal{X}$ be a set. By a *kernel* we mean a symmetric 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. 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 $\mathcal{A}(\mathcal{X})$ for the set of all kernels on the set $\mathcal{X}$ and $\mathcal{A}_+(\mathcal{X})$ for the set of kernels K such that, for each $m \in \mathbb{N}$ and each choice of $\mathbf{x}$, $K_{\mathbf{x}} \in \mathcal{L}_+(\mathbb{R}^m)$.

According to Aronszajn and Moore [1], every kernel is associated with an (essentially) *unique* Hilbert space $\mathcal{H}_K$ with inner product $\langle \cdot, \cdot \rangle_K$ such that K is its reproducing kernel. This means that for every $f \in \mathcal{H}_K$ and $x \in \mathcal{X}$, $\langle f, K_x \rangle_K = f(x)$, where $K_x(\cdot) := K(x, \cdot)$. Equivalently, any Hilbert space $\mathcal{H}$ of real-valued functions defined everywhere on $\mathcal{X}$ such that the point evaluation functionals $L_x(f) := f(x),\ f \in \mathcal{H}$ are continuous on $\mathcal{H}$, admits a reproducing kernel K.

Let $D := \{(x_j, y_j) : j \in \mathbb{N}_m\} \subseteq \mathcal{X} \times \mathbb{R}$ 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 let $\mathbb{R}_+ := [0, \infty)$, 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 $\|f\|_K^2 := \langle f, f \rangle_K$ and μ is a positive constant. Note that Q depends on y but we suppress it in our notation as it is fixed throughout our discussion. For example, in equation (1) we have, for $w = (w_j : j \in N_n)$, that $Q(w) = \sum_{j \in \mathbb{N}_m} q(y_j, w_j)$, where q is a loss function.

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

$$Q_\mu(K) := \inf\{Q_\mu(f, K) : f \in \mathcal{H}_K\} \tag{5}$$

which defines a functional $Q_\mu : \mathcal{A}(\mathcal{X}) \to \mathbb{R}_+$. We remark, in passing, that all of what we say about problem (5) applies to functions Q on $\mathbb{R}^m$ which are bounded from below as we can merely adjust the expression (4) by a constant independent of f and K. Note that if $Q : \mathbb{R}^m \to \mathbb{R}_+$ is continuous and μ is a positive number then the infimum in (5) is achieved because the unit ball in $\mathcal{H}_K$ is *weakly* compact. In particular, when Q is convex the minimum is unique since in this case the right hand side of equation (4) is a *strictly* convex functional of $f \in \mathcal{H}_K$. Moreover, if f is a solution to problem (5) then it has the form

$$f(x) = \sum_{j \in \mathbb{N}_m} c_j K(x_j, x), \quad x \in \mathcal{X} \tag{6}$$

for some real vector $c = (c_j : j \in \mathbb{N}_m)$. This result is known as the *Representer Theorem*, see, for example, [14]. Although it is simple to prove, this result is remarkable as it makes the variational problem (5) amenable to computations. In particular, if Q is convex, the unique minimizer of problem (5) can be found by replacing f by the right hand side of equation (6) in equation (4) and then optimizing with respect to the vector c. That is, we have the finite dimensional variational problem

$$Q_\mu(K) := \min\{Q(K_{\mathbf{x}} c) + \mu(c, K_{\mathbf{x}} c) : c \in \mathbb{R}^m\} \tag{7}$$

where $(\cdot,\cdot)$ is the standard inner product on $\mathbb{R}^m$. For example, when Q is the square loss defined for $w = (w_j : j \in \mathbb{N}_m) \in \mathbb{R}^m$ as $Q(w) = \|w - y\|^2 := \sum_{j\in\mathbb{N}_m}(w_j - y_j)^2$ the function in the right hand side of (7) is quadratic in the vector c and its minimizer is obtained by solving a linear system of equations.

The point of view of this paper is that the functional (5) can be used as a *design criterion to select the kernel* K. To this end, we specify an arbitrary convex subset $\mathcal{K}$ of $\mathcal{A}(\mathcal{X})$ and focus on the problem

$$Q_\mu(\mathcal{K}) := \inf\{Q_\mu(K) : K \in \mathcal{K}\}. \tag{8}$$

Every input set $\mathbf{x}$ and convex set $\mathcal{K}$ of kernels determines a convex set of matrices in $\mathcal{L}(\mathbb{R}^m)$, namely $\mathcal{K}(\mathbf{x}) := \{K_\mathbf{x} : K \in \mathcal{K}\}$. Obviously, it is this set of matrices that affects the variational problem (8). For this reason, we say that the set of kernels $\mathcal{K}$ is *compact and convex* provided that for all $\mathbf{x}$ the set of matrices $\mathcal{K}(\mathbf{x})$ is compact and convex. The following result is taken directly from [9].

Lemma 1. *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 (8) exists.*

The 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

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

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

There are two useful cases of the 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 case generalizes the above one to a compact Hausdorff space Ω (see, for example, [12]) 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 $\mathcal{P}(\Omega)$ be the set of all *probability measures* on Ω and observe that

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

is a compact and convex set of kernels in $\mathcal{A}_+(\mathcal{X})$. The compactness of this set is a consequence of the weak*-compactness of the unit ball in the dual space of $C(\Omega)$, the set of all continuous real-valued functions g on Ω with norm $\|g\|_\Omega := \max\{|g(\omega)| : \omega \in \Omega\}$, see [12]. For example, we choose $\Omega = [\omega_1, \omega_2]$, where $0 < \omega_1 < \omega_2$ and $G(\omega)(x,t) = e^{-\omega\|x-t\|^2}$, $x, t \in \mathbb{R}^d$, $\omega \in \Omega$, to obtain *radial kernels*, or $G(\omega)(x,t) = e^{\omega(x,t)}$ to obtain *dot product kernels*. Note that the choice $\Omega = \mathbb{N}_n$ corresponds to the first case.

Next, we establish that if the loss function $Q : \mathbb{R}^m \to \mathbb{R}$ is convex then the functional $Q_\mu : \mathcal{A}_+(\mathcal{X}) \to \mathbb{R}_+$ is convex as well, that is, the variational problem

(8) is a *convex optimization problem*. To this end, we recall that the conjugate function of Q, denoted by $Q^* : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, is defined, for every $v \in \mathbb{R}^m$, as

$$Q^*(v) = \sup\{(w, v) - Q(w) : w \in \mathbb{R}^m\} \tag{10}$$

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

$$Q(w) = \sup\{(w, v) - Q^*(v) : v \in \mathbb{R}^m\} \tag{11}$$

see [5]. See also [17] for a nice application of the conjugate function to linear statistical models. For example, for the square loss defined above, the conjugate function is given, for every $v \in \mathbb{R}^m$, by

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

Note that $Q^*(0) = -\inf\{Q(w) : w \in \mathbb{R}^m\} < \infty$ since Q is bounded from below. This observation is used in the proof of the lemma below.

Lemma 2. *If $K \in \mathcal{A}(\mathcal{X})$, $\mathbf{x}$ is a set of m 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*

$$Q_\mu(K) = \max\left\{-\frac{1}{4\mu}(v, K_{\mathbf{x}} v) - Q^*(v) : v \in \mathbb{R}^m\right\}. \tag{12}$$

The fact that the maximum above exists follows from the hypothesis that $K_{\mathbf{x}} \in \mathcal{L}_+(\mathbb{R}^m)$ and the fact that $Q^*(v) \geq -Q(0)$ for all $v \in \mathbb{R}^m$, which follows from equation (10). The proof of the lemma is based on a version of the von Neumann minimax theorem (see the appendix). This lemma implies that the functional $Q_\mu : \mathcal{A}_+(\mathcal{X}) \to \mathbb{R}_+$ is convex. Indeed, equation (12) expresses $Q_\mu(K)$ as the maximum of linear functions in the kernel K.

3 Characterization of an Optimal Kernel

Our discussion in Section 2 establishes that problem (8) reduces to the minimax problem

$$Q_\mu(\mathcal{K}) = -\max\{\min\{R(c, K) : c \in \mathbb{R}^m\} : K \in \mathcal{K}\} \tag{13}$$

where the function R is defined as

$$R(c, K) = \frac{1}{4\mu}(c, K_{\mathbf{x}} c) + Q^*(c), \quad c \in \mathbb{R}^m,\ K \in \mathcal{K}. \tag{14}$$

In this section we show that problem (13) admits a saddle point, that is, the minimum and maximum in (13) can be interchanged and describe the properties of this saddle point. We consider this problem in the general case that $\mathcal{K}$ is induced by a continuous mapping $G : \Omega \to \mathcal{A}_+(\mathcal{X})$ where Ω is a compact Hausdorff space, so we write $\mathcal{K}$ as $\mathcal{K}(G)$, see equation (9).

We assume that the conjugate function is differentiable everywhere and denote the gradient of Q^* at c by $\nabla Q^*(c)$.

Theorem 1. *If Ω is a compact Hausdorff topological space and $G : \Omega \to \mathcal{A}_+(\mathcal{X})$ 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 on Ω with at most $m+1$ atoms and, for any atom $\hat{\omega} \in \Omega$ of $\hat{p}$, we have that*

$$R(\hat{c}, G(\hat{\omega})) = \max\{R(\hat{c}, G(\omega)) : \omega \in \Omega\} \tag{15}$$

where $\hat{c}$ is the unique solution to the equation

$$\frac{1}{2\mu}\hat{K}_{\mathbf{x}}\hat{c} + \nabla Q^*(\hat{c}) = 0. \tag{16}$$

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

$$R(\hat{c}, K) \leq R(\hat{c}, \hat{K}) \leq R(c, \hat{K}). \tag{17}$$

Proof. Let us first comment on the nonlinear equation (16). For *any* kernel $K \in \mathcal{K}(G)$ the extremal problem

$$\min\{R(c, K) : c \in \mathbb{R}^m\}$$

has a *unique* solution, since the function $c \mapsto R(c, K)$ is strictly convex and $\lim_{\|c\| \to \infty} R(c, K) = \infty$. Moreover, if we let $c_K \in \mathbb{R}^m$ be the unique minimizer, it solves the equation

$$\frac{1}{2\mu}K_{\mathbf{x}}c_K + \nabla Q^*(c_K) = 0.$$

Hence, equation (16) says that $\hat{c} = c_{\hat{K}}$.

Now let us turn to the existence of the kernel $\hat{K}$ described above. First, we note the immediate fact that

$$\max\{R(c, K) : K \in \mathcal{K}(G)\} = \max\{R(c, G(\omega)) : \omega \in \Omega\}.$$

Next, we define the function $\varphi : \mathbb{R}^m \to \mathbb{R}$ by

$$\varphi(c) := \max\{R(c, G(\omega)) : \omega \in \Omega\}, \quad c \in \mathbb{R}^m.$$

According to the definition of the conjugate function in equation (10) and the hypotheses that G is continuous and $\{G(\omega) : \omega \in \Omega\} \subseteq \mathcal{A}_+(\mathcal{X})$ we see that $\lim_{\|c\| \to \infty} \varphi(c) = \infty$. Hence, φ has a minimum. We call a minimizer $\tilde{c}$. This vector is characterized by the fact that the right directional derivative of φ at $\tilde{c}$ in all directions $d \in \mathbb{R}^m$ is nonnegative. We denote this derivative by $\varphi'_+(\tilde{c}; d)$. Using Lemma 4 in the appendix, we have that

$$\varphi'_+(\tilde{c}; d) = \max\left\{ \frac{1}{2\mu}(d, G_{\mathbf{x}}(\omega)\tilde{c}) + (\nabla Q^*(\tilde{c}), d) : \omega \in \Omega^* \right\}$$

where the set Ω^* is defined as

$$\Omega^* := \{\omega : \omega \in \Omega,\ R(\tilde{c}, G(\omega)) = \varphi(\tilde{c})\}.$$

If we define the vectors $z(\omega) = \frac{1}{2\mu} G_{\mathbf{x}}(\omega)\tilde{c} + \nabla Q^*(\tilde{c})$, $\omega \in \Omega^*$, the condition that $\varphi'_+(\tilde{c}; d)$ is nonnegative for all $d \in \mathbb{R}^m$ means that

$$\max\{(z(\omega), d) : \omega \in \Omega^*\} \geq 0, \quad d \in \mathbb{R}^m.$$

Since G is continuous, the set $\mathcal{N} := \{z(\omega) : \omega \in \Omega^*\}$ is a closed subset of $\mathbb{R}^m$. Therefore, its convex hull $\mathcal{M} := co(\mathcal{N})$ is closed as well. We claim that $0 \in \mathcal{M}$. Indeed, if $0 \notin \mathcal{M}$ then there exists a hyperplane $\{c : c \in \mathbb{R}^m, (w, c) + \alpha = 0\}$, $\alpha \in \mathbb{R}$, $w \in \mathbb{R}^m$, which strictly separates 0 from $\mathcal{M}$, that is, $(w, 0) + \alpha > 0$ and $(w, z(\omega)) + \alpha \leq 0$, $\omega \in \Omega^*$, see [12]. The first condition implies that $\alpha > 0$ and, so we conclude that

$$\max\{(w, z(\omega)) : \omega \in \Omega^*\} < 0$$

which contradicts our hypothesis that $\tilde{c}$ is a minimum of φ.

By the Caratheodory theorem, see, for example, [5, Ch. 2], every vector in $\mathcal{M}$ can be expressed as a convex combination of at most $m+1$ of the vectors in $\mathcal{N}$. In particular, we have that

$$0 = \sum_{j \in \mathbb{N}_{m+1}} \lambda_j z(\omega_j) \tag{18}$$

for some $\{\omega_j : j \in \mathbb{N}_{m+1}\} \subseteq \Omega^*$ and nonnegative constants λ_j with $\sum_{j \in \mathbb{N}_{m+1}} \lambda_j = 1$. Setting

$$\hat{K} := \sum_{j \in \mathbb{N}_{m+1}} \lambda_j G(\omega_j) = \int_{\Omega} G(\omega) d\hat{p}(\omega)$$

where $\hat{p} = \sum_{j \in \mathbb{N}_{m+1}} \lambda_j \delta_{\omega_j}$, (we denote by δ_ω the Dirac measure at ω), we can rewrite equation (18) as

$$\frac{1}{2\mu} \hat{K}_{\mathbf{x}} \tilde{c} + \nabla Q^*(\tilde{c}) = 0.$$

Hence, we conclude that $\tilde{c} = \hat{c}$ which means that

$$\min\{R(c, \hat{K}) : c \in \mathbb{R}^m\} = R(\hat{c}, \hat{K}).$$

This establishes the upper inequality in (17). For the lower inequality we observe that

$$R(\hat{c}, \hat{K}) = \int_{\Omega} R(\hat{c}, G(\omega)) d\hat{p}(\omega) = \sum_{j \in \mathbb{N}_{m+1}} \lambda_j R(\hat{c}, G(\omega_j)).$$

Since $\hat{c} = \tilde{c}$, we can use the definition of the ω_j to conclude for any $K \in \mathcal{K}(G)$ by equation (18) and the definition of the function φ that

$$R(\hat{c}, \hat{K}) = \varphi(\hat{c}) \geq R(\hat{c}, K).$$

□

This theorem improves upon our earlier results in [9] where only the square loss function was studied in detail. Generally, not all saddle points $(\hat{c}, \hat{K})$ of R satisfy the properties stated in Theorem 1. Indeed, a maximizing kernel may be

represented as $\hat{K} = \int_\Omega G(\omega) d\hat{p}(\omega)$ where $\hat{p}$ may contain more than $m+1$ atoms or even have uncountable support (note, though, that the proof above provides a procedure for finding a kernel which is the convex combination of at most $m+1$ kernels). With this caveat in mind, we show below that the conditions stated in Theorem 1 are necessary and sufficient.

Theorem 2. *Let $\hat{c} \in \mathbb{R}^m$ and $\hat{K} = \int_\Omega G(\omega) d\hat{p}(\omega)$, where $\hat{p}$ is a probability measure with support $\hat{\Omega} \subseteq \Omega$. The pair $(\hat{c}, \hat{K})$ is a saddle point of problem (13) if and only if $\hat{c}$ solves equation (16) and every $\hat{\omega} \in \hat{\Omega}$ satisfies equation (15).*

Proof. If $(\hat{c}, \hat{K})$ is a saddle point of (13) then $\hat{c}$ is the unique minimizer of the function $R(\cdot, \hat{K})$ and solves equation (16). Moreover, we have that

$$\int_{\hat{\Omega}} R(\hat{c}, G(\omega)) d\hat{p}(\omega) = R(\hat{c}, \hat{K}) = \max\{R(\hat{c}, G(\omega)) : \omega \in \Omega\}$$

implying that equation (15) holds true for every $\hat{\omega} \in \hat{\Omega}$.

On the other hand, if $\hat{c}$ solves equation (16) we obtain the upper inequality in equation (17) whereas equation (15) brings the lower inequality. □

Theorem 1 can be specified to the case that Ω is a finite set, that is $\mathcal{K} = co(\mathcal{K}_n)$ where $\mathcal{K}_n = \{K_\ell : \ell \in \mathbb{N}_n\}$ is a prescribed set of kernels. Below, we use the notation $K_{\mathbf{x},\ell}$ for the matrix $(K_\ell)_{\mathbf{x}}$.

Corollary 1. *If $\mathcal{K}_n = \{K_j : j \in \mathbb{N}_n\} \subset \mathcal{A}_+(\mathcal{X})$ there exists a kernel $\hat{K} = \sum_{j \in \mathbb{N}_n} \lambda_j K_j \in co(\mathcal{K}_n)$ such that the set $J = \{j : j \in \mathbb{N}_n, \lambda_j > 0\}$ contains at most $min(m+1, n)$ elements and, for every $j \in J$ we have that*

$$R(\hat{c}, K_j) = \max\{(R(\hat{c}, K_\ell) : \ell \in \mathbb{N}_n\} \tag{19}$$

where $\hat{c}$ is the unique solution to the equation

$$\frac{1}{2\mu} \hat{K}_{\mathbf{x}} \hat{c} + \nabla Q^*(\hat{c}) = 0. \tag{20}$$

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

$$R(\hat{c}, K) \leq R(\hat{c}, \hat{K}) \leq R(c, \hat{K}). \tag{21}$$

In the important case that $\Omega = [\omega_1, \omega_2]$ for $0 < \omega_1 < \omega_2$ and $G(\omega)$ is a Gaussian kernel, $G(\omega)(x,t) = \exp(-\omega \|x-t\|^2), x, t \in \mathbb{R}^d$, Theorem 1 establishes that a mixture of at most $m+1$ Gaussian kernels provides an optimal kernel. What happens if we consider all possible Gaussians, that is, take $\Omega = \mathbb{R}_+$? This question is important because Gaussians generate the *whole class of radial kernels*. Indeed, we recall a beautiful result by I.J. Schoenberg [13].

Theorem 3. *Let h be a real-valued function defined on $\mathbb{R}_+$ such that $h(0) = 1$. We form a kernel K on $\mathbb{R}^d \times \mathbb{R}^d$ by setting, for each $x, t \in \mathbb{R}^d$, $K(x,t) =$*

$h(\|x-t\|^2)$. *Then* K *is positive definite for* any d *if and only if there is a probability measure* p *on* $\mathbb{R}_+$ *such that*

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

Note that the set $\mathbb{R}_+$ is *not* compact and the kernel $G(0)$ is not in $\mathcal{A}_+(\mathbb{R}^d)$. Therefore, on both accounts Theorem 1 does not apply in this circumstance. In general, 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 1 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 the discussion presented below is that $G(\infty) = \Delta$, the diagonal kernel. Furthermore, in the statement of the theorem below it is understood that when we say that $\hat{p}$ is a discrete probability measure on $\mathbb{R}_+$ we mean that $\hat{p}$ can have an atom not only at zero but also at infinity. Therefore, we can integrate any function relative to such a discrete measure over the extended positive real line provided such a function is defined therein.

Theorem 4. *Let* $G : \mathbb{R}_+ \to \mathcal{A}(\mathcal{X})$ *be defined as*

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

Then there exists a kernel $\hat{K} = \int_{\mathbb{R}_+} G(\omega) d\hat{p}(\omega) \in \mathcal{K}(G)$ *such that* $\hat{p}$ *is a discrete probability measure on* $\mathbb{R}_+$ *with at most* $m+1$ *atoms and, for any atom* $\hat{\omega} \in \mathbb{R}_+$ *of* $\hat{p}$, *we have that*

$$R(\hat{c}, G(\hat{\omega})) = \max\{R(\hat{c}, G(\omega)) : \omega \in \mathbb{R}_+\} \tag{22}$$

where $\hat{c}$ *is a solution to the equation*

$$\frac{1}{2\mu}\hat{K}_{\mathbf{x}}\hat{c} + \nabla Q^*(\hat{c}) = 0 \tag{23}$$

and the function Q^* *is continuously differentiable. Moreover, for every* $c \in \mathbb{R}^m$ *and* $K \in \mathcal{K}(G)$, *we have that*

$$R(\hat{c}, K) \leq R(\hat{c}, \hat{K}) \leq R(c, \hat{K}). \tag{24}$$

Proof. For every $\ell \in \mathbb{N}$ we consider the Gaussian kernel map on the interval $\Omega_\ell := [\ell^{-1}, \ell]$ and appeal to Theorem 1 to produce a sequence of kernels $\hat{K}_\ell = \int_{\Omega_\ell} G(\omega) d\hat{p}_\ell(\omega)$ and $\hat{c}_\ell \in \mathbb{R}^m$ with the properties described there. In particular, $\hat{p}_\ell$ is a discrete probability measure with at most $m+1$ atoms, a number *independent* of ℓ. Let us examine what may happen as ℓ tends towards infinity. Each of the atoms of $\hat{p}_\ell$ as well as 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 $\{\hat{p}_{n_\ell} : \ell \in \mathbb{N}\}$ of probability measures and kernels $\{\hat{K}_{n_\ell} : \ell \in \mathbb{N}\}$ such that $\lim_{\ell\to\infty} \hat{p}_{n_\ell} = \hat{p}$, $\lim_{\ell\to\infty} \hat{K}_{n_\ell} = \hat{K}$, and $\hat{K} = \int_{\mathbb{R}_+} G(\omega) d\hat{p}(\omega)$ with the provision that $\hat{p}$ may have atoms at either zero or infinity. In either case, we replace the Gaussian by its limit, namely $G(0)$, the identically one kernel, or $G(\infty)$, the delta kernel, in the integral which defines $\hat{K}$.

To establish that $\hat{K}$ is an optimal kernel, we turn our attention to the sequence of vectors $\hat{c}_{n_\ell}$. We claim that this sequence also has a convergent subsequence. Indeed, from equation (17) for every $K = \int_{\Omega_\ell} G(\omega) dp(\omega)$, $p \in \mathcal{P}(\Omega_\ell)$ we have that

$$R(\hat{c}_{n_\ell}, K) \leq R(0, \hat{K}_{n_\ell}) = Q^*(0) < \infty.$$

Using the fact that the function Q^* is bounded below (see our comments after the proof of Lemma 2) we see that the sequence $\hat{c}_{n_\ell}$ has Euclidean norm bounded independently of ℓ. Hence, it has a convergent subsequence whose limit we call $\hat{c}$. Passing to the limit we obtain equations (23) and (24) and, so, conclude that $(\hat{c}, \hat{K})$ is a saddle point. □

We remark that extensions of the results in this section also hold true for non-differentiable convex functions Q. The proofs presented above must be modified in this general case in detail but not in substance. We postpone the discussion of this issue to a future occasion.

4 A Greedy Algorithm for Learning the Kernel

The analysis in the previous section establishes necessary and sufficient conditions for a pair $(\hat{c}, \hat{K}) \in \mathbb{R}^m \times \mathcal{K}(G)$ to be a saddle point of the problem

$$-Q_\mu(G) := \max \{\min \{R(c, K) : c \in \mathbb{R}^m\} : K \in \mathcal{K}(G)\}.$$

The main step in this problem is to compute the optimal kernel $\hat{K}$. Indeed, once $\hat{K}$ has been computed, $\hat{c}$ is obtained as the unique solution c_K to the equation

$$\frac{1}{2\mu} K_{\mathbf{x}} c_K + \nabla Q^*(c_K) = 0 \tag{25}$$

for $K = \hat{K}$.

With this observation in mind, in this section we focus on the computational issues for the problem

$$-Q_\mu(G) = \max\{g(K) : K \in \mathcal{K}(G)\} \tag{26}$$

where the function $g : \mathcal{A}_+(\mathcal{X}) \to \mathbb{R}$ is defined as

$$g(K) := \min \{R(c, K) : c \in \mathbb{R}^m\}, \quad K \in \mathcal{A}_+(\mathcal{X}). \tag{27}$$

We present a greedy algorithm for learning an optimal kernel. The algorithm starts with an initial kernel $K^{(1)} \in \mathcal{K}(G)$ and computes iteratively a sequence of kernels $K^{(t)} \in \mathcal{K}(G)$ such that

$$g(K^{(1)}) < g(K^{(2)}) < \cdots < g(K^{(s)}) \tag{28}$$

Initialization: Choose $K^{(1)} \in \mathcal{K}(G)$
For $t = 1$ to T:

1. Compute $c^{(t)} = \text{argmin}\{R(c, K^{(t)}) : c \in \mathbb{R}^m\}$ using equation (25)
2. Find $\hat{\omega} \in \Omega : (c^{(t)}, G_{\mathbf{x}}(\hat{\omega})c^{(t)}) > (c^{(t)}, K_{\mathbf{x}}^{(t)}c^{(t)})$. If such $\hat{\omega}$ does not exist terminate
3. Compute $\hat{\lambda} = \text{argmax}\{g(\lambda G(\hat{\omega}) + (1-\lambda)K^{(t)}) : \lambda \in (0, 1]\}$
4. Set $K^{(t+1)} = \hat{\lambda}G(\hat{\omega}) + (1-\hat{\lambda})K^{(t)}$

Fig. 1. Algorithm to compute an optimal convex combination of kernels in the set $\{G(\omega) : \omega \in \Omega\}$.

where s is the number of iterations. At each iteration t, $1 \leq t \leq s$, the algorithm searches for a value $\hat{\omega} \in \Omega$, if any, such that

$$(c^{(t)}, G_{\mathbf{x}}(\hat{\omega})c^{(t)}) > (c^{(t)}, K_{\mathbf{x}}^{(t)}c^{(t)}) \tag{29}$$

where we have defined $c^{(t)} := c_{K^{(t)}}$. If such $\hat{\omega}$ is found then a new kernel $K^{(t+1)}$ is computed to be the optimal convex combination of the kernels $G(\hat{\omega})$ and $K^{(t)}$, that is,

$$g(K^{(t+1)}) = \max\left\{g(\lambda G(\hat{\omega}) + (1-\lambda)K^{(t)}) : \lambda \in [0, 1]\right\}. \tag{30}$$

If no $\hat{\omega} \in \Omega$ satisfying inequality (29) can be found, the algorithm terminates. The algorithm is summarized in Figure 1.

Step 2 of the algorithm is implemented with a local gradient ascent in Ω. If the value of ω found locally does not satisfy inequality (29), the smallest hyperrectangle containing the search path is removed and a new local search is started in the yet unsearched part of Ω, continuing in this way until either the whole of Ω is covered or inequality (29) is satisfied. Although in the experiments below we will apply this strategy when Ω is an interval, it also naturally applies to more complex parameter spaces, for example a compact set in a Euclidean space. Step 3 is a simple maximization problem which we solve using Newton method, since the function $g(\lambda G(\hat{\omega}) + (1-\lambda)K^{(t)})$ is concave in λ and its derivative can be computed by applying Lemma 4. We also use a tolerance parameter ϵ to enforce a non-zero gap in inequality (29). A version of this algorithm for the case when $\Omega = \mathbb{N}_n$ has also been implemented (below, we refer to this version as the "finite algorithm"). It only differs from the continuous version in Step 2, in that inequality (29) is tested by trial and error on randomly selected kernels from $\mathcal{K}_n$.

We now show that after each iteration, either the objective function g increases or the algorithm terminates, that is, inequality (28) holds true. To this end, we state the following lemma whose proof follows immediately from Theorem 2.

Lemma 3. *Let $K_1, K_2 \in \mathcal{A}_+(\mathcal{X})$. Then, $\lambda = 0$ is* not *a solution to the problem*

$$\max\{g(\lambda K_1 + (1-\lambda)K_2) : \lambda \in [0, 1]\}$$

Table 1. Misclassification error percentage for the continuous and finite versions of the algorithm and the SVM on different handwritten digit recognition tasks. See text for description.

Task \ Method	Cont.	Finite	SVM	Cont.	Finite	SVM	Cont.	Finite	SVM
	$\sigma \in [75, 25000]$			$\sigma \in [100, 10000]$			$\sigma \in [500, 5000]$		
odd vs. even	6.6	18.0	11.8	6.6	10.9	8.6	6.5	6.7	6.9
3 vs. 8	3.8	6.9	6.0	3.8	4.9	5.1	3.8	3.7	3.8
4 vs. 7	2.5	4.2	2.8	2.5	2.7	2.6	2.5	2.6	2.3
1 vs. 7	1.8	3.9	1.8	1.8	1.8	1.8	1.8	1.9	1.8
2 vs. 3	1.6	3.9	3.1	1.6	2.8	2.3	1.6	1.7	1.6
0 vs. 6	1.6	2.2	1.7	1.6	1.7	1.5	1.6	1.6	1.5
2 vs. 9	1.5	3.2	1.9	1.5	1.9	1.8	1.5	1.4	1.4
0 vs. 9	0.9	1.2	1.1	0.9	1.0	1.0	0.9	0.9	1.0

if and only if $R(c_{K_2}, K_1) > R(c_{K_2}, K_2)$.

Applying this lemma to the case that $K_1 = G(\hat{\omega})$ and $K_2 = K^{(t)}$ we conclude that

$$g(K^{(t+1)}) > g(K^{(t)})$$

if and only if $\hat{\omega}$ satisfies the inequality

$$R(c^{(t)}, G(\hat{\omega})) > R(c^{(t)}, K^{(t)})$$

which is equivalent to inequality (29).

4.1 Experimental Validation

We have tested the above algorithm on eight handwritten digit recognition tasks of varying difficulty from the MNIST data-set[3]. The data are 28×28 images with pixel values ranging between 0 and 255. We used Gaussian kernels as the basic kernels, that is, $G(\sigma)(x,t) = \exp(-\|x-t\|^2/\sigma^2)$, $\sigma \in [\sigma_1, \sigma_2]$. In all the experiments, the test error rates were measured over 1000 points from the MNIST test set.

The continuous and finite algorithms were trained using the square loss and compared to an SVM[4]. In all experiments, the training set consisted of 500 points. For the finite case, we chose ten Gaussian kernels with σ's equally spaced in an interval $[\sigma_1, \sigma_2]$. For both versions of our algorithm, the starting value of the kernel was the average of these ten kernels and the regularization parameter was set to 10^{-7}. This value typically provided the best test performance among the nine values $\mu = 10^{-\ell}$, $\ell \in \{3, \ldots, 11\}$. The performance of the SVM was obtained as the best among the results for the above ten kernels and nine values of μ. This strategy slightly favors the SVM but compensates for the fact that

[3] Available at: `http://yann.lecun.com/exdb/mnist/index.html`

[4] Trained using SVM-light, see: `http://www.cs.cornell.edu/People/tj/svm_light`

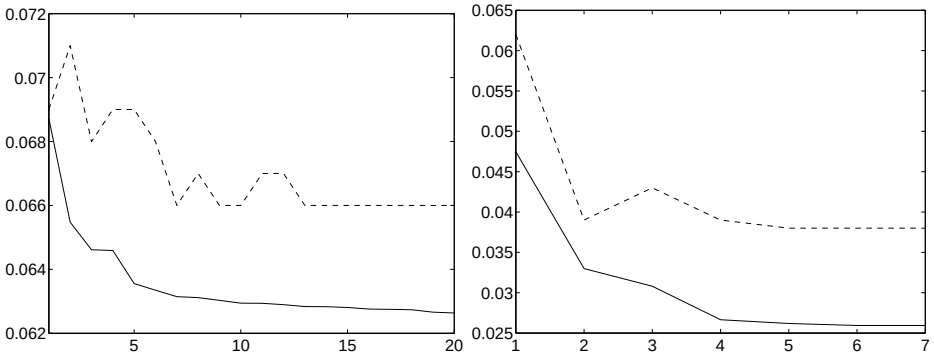

Fig. 2. Functional Q_μ (*solid line*) and misclassification error (*dotted line*) after the first iteration of the algorithm of Figure 1 for even vs. odd (*left*) and 3 vs. 8 (*right*).

the loss functions are different. The parameters ϵ and T of our algorithm were chosen to be 10^{-3} and 100 respectively.

Table 1 shows the results obtained. The range of σ is $[75, 25000]$ in columns 2–4, $[100, 10000]$ in columns 5–7 and $[500, 5000]$ in columns 8–10. Note that, in most cases, the continuous algorithm finds a better combination of kernels than the finite version. In general, the continuous algorithm performs better than the SVM, whereas most of the time the finite algorithm is worse than the SVM. Moreover, the results indicate that the continuous algorithm *is not affected by the range of* σ, unlike the other two methods.

Typically, the continuous algorithm requires less than 20 iterations to terminate whereas the finite algorithm may require as much as 100 iterations. Figure 2 depicts the convergence behavior of the continuous algorithm on two different tasks. In both cases $\sigma \in [100, 10000]$. The actual values of Q_μ are six orders of magnitude smaller, but they were rescaled to fit the plot. Note that, in agreement with inequality (28), Q_μ decreases and eventually converges. The misclassification error also converges to a lower value, indicating that Q_μ provides a good learning criterion.

5 Conclusion

We have studied the problem of learning a kernel which minimizes a convex error functional over the convex hull of prescribed basic kernels. The main contribution of this paper is a general analysis of this problem when the basic kernels are continuously parameterized by a compact set. In particular, we have shown that there always exists an optimal kernel which is a finite combination of the basic kernels and presented a greedy algorithm for learning a suboptimal kernel. The algorithm is simple to use and our preliminary findings indicate that it typically converges in a small number of iterations to a kernel with a competitive statistical performance. In the future we shall investigate the convergence properties of the

algorithm, compare it experimentally to previous related methods for learning the kernel, such as those in [3, 6, 8], and study generalization error bounds for this problem. For the latter purpose, the results in [4, 18] may be useful.

References

1. N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 686, pp. 337–404, 1950.
2. J.P. Aubin. *Mathematical Methods of Game and Economic Theory.* Studies in Mathematics and its applications, Vol. 7, North-Holland, 1982.
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, 2004.
4. O. Bousquet and D.J.L. Herrmann. On the complexity of learning the kernel matrix. *Advances in Neural Information Processing Systems*, 15, 2003.
5. J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization. Theory and Examples.* CMS (Canadian Math. Soc.) Springer-Verlag, New York, 2000.
6. O. Chapelle, V.N. Vapnik, O. Bousquet and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1), pp. 131–159, 2002.
7. 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.
8. G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui and M.I. Jordan. Learning the kernel matrix with semi-definite programming. *J. of Machine Learning Research*, 5, pp. 27–72, 2004.
9. C.A. Micchelli and M. Pontil. Learning the kernel function via regularization. To appear in *J. of Machine Learning Research* (see also Research Note RN/04/11, Department of Computer Science, UCL, June 2004)..
10. C. A. Micchelli and T. J. Rivlin. Lectures on optimal recovery. In Lecture Notes in Mathematics, Vol. 1129, P. R. Turner (Ed.), Springer Verlag, 1985.
11. 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.
12. H.L. Royden. *Real Analysis.* Macmillan Publ. Company, New York, 3rd ed., 1988.
13. I.J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39, pp. 811–841, 1938.
14. J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.
15. V.N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.
16. G. Wahba. *Spline Models for Observational Data.* Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.
17. T. Zhang. On the dual formulation of regularized linear systems with convex risks. *Machine Learning*, 46, pp. 91–129, 2002.
18. Q. Wu, Y. Ying and D.X. Zhou. Multi-kernel regularization classifiers. *Preprint*, City University of Hong Kong, 2004.

A Appendix

The first result we record here is a useful version of the classical von Neumann minimax theorem we have learned from [2, Ch. 7].

Theorem 5. *Let $h : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ 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 $x \mapsto h(x,y)$ is convex and lower semi-continuous for every $y \in \mathcal{B}$, the function $y \mapsto h(x,y)$ 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 $\{x : x \in \mathcal{A},\ h(x,y_0) \leq \lambda\}$ is compact then there is an $x_0 \in \mathcal{A}$ such that*

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

In particular, we have that

$$\min\{\sup\{h(x,y) : y \in \mathcal{B}\} : x \in \mathcal{A}\} = \sup\{\inf\{h(x,y) : x \in \mathcal{A}\} : y \in \mathcal{B}\}. \quad (31)$$

The hypothesis of lower semi-continuity means, for all $\lambda \in \mathbb{R}$ and $y \in \mathcal{B}$, that the set $\{x : x \in \mathcal{A},\ h(x,y) \leq \lambda\}$ is a closed subset of $\mathcal{A}$.

The next result concerns differentiation of a "max" function. Its proof can be found in [9].

Lemma 4. *Let $\mathcal{X}$ be a topological vector space, $\mathcal{T}$ 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 $\mathcal{X}$ as*

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

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 of Lemma 2. Theorem 5 applies since $K_{\mathbf{x}} \in \mathcal{L}_+(\mathbb{R}^m)$. Indeed, we let $h(c,v) = (K_{\mathbf{x}}c, v) - Q^*(v) + \mu(c, K_{\mathbf{x}}c)$, $\mathcal{A} = \mathbb{R}^m$, $\mathcal{B} = \{v : Q^*(v) < \infty, v \in \mathbb{R}^m\}$ and $v_0 = 0$. Then, $\mathcal{B}$ is convex and, for any $\lambda \in \mathbb{R}$, the set $\{c : c \in \mathbb{R}^m, h(c,v_0) \leq \lambda\}$ is compact. Therefore, all the hypotheses of Theorem 5 hold. Consequently, using (11) in (7) we have that

$$Q_\mu(K) = \sup\{\min\{(K_{\mathbf{x}}c, v) - Q^*(v) + \mu(c, K_{\mathbf{x}}c) : c \in \mathbb{R}^m\} : v \in \mathcal{B}\}. \quad (32)$$

For each $v \in \mathcal{B}$, the minimum over c satisfies the equation $K_{\mathbf{x}}v + 2\mu K_{\mathbf{x}}c = 0$, implying that

$$\min\{(K_{\mathbf{x}}c, v) - Q^*(v) + \mu(c, K_{\mathbf{x}}c) : c \in \mathbb{R}^m\} = -\frac{(v, K_{\mathbf{x}}v)}{4\mu} - Q^*(v)$$

and the result follows. □