Characterizing the Representer Theorem

Yao-Liang Yu
Hao Cheng
Dale Schuurmans
Csaba Szepesvári
Dept. of Computing Science, University of Alberta, Edmonton, AB, T6G 2E8 CANADA

Abstract

The representer theorem assures that kernel methods retain optimality under penalized empirical risk minimization. While a sufficient condition on the form of the regularizer guaranteeing the representer theorem has been known since the initial development of kernel methods, necessary conditions have only been investigated recently. In this paper we completely characterize the necessary and sufficient conditions on the regularizer that ensure the representer theorem holds. The results are surprisingly simple yet broaden the conditions where the representer theorem is known to hold. Extension to the matrix domain is also addressed.

1. Introduction

Reproducing kernel Hilbert spaces (RKHSs) are an important construction that has been studied in several fields, including functional analysis (Aronszajn, 1950; Schwartz, 1964), statistics (Wahba, 1990), computational mathematics (Kirsch, 2011) and, more recently, machine learning (Schölkopf & Smola, 2001; Shawe-Taylor & Cristianini, 2004). Methods that operate on an RKHS, so called kernel methods, are so compelling that one can witness their impact in virtually every area of machine learning.

A key property that underlies the successful application of kernel methods is the representer theorem (Kimeldorf & Wahba, 1971; Schölkopf et al., 2001), which allows one to conduct all optimization in a space whose dimension does not exceed the number of data points. In particular, consider the problem of penalized empirical risk minimization:

\[
\min_{f \in \mathcal{H}} L_n(f(x_1), \ldots, f(x_n)) + \lambda_n \|f\|^2,
\]

where \(x_i \in \mathcal{X}, i = 1, \ldots, n\), \(L_n : \mathbb{R}^n \rightarrow \mathbb{R}\) is some loss function, \(\mathcal{H} \subseteq \mathbb{R}^\mathcal{X}\) (with its Hilbert norm \(\| \cdot \|\)) is the RKHS induced by some kernel \(\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}\). Notice that the optimization problem (1) is expressed in the RKHS \(\mathcal{H}\), a vector space whose dimension is usually very high or even infinite.

Kimeldorf & Wahba (1971) were perhaps among the first to realize that (1) could actually be reduced, without loss of generality, to an optimization problem that is much more computationally friendly. They showed that minimizers of (1) must be of the form \(f(\cdot) = \sum_{i=1}^n \alpha_i \kappa(\cdot, x_i)\) for some \(\alpha \in \mathbb{R}^n\). Then using the reproducing property of the kernel one can equivalently re-express the problem (1) as merely finding the coefficients \(\alpha_i\) that solve:

\[
\min_{\alpha \in \mathbb{R}^n} L_n((K\alpha)_1, \ldots, (K\alpha)_n) + \lambda_n \alpha^\top K \alpha,
\]

where the kernel matrix \(K = (K_{ij})\) is defined with \(K_{ij} = \kappa(x_i, x_j)\). Note that (2) is a finite dimensional problem. Moreover, it belongs to the pleasant category of convex programs if the loss \(L_n\) is convex.

Although Kimeldorf & Wahba (1971) considered the simple squared Hilbert norm regularizer, as given in (1), this computational reduction is possible for more general regularizers. Let \(\Omega : \mathcal{H} \rightarrow \overline{\mathbb{R}}\) (where we use \(\overline{\mathbb{R}}\) to denote \(\mathbb{R} \cup \{\infty\}\)) and consider

\[
\min_{f \in \mathcal{H}} L_n(f(x_1), \ldots, f(x_n)) + \lambda_n \Omega(f).
\]

For this more general form Schölkopf et al. (2001) proved the following result.

Readers who are interested in the origins of RKHS can consult Aronszajn (1950), although this is not needed for reading this paper.
Finally, we conclude the paper in Section 5.

In retrospect, it is somewhat surprising that this result has not been discovered earlier given its directness and simplicity, and the wide applicability of kernel methods. Note that this complete characterization of the representer theorem has not been discovered earlier given its directness and simplicity, and the wide applicability of kernel methods. Henceforth, we will merely assume that $H$ is an inner product space.

Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ denote the set of extended reals. Following Argyriou et al. (2009), we adopt the following definition of admissibility:

**Definition 1** The function $\Omega : H \to \bar{\mathbb{R}}$ is admissible if for all $n, (f_i)_{i=1}^n$ and $(y_i)_{i=1}^n$, some minimizer of (5) admits the form

$$f = \sum_{i=1}^n \alpha_i f_i$$

for some $\alpha \in \mathbb{R}^n$. The function $\Omega$ is strictly admissible if all minimizers of (5) admit this form.

We will consider the statement to be vacuously true if (5) has no minimizer.

It is easy to see that if $\Omega$ is admissible in the sense of Definition 1 then the representer theorem also must hold for the penalized problem (3) (for any loss $L_n$, not necessarily convex). The other implication (if the representer theorem holds for (3), it must also hold for (5)) is true as well under mild assumptions, see (Argyriou et al., 2009).

Argyriou et al. (2009) proved that the sufficient condition in Theorem 1 is also necessary for the regularizer $\Omega$ to be admissible, provided that $\Omega$ is Gâteaux differentiable. A key step in (Argyriou et al., 2009) is to establish the following proposition, whose proof we reproduce here to keep the presentation self-contained.

**Proposition 1** Let $H$ be an inner product space. The

\[
\text{Proposition 1 (Representer theorem)} \quad \text{If} \quad \Omega = h(\|f\|) \quad \text{for some increasing function} \quad h : \mathbb{R}_+ \to \bar{\mathbb{R}}, \quad \text{then some minimizer of (3) must admit the form} \quad f(\cdot) = \sum_{i=1}^n \alpha_i \kappa(\cdot, x_i) \quad \text{for some} \quad \alpha \in \mathbb{R}^n. \quad \text{If} \quad h \quad \text{is strictly increasing, all minimizers admit this form.}
\]

This theorem gives a sufficient condition on the regularizer $\Omega$ that assures the representer theorem holds. Some recent effort has been devoted to understanding what conditions are necessary. Notably Argyriou et al. (2009) considered a close variant of (3) (the interpolation problem, see (5) below) and showed that the sufficient condition in Theorem 1 is also necessary provided that the regularizer $\Omega$ is differentiable. More recently, Dinuzzo & Schölkopf (2012) managed to relax the differentiability assumption to mere lower semicontinuity. However, a complete, easily verifiable, necessary and sufficient characterization of the representer theorem is still lacking. We notice that (Warmuth & Vishwanathan, 2005; Warmuth et al., 2012) gave another characterization of the representer theorem, under a somewhat different formulation hence is not directly comparable.

Building on the pioneering work of Argyriou et al. (2009) and the recent work of Dinuzzo & Schölkopf (2012), we prove that the representer theorem (for the interpolation problem) holds if and only if the regularizer $\Omega$ is a weakly increasing function of the Hilbert norm, i.e.,

$$\forall f, g \in H, \|g\| > \|f\| \implies \Omega(g) \geq \Omega(f).$$

Since we work with interpolation problems, we in fact prove this result for inner product spaces (not even Hilbert spaces).

In retrospect, it is somewhat surprising that this result has not been discovered earlier given its directness and simplicity, and the wide applicability of kernel methods. Note that this complete characterization of the representer theorem has practical consequence: the desire to enjoy the representer theorem must prevent one from designing interesting regularizers that are not (almost) an increasing function of the RKHS norm.

To establish the main result, in Section 2, we first recall some definitions and an important proposition due to Argyriou et al. (2009). Then in Section 3 we establish our main result and provide some discussion of its consequences. We point out in Section 4 that an enhanced problem in the matrix domain can be treated similarly, although the results there are less complete. Finally, we conclude the paper in Section 5.
function $\Omega : \mathcal{H} \to \mathbb{R}$ is admissible iff
\[
\forall f, g \in \mathcal{H}, (f, g) = 0 \Rightarrow \Omega(f + g) \geq \Omega(f). \quad (7)
\]

It is strictly admissible iff
\[
\forall f, g \in \mathcal{H}, (f, g) = 0, g \neq 0 \Rightarrow \Omega(f + g) > \Omega(f). \quad (8)
\]

Proof: $\Rightarrow$: Suppose $\Omega$ is admissible. Consider the following instance of (5):
\[
\min_{h \in \mathcal{H}} \Omega(h) \text{ s.t. } (h, f) = \|f\|^2. \quad (9)
\]
The admissibility of $\Omega$ implies that $f$ is a minimizer of (9). Since $f + g$ for any $g \perp f$ is feasible for (9), we have $\Omega(f + g) \geq \Omega(f)$ from the optimality of $f$.

$\Leftarrow$: Suppose (7) holds and (5) has a minimizer $f$. Then, we have $\Omega(f + g) \geq \Omega(f)$, which the admissibility of $\Omega$ will follow. Pick any $f,g$ such that $(f,g) = 0 \neq 0$, we have $\Omega(f + g) > \Omega(f)$. Then we have $\|f + g\| > \|f\|$, which follows from (7) and (10). Therefore, for the remainder of this section we will exclude this trivial case and assume $\dim(\mathcal{H}) \geq 2$ henceforth.

The main result of this paper, which, in retrospect could be considered to be the “correct form” of the representer theorem, is the following:

**Theorem 2** Let $\mathcal{H}$ be an inner product space. A function $\Omega : \mathcal{H} \to \mathbb{R}$ is admissible iff
\[
\forall f, g \in \mathcal{H}, (f, g) > \|f\| \Rightarrow \Omega(g) \geq \Omega(f). \quad (10)
\]

It is strictly admissible iff
\[
\forall f, g \in \mathcal{H}, (f, g) > \|f\| \Rightarrow \Omega(g) > \Omega(f). \quad (11)
\]

Note that the lower semicontinuity of $\Omega$ is not required for the statement to hold.

The condition equivalent to strict admissibility (11) can be stated concisely as the requirement that $\Omega$ be strictly increasing as a function of the norm of its argument. On the other hand, (10) is not the usual “increasing” property, but instead is a weaker requirement—we henceforth refer to this property as the weak increasing property. Then, the condition equivalent to admissibility can be stated concisely as the requirement that $\Omega$ has to be weakly increasing as a function of the norm of its argument.

**Proof:** $\Leftarrow$: Suppose (10) holds. We verify (7), from which the admissibility of $\Omega$ will follow. Pick any $f, g \in \mathcal{H}$ such that $(f, g) = 0 \neq 0$. Then we have $\|f + g\| > \|g\|$ and thus by (10), $\Omega(f + g) \geq \Omega(g)$. Noting that the case $f = 0$ also trivially holds, we see that (7) holds. By Proposition 1, we get that $\Omega$ is admissible.

$\Rightarrow$: Suppose now that $\Omega$ is admissible. Then, by Proposition 1, (7) holds. Note that in the special case when $f = 0$ and $g \neq 0$ (so that $\|g\| > 0$), we have $\Omega(g) = \Omega(0 + g) \geq \Omega(0) = \Omega(f)$. Therefore, in what follows we need only deal with the case when $f \neq 0$. To prove (10), we start with a claim.
Claim: The admissibility of $\Omega$ implies $\Omega(\cdot)$ is increasing along any ray $R_g = \{tg : t \geq 0\},\ g \neq 0 \in \mathcal{H}$. By the above reasoning it suffices to prove this claim for $R_g \setminus \{0\}$. We prove the claim using a geometric argument depicted in the left panel of Figure 1. For a fixed vector $g \in \mathcal{H}$ and an angle $\theta \in [0,\pi/2]$ choose some $f \in \mathcal{H}$ such that $f$ is not parallel to $g$. Such an $f$ exists since $\dim(\mathcal{H}) \geq 2$. Now, let $g_0$ be the rotation of $g$ in the plane (subspace) $P$ spanned by $f$ and $g$. The direction of rotation can be chosen arbitrarily. Take the line in the plane $P$ that passes through $g_0$ and which is orthogonal to $g_0$. Let $t(\theta)g$ be the point where the ray $R_g$ and the line intersect and let the vector $p_0$ be defined as $g_0 + p_0 = t(\theta)g$. Note that $t(\theta) = (1 + \tan^2(\theta))^{1/2} \geq 1$ for all $\theta \in [0,\pi/2)$. Thus, $p_0$ is orthogonal to $g_0$: $p_0 \perp g_0$. Further, let $s(\theta)g$ be the orthogonal projection of $g_0$ to the ray $R_g$ and call $q_0$ the vector that satisfies $s(\theta)g + q_0 = g_0$. Thus, $q_0 \perp s(\theta)g$. Further, $s(\theta) = \cos(\theta) \leq 1$ for all $\theta \in [0,\pi/2)$. Applying (7) from Proposition 1 twice we get

$$
\Omega(t(\theta)g) = \Omega(g_0 + p_0) = \Omega(g_0)
= \Omega(s(\theta)g + q_0) \geq \Omega(s(\theta)g).
$$

(12)

Thus, the claim and (7), we get

$$
\Omega(g) \geq \Omega(pg)
= \Omega(g_0) = \Omega(f_n - f_{n-1})
= \Omega(f_n - f_{n-1}) = \Omega(f_{n-2} + (f_{n-1} - f_{n-2}))
= \cdots
\geq \Omega(f_0) = \Omega(f),
$$

finishing the proof of (10).

The above proof can be easily mimicked for the strictly admissible case.

A significant portion of our proof is devoted to proving that any function satisfying (7) is necessarily increasing along any ray starting from the origin. We note that Dimuzzo & Schölkopf (2012) presented a concise algebraic proof of this fact (cf. the proof of Theorem 1 in their paper). Giving a geometric interpretation to their proof leads to the proof presented above, which we prefer as it leads nicely to the geometric proof of the necessity of (10). Furthermore, the geometric interpretation will offer a convenient approach for understanding the matrix case as well.

The reason the continuity conditions can be avoided in Theorem 2, making the result simpler and more elegant, is that the necessary condition for the admissibility of $\Omega$ avoids stipulating $\Omega$’s behavior on the surface of balls. In fact, if one modified (2) to include the case when $||f|| = ||g||$, this would imply that $\Omega$ is radial (i.e., $\Omega(f)$ depends on the argument $f$ only through $||f||$). The next example demonstrates that one can have an admissible regularizer that is not radial (of course, such an $\Omega$ cannot also be semicontinuous).

**Example 1** Figure 2 shows an admissible regularizer $\Omega$ that is not radial. The gray area denotes, say, the
region \( \{ \| f \| \leq 1 \} \) and the red point represents some \( g \) on \( \{ \| f \| = 1 \} \). It is clear that \( \Omega \) is neither l.s.c. nor u.s.c.\(^4\). Note also that \( \Omega \) is in fact a convex admissible regularizer (demonstrating that convex functions can be “ugly” on boundary points).

**Remark 1** As the previous example demonstrates, there exist non-radial, but admissible regularizers. However, Theorem 2 also implies that every admissible function is equal to an admissible radial function except for a set whose cardinality is at most “countable”. To see this consider the function \( I(r) = \inf \{ \Omega(f) : \| f \| = r \} \). Clearly \( I : \mathbb{R}_+ \to \mathbb{R} \) is an increasing function, hence it can have only at most countably many discontinuity points. But it is easily seen that for any continuity point \( r \) of \( I \) and any \( f, g \in H \) on the \( H \)-sphere of radius \( r \), it follows that \( \Omega(f) = \Omega(g) \). Thus, \( \Omega \) is radial except for at most countably many spheres.

Before refining Theorem 2, let us remark that there is a useful result that a function \( \Omega : H \to \mathbb{R} \) is u.s.c. iff for all \( f \in H \), \( \Omega(f) = \limsup \Omega(f_n) \). Of course, \( \Omega \) is continuous iff it is both l.s.c. and u.s.c. Another equivalent characterization of l.s.c. (u.s.c.) is the closedness (openness) of the sublevel sets.

**Remark 2** One should not confuse the l.s.c. (u.s.c.) of \( \Omega : H \to \mathbb{R} \) with the l.s.c. (u.s.c.) of \( \Omega : \text{dom } \Omega \to \mathbb{R} \). The former condition, used throughout this paper, is strictly stronger than the latter condition; refer to Figure 2 for an example.

As noted in the introduction, under the assumption that \( \Omega \) is l.s.c., Dinuzzo & Schölkopf (2012) proved that the sufficient condition in Theorem 1 is also necessary. We now show that this statement remains true even if the u.s.c. requirement is replaced by l.s.c., which is essentially the main result of Dinuzzo & Schölkopf (2012).

**Remark 3** Another easy way to see the result in Theorem 3 is to notice that the function \( I(r) \) defined in Remark 1 is in fact continuous when \( \Omega \) satisfies (14) (or equivalently (7)) and is either l.s.c. or u.s.c. Based on Theorem 2 and 3, it can also be shown that the lower or upper semicontinuous hull\(^6\) of (strictly) admissible regularizers remain (strictly) admissible, although the reverse implication is false.

It turns out that positive homogeneity, other than semicontinuity, also forces admissible regularizers to be radial. Notice that both properties imply that the function \( I(r) \) discussed in Remark 1 is continuous.

**Theorem 4** Let \( H \) be an inner product space with (the induced) norm \( \| \cdot \| \). If \( \Omega \) is admissible and positively homogeneous, then it is a positive multiple of the norm \( \| \cdot \| \).

**Proof:** We prove first that \( \Omega \) must be an increasing function of the norm. Note that due to positive homogeneity, we have \( \Omega(0) = 0 \) hence \( \Omega \geq 0 \) by the admissibility. Suppose to the contrary there exist \( x, y \in H \) such that \( \| x \| = \| y \| \neq 0 \) but \( \Omega(x) > \Omega(y) \). Then for all \( 1 < \lambda < \frac{\Omega(x)}{\Omega(y)} \), \( \| \lambda y \| = \lambda \| y \| > \| x \| \), hence \( \Omega(\lambda y) > \Omega(x) \) by the admissibility. Due to positive homogeneity, \( \lambda > \Omega(x)/\Omega(y) \), contradiction.

Take an arbitrary \( x_0 \) with unit norm (i.e., \( \| x_0 \| = 1 \)), then apparently \( \Omega(x) = \| x \| \cdot \Omega(x_0) \). The proof is now complete.

\(\text{The consequence of Theorem 4 is immediate: Essentially, any other (semi)norm defined on } H \text{ (which may or may not be compatible with the topology of } H \text{) can not be admissible. Obviously if } \Omega \text{ is admissible and positively homogeneous with degree } r \text{ (i.e., } \Omega(\lambda x) = \lambda^r \cdot \Omega(x) \text{)} \text{ then we have } \Omega(x) = \| x \|^r \cdot \Omega(x_0) \text{ for some (arbitrary) } x_0 \text{ having unit norm.}

The next example shows that neither l.s.c. nor u.s.c.
is necessary for $\Omega$ to be an increasing radial function and hence satisfy (14).

**Example 2** Figure 3 shows a regularizer that is an increasing radial function, but is neither l.s.c. nor u.s.c. Here, the gray region denotes $\{\|f\| < 1\}$, while the red circle represents $\{\|f\| = 1\}$. This example, despite of its triviality, motivates our next development. Note that we can write

$$\Omega = [1 \cdot \mathbb{1}_{\|f\| \leq 1} + 2 \cdot \mathbb{1}_{\|f\| > 1}] \land [0 \cdot \mathbb{1}_{\|f\| < 1} + 2 \cdot \mathbb{1}_{\|f\| \geq 1}],$$

where $\land$ denotes pointwise infimum. Observe that the first function is l.s.c. while the second is u.s.c.

Needless to say, if $\Omega$, satisfies (10) or (14), then so do $\sum_i \Omega_i$, $\inf_i \Omega_i$, and $\sup_i \Omega_i$ respectively (whenever they are well-defined). In fact, much more can be said. We end this section with a necessary and sufficient characterization of (14).

**Theorem 5** $\Omega$ satisfies (14) iff $\Omega = \Omega_1 \lor \Omega_2$, or $\Omega = \Omega_1 \lor \Omega_2$, or $\Omega = \Omega_1 + \Omega_2$, where $\Omega_1$ is l.s.c. and admissible while $\Omega_2$ is u.s.c. and admissible.

**Proof:** $\Rightarrow$: Suppose $\Omega$ satisfies (14), then there exists an increasing function $h : \mathbb{R}_+ \to \mathbb{R}$ so that $\Omega(f) = h(\|f\|)$. Obviously $h$ has at most countably many discontinuous points, which we denote as $D = \{t_i\}_{i \in \mathbb{N}}$ (arranged according to their magnitude).

Let us first consider the case $\Omega = \Omega_1 \land \Omega_2$. If $D = \emptyset$, then simply take $\Omega_i(f) = h(\|f\|), i = 1, 2$; otherwise define

$$h_1(t) = \sum_{i \in \mathbb{N}} [h(t_i) + h(t_i) - h(t_i - 1)] \cdot \mathbb{1}_{t_i-1 < t < t_i},$$

which is l.s.c. and increasing. Next define (the so-called u.s.c. hull)$^8$

$$h_2(t) = \sup \{s : t \in \text{cl}\{s \geq t\}\}.$$
the vector case: essentially we are freed from considering details of the loss term while at the same time losing little generality.

We extend the definition of admissibility in a straightforward manner:

**Definition 2** The matrix regularizer $\Omega : \mathcal{M} \to \mathbb{R}$ is (strictly) admissible if for all $n$, $X_i = (x_{i,1}, \ldots, x_{i,k}) \in \mathbb{R}^{d \times k}$, $y_i \in \mathbb{R}^k$, $1 \leq i \leq n$, one (respectively, all) of the minimizers $W = (w_1, \ldots, w_k)$ of (15) satisfies that for all $1 \leq p \leq k$, the $p$th column of $W$, $w_p$, is in the linear subspace in $\mathbb{R}^d$ spanned by the columns of the matrices $X_i$:

$$w_p = \sum_{i=1}^n \sum_{q=1}^k \alpha^{(p)}_{i,q} x_{i,q}, \quad 1 \leq p \leq k$$

(16)

for some real numbers $(\alpha^{(p)}_{i,q})_{i,q,p}$.

The important thing to notice here is that $w_p$ depends on all $\{x_{i,q}\}_{1 \leq i \leq n, 1 \leq q \leq k}$, not just $\{x_{i,q}\}_{1 \leq i \leq n}$, even though the latter are all the vectors that constrain $w_p$ in (15).

As in the previous section, an important tool for studying admissibility is the following generalization of Proposition 1, established by Argyriou et al. (2009):

**Proposition 2** $\Omega : \mathcal{M} \to \mathbb{R}$ is admissible iff it satisfies

$$\forall W, P \in \mathcal{M}, W^T P = 0 \Rightarrow \Omega(W + P) \geq \Omega(W).$$

(17)

It is strictly admissible iff

$$\forall W, P \in \mathcal{M}, W^T P = 0 \Rightarrow \Omega(W + P) > \Omega(W).$$

(18)

The proof is quite similar to that of Proposition 1, hence it is omitted. We note that Argyriou et al. (2009) also point out that Proposition 2 remains true even when the $X_i$ are restricted to be of rank 1.

We are now ready to characterize (17), which turns out to be more involved than one might expect. In the statement below all orderings between matrices are with respect to the Löwner partial ordering. Further, the symbol $A \succeq B$ means that $A \geq B$ and $A \neq B$.

**Theorem 6** Let $\Omega : \mathcal{M} \to \mathbb{R}$. Consider the following statements:

(a) $(A + B)^\top (A + B) \succeq A^\top A \Rightarrow \Omega(A + B) \geq \Omega(A)$;

(b) $(A + B)^\top (A + B) \preceq A^\top A \Rightarrow \Omega(A + B) \geq \Omega(A)$;

(c) $\Omega$ is admissible, i.e. $A^\top B = 0 \Rightarrow \Omega(A + B) \geq \Omega(A)$;

(d) $A^\top B = 0$ and $B^\top B \succ 0 \Rightarrow \Omega(A + B) \geq \Omega(A)$.

Then, (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d). Moreover, if $d \geq k$ then (d) together with $\Omega$ being u.s.c. imply (c). If $d \geq 2k$ and $\Omega$ is either l.s.c. or u.s.c. then (c) implies (a).

**Proof:** Clearly we have (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d).

Let us now show (d) $\Rightarrow$ (c) under the assumption that $\Omega$ is u.s.c. Fix $A$ and $B$ such that $A^\top B = 0$. If $B^\top B > 0$, then we have $\Omega(A + B) \geq \Omega(A)$; otherwise, thanks to $d \geq k$, we can find $B_n, B_n > 0$, $A^\top B_n = 0$ and $B_n \to B$ as $n \to \infty$. Since $\Omega$ is u.s.c., $\Omega(A + B) \geq \lim sup_{n \to \infty} \Omega(A + B_n) \geq \Omega(A)$. Therefore we conclude that (d) $\Rightarrow$ (c).

Finally, we prove (c) $\Rightarrow$ (a) when $\Omega$ is either l.s.c. or u.s.c. and $d \geq 2k$. We claim that $\Omega(A)$ is independent of the left singular vectors of $A$ in the sense that for any $r \leq k$, $\sigma_j \geq 0$, $(u_j)$, $(z_j)$ are orthonormal systems of $\mathbb{R}^d$, $(v_j)$ orthonormal system of $\mathbb{R}^k$, $1 \leq j \leq r$, if $A = \sum_{j=1}^r \sigma_j u_j v_j^\top$ and $B = \sum_{j=1}^r \sigma_j z_j v_j^\top$, then $\Omega(A) = \Omega(B)$. To see this, it suffices to show that $\Omega(A) \leq \Omega(B)$ because reversing the roles of $A$ and $B$ gives $\Omega(A) = \Omega(B)$.

Thus, we can be of the above form. Thanks to $2r \leq 2k \leq d$, it is possible to find unit vectors $o_1, o_2, \ldots, o_r \in \mathbb{R}^d$ such that

$$o_1 = u_1 \perp \{u_2, \ldots, u_r\},$$

$$o_2 \perp \{o_1, u_3, \ldots, u_r, z_1\},$$

$$\ldots \perp \ldots$$

$$o_r \perp \{o_1, \ldots, o_{r-1}, z_1, \ldots, z_{r-1}\}.$$  

Suppose now that $\Omega$ is u.s.c. Then, using the rotation idea presented in the right part of Figure 1, we get

$$\Omega(A) = \Omega(\sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top + \sum_{i=3}^r \sigma_i u_i v_i^\top)$$

$$\leq \Omega(\sigma_1 o_1 v_1^\top + \sigma_2 o_2 v_2^\top + \sum_{i=3}^r \sigma_i o_i v_i^\top)$$

$$\vdots$$

$$\leq \Omega(\sigma_1 z_1 v_1^\top + \sigma_2 o_2 v_2^\top + \sum_{i=3}^r \sigma_i o_i v_i^\top)$$

$$\vdots$$

$$\leq \Omega(\sigma_1 z_1 v_1^\top + \sigma_2 o_2 v_2^\top + \sum_{i=3}^r \sigma_i z_i v_i^\top)$$

$$= \Omega(B).$$
For instance, consider the first inequality: Let \( f = u_2, g = o_2 \). Note that by construction \( \text{span}\{u_2, o_2\} \perp u_j \) for any \( 1 \leq j \leq r, j \neq 2 \). Let \( A' = \sum_{j \neq 2} \sigma_j u_j v_j^\top \).

Hence, given the vectors \( p_i \), the matrices \( P_i = \sigma_2 p_i v_2^\top \) are such that \( P_i^\top (\sigma_2 f_i v_2^\top + A') = 0 \) while \( P_i + \sigma_2 f_i v_2^\top + A' = \sigma_2 f_{i+1} v_2^\top + A', i = 0, \ldots, n - 1 \). Thus, by (c),

\[
\Omega(\sigma_2 u_2 v_2^\top + A') = \Omega(\sigma_2 f_0 v_2^\top + A') \\
\vdots \\
= \Omega(\sigma_2 f_i v_2^\top + A') \\
\leq \Omega(\sigma_2 f_i + v_2^\top + A') \\
= \Omega(\sigma_2 f_{i+1} v_2^\top + A') \\
\vdots \\
\leq \Omega(\sigma_2 f_n v_2^\top + A').
\]

Now, notice that \( f_n \to o_2 \) as \( n \to \infty \). Thus, by the u.s.c. of \( \Omega \), \( \limsup_{n \to \infty} \Omega(\sigma_2 f_n v_2^\top + A') \leq \Omega(\sigma_2 o_2 v_2^\top + A') \). This, together with the previous inequality gives \( \Omega(\sigma_2 u_2 v_2^\top + A') \leq \Omega(\sigma_2 o_2 v_2^\top + A') \), which was the inequality to be proven.

If \( \Omega \) is l.s.c., then use a similar rotation idea and change accordingly the direction of the above inequalities. This finishes the proof that \( \Omega(A) \) is independent of the left singular vectors of \( A \).

Clearly, it suffices to show that for any \( A, B \in \mathcal{M} \), such that \( A^\top A \succeq B^\top B \), \( \Omega(A) \succeq \Omega(B) \). Thus, fix \( A, B \) with these properties. For a matrix \( X \in \mathcal{M} \) let \( U_X \in \mathcal{M} \) be the matrix obtained from the left singular vectors of \( X \). Note that for any matrix \( Y \in \mathcal{M}, U_X Y \) and \( Y \) have the same singular values and right singular vectors. Since we have shown \( \Omega(X) \) is invariant to the left singular vectors of \( X \), it follows that for any \( Y \in \mathcal{M}, \Omega(Y) = \Omega(U_X (Y^\top Y)^{1/2}) \). Thus, \( \Omega(A) = \Omega(U_A (A^\top A)^{1/2}) = \Omega(U_A [(C + D)^\top (C + D)]^{1/2}) = \Omega(C + D), \) where \( C^\top = ((B^\top B)^{1/2}, \ldots), D^\top = (A^\top A - B^\top B)^{1/2} \). We have padded necessary zeros in \( C \) and \( D \) so that they belong to \( \mathcal{M} \) (hence \( \Omega \) can be applied on them). By construction \( C^\top D = 0 \), hence (c) gives \( \Omega(C + D) \geq \Omega(C) = \Omega(U_C (C^\top C)^{1/2}) = \Omega(U_C (B^\top B)^{1/2}) = \Omega(B) \).

The proof is now complete.

\[\square\]

Our proof of (c) \( \Rightarrow \) (a) closely follows that of Theorem 15 in the paper by Argyriou et al. (2009), except that we have managed to relax their differentiability assumption to semicontinuity.

Next, we show by means of some examples that the implications in Theorem 6 cannot be improved in general.

**Example 3** (b) \( \not\Rightarrow \) (a): Setting \( k = 1 \) makes (b) the same as (10) while (a) the same as (14). Example 1 then consists of a counterexample.

(c) \( \not\Rightarrow \) (b): Let \( d = 4, k = 2 \) hence \( d \geq 2k \) is met. Take an arbitrary rank-1 matrix \( X \) and set \( \Omega(X) = 1.5 \) while \( \Omega(A) = \text{rank}(A) \) at all other points \( A \). Apparently under this specification \( \Omega \) is admissible but on the other hand \( \Omega(X + X) = \Omega(2X) = 1 < 1.5 = \Omega(X) \) hence (b) is false. Needless to say that this example also demonstrates that (c) \( \not\Rightarrow \) (a).

(d) \( \not\Rightarrow \) (d) and l.s.c. \( \not\Rightarrow \) (c): Let \( d = 4, k = 2 \) hence \( d \geq 2k \) is met. Set \( \Omega() = 1, \Omega(X) = 0 \) where \( X \) is an arbitrary rank-1 matrix. Put \( \Omega = \infty \) at all other points. One may verify that \( \Omega \) is indeed l.s.c. and satisfies (d). But \( \Omega \) is not admissible since \( \Omega(X + X) = \Omega(X) < \Omega() \).

**Remark 4** Example 3 is a bit surprising once we realize that when \( k = 1 \) (i.e., we go back to the case considered in Section 3), then (b), (c) and (d) are actually all equivalent. Clearly the matrix case exhibits some difficulty that is not present for inner product spaces. Considering this new difficulty, perhaps one should not be too disappointed with the incomplete characterization in Theorem 6. We also observe that u.s.c. and l.s.c. no longer play similar roles in the matrix domain.

**5. Conclusion**

We have proved that for the interpolation problem, the representer theorem holds if and only if the regularizer is a weakly increasing function of the inner product induced norm. This complete characterization of the representer theorem excludes the possibility of designing (non-standard) regularizers that enjoy the representer theorem without being (almost) an increasing radial function. Extension to the matrix domain is also given, although the results are less complete in this case due to some new complexities we have identified.

Finally we mention that for vector-valued kernels (Micchelli & Pontil, 2005; Carmeli et al., 2006), our results continue to hold as a sufficient condition, while a complete characterization seems to require a substantially new idea.
References


