Characterizing the Representer Theorem

Yao-Liang Yu YAOLIANG@CS.UALBERTA.CA
Hao Cheng HCHENG2@CS.UALBERTA.CA
Dale Schuurmans DALE@CS.UALBERTA.CA
Csaba Szepesvári SZEPESVA@CS.UALBERTA.CA
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:

$$\min_{f\in\mathcal{H}} L_n\big(f(x_1),\ldots,f(x_n)\big) + \lambda_n\|f\|^2, \tag{1}$$

where $x_i \in \mathcal{X}, i = 1,\ldots,n$, $L_n : \mathbb{R}^n \to \mathbb{R}$ is some loss function, $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ (with its Hilbert norm $\|\cdot\|$) is the RKHS induced by some kernel $\kappa : \mathcal{X}\times\mathcal{X} \mapsto \mathbb{R}$.[1] 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 $\boldsymbol{\alpha} \in \mathbb{R}^n$. Then using the reproducing property of the kernel one can equivalently re-express the problem (1) as merely finding the coefficients α_i that solve:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} L_n\Big((K\boldsymbol{\alpha})_1,\ldots,(K\boldsymbol{\alpha})_n\Big) + \lambda_n\boldsymbol{\alpha}^\top K\boldsymbol{\alpha}, \tag{2}$$

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 $\Omega : \mathcal{H} \to \bar{\mathbb{R}}$ (where we use $\bar{\mathbb{R}}$ to denote $\mathbb{R}\cup\{\infty\}$) and consider

$$\min_{f\in\mathcal{H}} L_n\big(f(x_1),\ldots,f(x_n)\big) + \lambda_n\Omega(f). \tag{3}$$

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

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

[1] Readers who are interested in the origins of RKHS can consult Aronszajn (1950), although this is not needed for reading this paper.

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

This theorem gives a sufficient condition on the regularizer Ω 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 Ω 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 Ω is a *weakly* increasing function of the Hilbert norm, *i.e.*,

$$\forall f, g \in \mathcal{H}, \|g\| > \|f\| \implies \Omega(g) \geq \Omega(f). \tag{4}$$

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.

2. Previous Work

We recall in this section some previous work on characterizing the representer theorem.

As pointed out by Argyriou et al. (2009), to study the representer theorem, one can (and perhaps should) focus on the interpolation problem:

$$\min_{f \in \mathcal{H}} \Omega(f) \quad \text{s.t. } \langle f, f_i \rangle = y_i, \ i = 1, \ldots, n. \tag{5}$$

Usually we take $f_i = \kappa(\cdot, x_i)$ hence the constraint in (5) becomes $f(x_i) = y_i$ thanks to the reproducing property of the kernel κ. The advantage of considering interpolation is that the loss function L_n no longer plays any role in the specification. Note that the reproducing property of the kernel is only used for the loss term, therefore if we restrict attention to interpolation (5) the results in this section will continue to hold provided only that $\mathcal{H}$ be an inner product space. Henceforth, *we will merely assume that $\mathcal{H}$ 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 $\Omega : \mathcal{H} \to \bar{\mathbb{R}}$ is admissible if for all n, $(f_i)_{i=1}^n$ and $(y_i)_{i=1}^n$, some minimizer of (5) admits the form*

$$f = \sum_{i=1}^n \alpha_i f_i \tag{6}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^n$. The function Ω 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 Ω 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 Ω to be admissible, provided that Ω 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 $\mathcal{H}$ be an inner product space. The* *function* $\Omega : \mathcal{H} \to \bar{\mathbb{R}}$ *is admissible iff*

$$\forall f, g \in \mathcal{H}, \langle f, g\rangle = 0 \;\Rightarrow\; \Omega(f+g) \geq \Omega(f). \tag{7}$$

It is strictly admissible iff

$$\forall f, g \in \mathcal{H}, \langle f, g\rangle = 0,\; g \neq 0 \Rightarrow \Omega(f+g) > \Omega(f). \tag{8}$$

Proof: $\Rightarrow$: Suppose Ω is admissible. Consider the following instance of (5):

$$\min_{h\in\mathcal{H}} \;\Omega(h) \;\text{ s.t. }\; \langle h, f\rangle = \|f\|^2. \tag{9}$$

The admissibility of Ω 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 $f^* = g + g^\perp$, where $g \in \mathcal{H}_m \doteq \operatorname{span}\{f_1, \ldots, f_m\}$ and $g^\perp$ is in the orthogonal complement of $\mathcal{H}_m$.[2] Clearly $\langle f^*, f_i\rangle = \langle g, f_i\rangle = y_i$, hence g is feasible. Invoking (7) we know $\Omega(f^*) \geq \Omega(g)$, therefore g is also a minimizer, proving the admissibility of Ω.

The proof for strict admissibility is similar. ∎

Although Proposition 1 gives a complete characterization of admissibility, the verification of its conditions can be cumbersome. Our goal is to prove that admissibility is equivalent to the much simpler to understand and state condition that the regularizer Ω be a weakly increasing function of the norm, in the sense of (4).

A recent result that comes very close to providing a complete and easy to verify characterization of admissibility is due to Dinuzzo & Schölkopf (2012), who showed that as long as Ω is lower semicontinous (l.s.c.),[3] the sufficient condition in Theorem 1 is also necessary. We now demonstrate that a modification of their proof yields a more complete and far simpler characterization of admissibility that removes the l.s.c. requirement.

3. Complete Characterization

We first make an easy observation. If the vector space $\mathcal{H}$ has unit dimension, *i.e.* $\dim(\mathcal{H}) = 1$, then the condition (7) is equivalent to requiring $\Omega(g) \geq \Omega(0)$ for all $g \in \mathcal{H}$; similarly the condition (8) is equivalent to requiring $\Omega(g) > \Omega(0)$ for all $g \neq 0$. Therefore, for the remainder of this section we will exclude this trivial case and assume $\dim(\mathcal{H}) \geq 2$ henceforth.

[2] The existence of such a decomposition depends only on the completeness of $\mathcal{H}_m$, not on that of $\mathcal{H}$. Note that $\mathcal{H}_m$ is indeed complete since it is of finite dimension.

[3] A function $f : \mathcal{H} \to \bar{\mathbb{R}}$ is lower semicontinuous if and only if its epigraph (the set of points lying on or above its graph) as a subset of $\mathcal{H} \times \bar{\mathbb{R}}$ is closed.

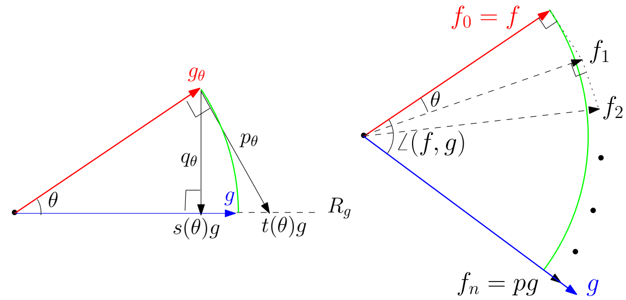

Figure 1. Pictorial illustration of the main idea presented in the proofs of Theorem 2 and Theorem 6.

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* $\Omega : \mathcal{H} \to \bar{\mathbb{R}}$ *is admissible iff*

$$\forall f, g \in \mathcal{H}, \|g\| > \|f\| \;\Rightarrow\; \Omega(g) \geq \Omega(f). \tag{10}$$

It is strictly admissible iff

$$\forall f, g \in \mathcal{H}, \|g\| > \|f\| \;\Rightarrow\; \Omega(g) > \Omega(f). \tag{11}$$

Note that the lower semicontinuity of Ω is not required for the statement to hold.

The condition equivalent to strict admissibility (11) can be stated concisely as the requirement that Ω 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 Ω 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 Ω will follow. Pick any $f, g \in \mathcal{H}$ such that $\langle f, g\rangle = 0, f \neq 0$. 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 Ω is admissible.

$\Rightarrow$: Suppose now that Ω 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 Ω implies $\Omega(\cdot)$ is increasing along any ray $R_g = \{tg : t \geq 0\}$, $0 \neq g \in \mathcal{H}$.

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_θ 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_θ and which is orthogonal to g_θ. Let $t(\theta)g$ be the point where the ray R_g and the line intersect and let the vector p_θ be defined as $g_\theta + p_\theta = t(\theta)g$. Note that $t(\theta) = (1 + \tan^2(\theta))^{1/2} \geq 1$ for all $\theta \in [0, \pi/2)$. Thus, p_θ is orthogonal to g_θ: $p_\theta \perp g_\theta$. Further, let $s(\theta)g$ be the orthogonal projection of g_θ to the ray R_g and call q_θ the vector that satisfies $s(\theta)g + q_\theta = g_\theta$. Thus, $q_\theta \perp s(\theta)g$. Further, $s(\theta) = \cos(\theta) \leq 1$ for all $\theta \in [0, \pi/2)$. Applying (7) from Proposition 1 twice we get

$$\begin{aligned}\Omega(t(\theta)g) &= \Omega(g_\theta + p_\theta) \geq \Omega(g_\theta) \\ &= \Omega(s(\theta)g + q_\theta) \geq \Omega(s(\theta)g).\end{aligned} \tag{12}$$

Note that this holds for any $g \in \mathcal{H}$, $g \neq 0$ and $\theta \in [0, \pi/2)$.

Now, take any $0 < \tau_1 < \tau_2$. Since $t(\theta)/s(\theta)$ is continuous on $[0, \pi/2)$ and its range is $[1, \infty)$, there exists a value $\theta' \in [0, \pi/2)$ such that

$$\frac{t(\theta')}{s(\theta')} = \frac{\tau_2}{\tau_1}\,. \tag{13}$$

Define $c = \tau_2/t(\theta')$. Note that we also have that $c = \tau_1/s(\theta')$ thanks to (13). Hence, applying (12) to cg and θ', we get

$$\Omega(\tau_2 g) = \Omega\big(t(\theta')(cg)\big) \geq \Omega\big(s(\theta')(cg)\big) = \Omega(\tau_1 g)\,,$$

finishing the proof of the claim.

Now if $\|g\| > \|f\|$ and f is not aligned with g, it is not hard to see (cf. Figure 1, right panel) that one can find a sufficiently large $n \geq 1$, a real number $p \in (0,1)$ and a sequence $f_0 = f, f_1, \ldots, f_n = pg$ such that for any $0 \leq i \leq n-1$, $\angle(f_i, f_{i+1}) = \theta \doteq \angle(f,g)/n$ and $(f_{i+1} - f_i) \perp f_i$. Indeed, n defines the above sequence uniquely with some $p = p_n > 0$. In particular, $p_n\|g\| = \|f_n\| = t(\theta/n)^n\|f\|$, so $p_n = t(\theta/n)^n \frac{\|f\|}{\|g\|}$. Since $t(\theta/n)^n \sim (1 + (\theta/n)^2)^{n^2/\theta^2 \cdot \theta^2/n} \sim e^{\theta^2/n} \to 1$ as $n \to \infty$, $p_n \to \frac{\|f\|}{\|g\|} < 1$ and so the existence of (n,p) with the said properties is guaranteed. Therefore, using the claim and (7), we get

$$\begin{aligned}\Omega(g) &\geq \Omega(pg) \\ &= \Omega(f_n) = \Omega(f_{n-1} + (f_n - f_{n-1})) \\ &\geq \Omega(f_{n-1}) = \Omega(f_{n-2} + (f_{n-1} - f_{n-2})) \\ &\;\;\vdots \\ &\geq \Omega(f_0) = \Omega(f)\,,\end{aligned}$$

thus finishing the proof of (10).

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

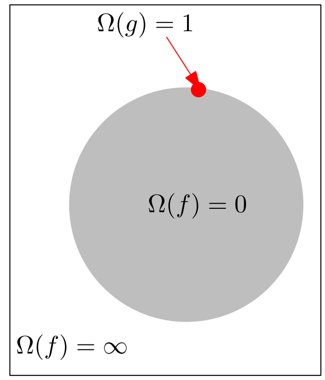

Figure 2. An admissible regularizer Ω that is *not* an increasing function of the RKHS norm.

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 Dinuzzo & 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 Ω avoids stipulating Ω's behavior on the surface of balls. In fact, if one modified (2) to include the case when $\|f\| = \|g\|$, this would imply that Ω 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 Ω cannot also be semicontinuous).

Example 1 *Figure 2 shows an admissible regularizer Ω that is not radial. The gray area denotes, say, the* *region $\{\|f\| \le 1\}$ and the red point represents some g on $\{\|f\| = 1\}$. It is clear that Ω is neither l.s.c. nor u.s.c.*[4] *Note also that Ω 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 $I(r) \doteq \inf\{\Omega(f) : \|f\| = r\}$. Clearly $I : \mathbb{R}_+ \to \bar{\mathbb{R}}$ 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 $f, g \in \mathcal{H}$ on the $\mathcal{H}$-sphere of radius r, it follows that $\Omega(f) = \Omega(g)$. Thus, Ω 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 $\Omega : \mathcal{H} \to \bar{\mathbb{R}}$ is u.s.c. iff for all $f \in \mathcal{H}$, $\Omega(f) = \limsup_{f_n \to f} \Omega(f_n)$. Of course, Ω 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 $\Omega : \mathcal{H} \to \bar{\mathbb{R}}$ with the l.s.c. (u.s.c.) of $\Omega : \operatorname{dom}\Omega \to \mathbb{R}$. 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 Ω 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 we replace l.s.c. by u.s.c.[5].

Theorem 3 *Let Ω be u.s.c. Then Ω is admissible iff*

$$\forall f, g \in \mathcal{H}, \|f\| \ge \|g\| \;\Rightarrow\; \Omega(f) \ge \Omega(g), \tag{14}$$

or, in other words, Ω is an increasing radial function. Further, Ω is strictly admissible iff it is a strictly increasing radial function.

Proof: $\Leftarrow$: (14) apparently implies (10) hence the admissibility of Ω.

$\Rightarrow$: Assume that Ω is u.s.c. and admissible. Thanks to Theorem 2, we need only prove that if $\|f\| = \|g\|$ then $\Omega(f) \ge \Omega(g)$. To see this, take a sequence $(f_n)_n$ that converges to f and that satisfies $\|f_n\| > \|f\|$. Then, $\|f_n\| > \|f\| = \|g\|$ also holds; therefore, by Theorem 2, $\Omega(f_n) \ge \Omega(g)$ holds for all n. Taking the $\limsup$ of both sides, we get $\Omega(f) \ge \limsup_{n\to\infty} \Omega(f_n) \ge \Omega(g)$.

The strictly admissible case follows immediately. ■

Note that an entirely analogous argument establishes that the theorem remains true 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 Ω 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 hulls*[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 $\mathcal{H}$ be an inner product space with (the induced) norm $\|\cdot\|$. If Ω is admissible and positively homogeneous, then it is a positive multiple of the norm $\|\cdot\|$.*

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

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

The next example shows that neither l.s.c. nor u.s.c.

[4] A function is u.s.c. when $-f$ is l.s.c.

[5] Note that one cannot naively negate an u.s.c. function here since our starting tool (7) is *not* invariant to negation.

[6] The lower semicontinuous hull of a function f is defined as $f_{\text{lsc}}(x) = \sup_{g \le f} g(x)$, subject to g is continuous.

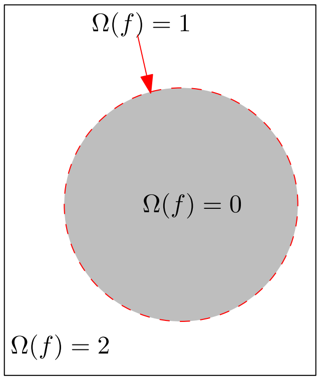

Figure 3. A regularizer Ω that is an increasing radial function without being either l.s.c. or u.s.c.

is necessary for Ω 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, say* $\{\|f\| < 1\}$, *while the red circle represents* $\{\|f\| = 1\}$. *This example, despite of its triviality, motivates our next development. Note that we can write*

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

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

Needless to say, if Ω_γ satisfies (10) or (14), then so do $\sum_\gamma \Omega_\gamma, \inf_\gamma \Omega_\gamma$ and $\sup_\gamma \Omega_\gamma$ 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 Ω *satisfies* (14) *iff* $\Omega = \Omega_1 \wedge \Omega_2$, *or* $\Omega = \Omega_1 \vee \Omega_2$, *or* $\Omega = \Omega_1 + \Omega_2$, *where* Ω_1 *is l.s.c. and admissible while* Ω_2 *is u.s.c. and admissible.*

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

Let us first consider the case $\Omega = \Omega_1 \wedge \Omega_2$. If $D = \emptyset$, then simply take $\Omega_i(f) = h(\|f\|), i = 1, 2$; otherwise define[7]

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

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

$$h_2(t) \doteq \sup\{s : t \in \mathrm{cl}\{h \geq s\}\}.$$

It is easily verified that h_2 is the smallest u.s.c. function that majorizes h. (In our case this amounts to simply modifying h at the discontinuous points D so that it becomes right continuous.) Clearly h_2 is non-decreasing (since h is so). Finally put $\Omega_i(f) = h_i(\|f\|)$ and then Theorem 5 can be verified without difficulty.

The case $\Omega = \Omega_1 \vee \Omega_2$ is proved similarly.

For the last case when $\Omega = \Omega_1 + \Omega_2$, define

$$\begin{aligned} h_1(t) &\doteq \sum_{i \in \mathbb{N}} \Big[h(t) - \sum_{j=1}^{i} (h(t_j+) - h(t_j))\Big] \cdot \mathbb{1}_{t_i<t\leq t_{i+1}}, \\ &\quad + \quad h(t) \cdot \mathbb{1}_{t \leq t_1}, \\ h_2(t) &\doteq \sum_{i \in \mathbb{N}} [h(t_i+) - h(t_i)] \cdot \mathbb{1}_{t_i<t}. \end{aligned}$$

One can easily verify that h_1 is increasing and u.s.c. while h_2 is increasing and l.s.c. Unsurprisingly, putting $\Omega_i(f) = h_i(\|f\|), i = 1, 2$ completes the construction.

$\Leftarrow$: Immediate consequence of the remark made before Theorem 5. ∎

4. Extension to Matrices

In this section we generalize our results to the matrix domain. Matrix norm regularizers have been proven very useful in multi-task learning (Argyriou et al., 2008) and collaborative filtering (Abernethy et al., 2009). We will restrict ourselves to finite dimensional matrices, since the extension to linear operators does not bring anything conceptually new (while the functional analysis technicalities might obscure the main ideas).

Fix the integers k, d and let $\mathcal{M} = \mathbb{R}^{d \times k}$ be the space of $d \times k$ matrices. For a square matrix S, let $\mathrm{diag}(S)$ denote the vector formed from the diagonal elements of S. Consider the matrix interpolation problem:

$$\min_{W \in \mathcal{M}} \Omega(W) \text{ s.t. } \mathrm{diag}(W^\top X_i) = y_i, \; 1 \leq i \leq n. \quad (15)$$

Here, $X_i \in \mathbb{R}^{d \times k}$, $y_i \in \mathbb{R}^k$. Note that $k = 1$ corresponds to the previously considered vector case. If $n > d$, the problem is overconstrained and might not have a solution. The motivation to focus on the interpolation problem (15) in this case is similar to that in

[7] In the case $D = \{t_1, \ldots, t_n\}$, we add $t_0 = 0$ and $t_{n+1} = \infty$ to the set D, in order to handle the boundary.

[8] cl means taking the topological closure of the set.

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, $X_i = (x_{i,1}, \ldots, x_{i,k}) \in \mathbb{R}^{d\times k}$, $y_i \in \mathbb{R}^k$, $1 \le i \le n$, one (respectively, all) of the minimizers $W = (w_1, \ldots, w_k)$ of (15) satisfies that for all $1 \le p \le k$, the pth column of W, w_p is in the linear subspace in $\mathbb{R}^d$ spanned by the columns of the matrices X_i:*

$$w_p = \sum_{i=1}^{n}\sum_{q=1}^{k} \alpha_{i,q}^{(p)} x_{i,q}, \quad 1 \le p \le k \tag{16}$$

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

The important thing to notice here is that w_p depends on all $\{x_{i,q}\}_{1\le i\le n, 1\le q\le k}$, not just $\{x_{i,p}\}_{1\le i\le n}$, 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*

$$\forall W, P \in \mathcal{M},\ W^\top P = 0 \Rightarrow \Omega(W + P) \ge \Omega(W). \tag{17}$$

It is strictly admissible iff

$$\forall W, P \in \mathcal{M},\ W^\top P = 0 \Rightarrow \Omega(W + P) > \Omega(W). \tag{18}$$

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 $A \succneqq B$ means that $A \succeq B$ and $A \ne B$.

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

(a). $(A + B)^\top(A + B) \succeq A^\top A \Rightarrow \Omega(A + B) \ge \Omega(A)$*;*

(b). $(A + B)^\top(A + B) \succneqq A^\top A \Rightarrow \Omega(A + B) \ge \Omega(A)$*;*

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

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

Then, (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d). Moreover, if $d \ge k$ then (d) together with Ω being u.s.c. imply (c). If $d \ge 2k$ and Ω 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 Ω is u.s.c. Fix A and B such that $A^\top B = 0$. If $B^\top B \succ 0$, then we have $\Omega(A+B) \ge \Omega(A)$; otherwise, thanks to $d \ge k$, we can find $B_n^\top B_n \succ 0, A^\top B_n = 0$ and $B_n \to B$ as $n \to \infty$. Since Ω is u.s.c., $\Omega(A+B) \ge \limsup_{n\to\infty} \Omega(A + B_n) \ge \Omega(A)$. Therefore we conclude that $(d) \Rightarrow (c)$.

Finally, we prove $(c) \Rightarrow (a)$ when Ω is either l.s.c. or u.s.c. and $d \ge 2k$. We claim that $\Omega(A)$ is independent of the left singular vectors of A in the sense that for any $r \le k$, $\sigma_j \ge 0$, (u_j), (z_j) are orthonormal systems of $\mathbb{R}^d$, (v_j) orthonormal system of $\mathbb{R}^k$, $1 \le j \le r$, 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 $\Omega(A) \le \Omega(B)$ because reversing the roles of A and B gives $\Omega(A) = \Omega(B)$. Thus, fix the matrices, A, B, to be of the above form. Thanks to $2r \le 2k \le d$, it is possible to find unit vectors $o_1, o_2, \ldots, o_r \in \mathbb{R}^d$ such that

$$\begin{aligned}
&o_1 = u_1 \perp \{u_2, \ldots, u_r\},\\
&o_2 \perp \{o_1, u_3, \ldots, u_r, z_1\},\\
&\ldots \perp \ldots\\
&o_r \perp \{o_1, \ldots, o_{r-1}, z_1, \ldots, z_{r-1}\}.
\end{aligned}$$

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

$$\begin{aligned}
\Omega(A) &= \Omega(\sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top + \sum_{i=3}^r \sigma_i u_i v_i^\top)\\
&\le \Omega(\sigma_1 o_1 v_1^\top + \sigma_2 o_2 v_2^\top + \sum_{i=3}^r \sigma_i u_i v_i^\top)\\
&\ \ \vdots\\
&\le \Omega(\sigma_1 o_1 v_1^\top + \sigma_2 o_2 v_2^\top + \sum_{i=3}^r \sigma_i o_i v_i^\top)\\
&\le \Omega(\sigma_1 z_1 v_1^\top + \sigma_2 o_2 v_2^\top + \sum_{i=3}^r \sigma_i o_i v_i^\top)\\
&\ \ \vdots\\
&\le \Omega(\sigma_1 z_1 v_1^\top + \sigma_2 z_2 v_2^\top + \sum_{i=3}^r \sigma_i z_i v_i^\top)\\
&= \Omega(B).
\end{aligned}$$

For instance, consider the first inequality: Let $f = u_2, g = o_2$. Note that by construction $\mathrm{span}\{u_2, o_2\} \perp u_j$ for any $1 \le j \le r$, $j \neq 2$. 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 $P_i + \sigma_2 f_i v_2^\top + A' = \sigma_2 f_{i+1} v_2 + A'$, $i = 0, \ldots, n-1$. Thus, by (c),

$$
\begin{aligned}
\Omega(\sigma_2 u_2 v_2^\top + A') &= \Omega(\sigma_2 f_0 v_2^\top + A') \\
&\vdots \\
&= \Omega(\sigma_2 f_i v_2^\top + A') \\
&\le \Omega(P_i + \sigma_2 f_i v_2^\top + A') \\
&= \Omega(\sigma_2 f_{i+1} v_2^\top + A') \\
&\vdots \\
&\le \Omega(\sigma_2 f_n v_2^\top + A').
\end{aligned}
$$

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

If Ω 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 $A^\top A \succeq B^\top B$, $\Omega(A) \ge \Omega(B)$. 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 $C^\top = ((B^\top B)^{1/2}, ,), D^\top = (, (A^\top A - B^\top B)^{1/2},)$. We have padded necessary zeros in C and D so that they belong to $\mathcal{M}$ (hence Ω can be applied on them). By construction $C^\top D = 0$, hence (c) gives $\Omega(C+D) \ge \Omega(C) = \Omega(U_C (C^\top C)^{1/2}) = \Omega(U_C (B^\top B)^{1/2}) = \Omega(B)$. The proof is now complete.

■

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 $d \ge 2k$ is met. Take an arbitrary rank-1 matrix X and set $\Omega(X) = 1.5$ while $\Omega(A) = \mathrm{rank}(A)$ at all other points A. Apparently under this specification Ω 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) + l.s.c. $\not\Rightarrow (c)$: Let $d = 4, k = 2$ hence $d \ge 2k$ 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 Ω is indeed l.s.c. and satisfies (d). But Ω is not admissible since $\Omega(X +) = \Omega(X) < \Omega()$.

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

Abernethy, Jacob, Bach, Francis, Evgeniou, Theodoros, and Vert, Jean-Philippe. A New Approach to Collaborative Filtering: Operator Estimation with Spectral Regularization. *Journal of Machine Learning Research*, 10:803–826, December 2009.

Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, January 2008.

Argyriou, Andreas, Micchelli, Charles A., and Pontil, Massimiliano. When is there a representer theorem? vector versus matrix regularizers. *Journal of Machine Learning Research*, 10:2507–2529, 2009.

Aronszajn, Nachman. Theory of reproducing kernels. *Transactions of the American Mathematical Sociery*, 68:337–404, 1950.

Carmeli, Claudio, Vito, Ernesto De, and Toigo, Alessandro. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4:377–408, 2006.

Dinuzzo, Francesco and Schölkopf, Bernhard. The representer theorem for Hilbert spaces: a necessary and sufficient condition, 2012. URL http://arxiv.org/abs/1205.1928.

Kimeldorf, George and Wahba, Grace. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

Kirsch, Andreas. *An Introduction to the Mathematical Theory of Inverse Problems*, volume 120 of *Applied Mathematical Sciences*. Springer, 2nd edition, August 2011.

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

Schölkopf, Bernhard and Smola, Alex J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

Schölkopf, Bernhard, Herbrich, Ralf, and Smola, Alex J. A generalized representer theorem. In *Conference on Computational Learning Theory*, 2001.

Schwartz, Laurent. Sous-espaces Hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d'Analyse Mathématique*, 13:115–256, 1964.

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

Wahba, Grace. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

Warmuth, Manfred K. and Vishwanathan, S.V.N. Leaving the span. In *Conference on Computational Learning Theory*, 2005.

Warmuth, Manfred K., Kotlowski, Wojciech, and Zhou, Shuisheng. Kernelization of matrix updates, when and how? In *Conference on Algorithmic Learning Theory*, 2012.