Journal of Global Optimization (2019) 74:401–415
https://doi.org/10.1007/s10898-019-00767-0

When is there a representer theorem?

Nondifferentiable regularisers and Banach spaces

Kevin Schlegel[1]

Received: 25 July 2018 / Accepted: 28 March 2019 / Published online: 8 April 2019

Abstract
We consider a general regularised interpolation problem for learning a parameter vector from data. The well known representer theorem says that under certain conditions on the regulariser there exists a solution in the linear span of the data points. This is at the core of kernel methods in machine learning as it makes the problem computationally tractable. Necessary and sufficient conditions for differentiable regularisers on Hilbert spaces to admit a representer theorem have been proved. We extend those results to nondifferentiable regularisers on uniformly convex and uniformly smooth Banach spaces. This gives a (more) complete answer to the question when there is a representer theorem. We then note that for regularised interpolation in fact the solution is determined by the function space alone and independent of the regulariser, making the extension to Banach spaces even more valuable.

Keywords Representer theorem · Regularised interpolation · Regularisation · Semi-inner product spaces · Kernel methods

Mathematics Subject Classification 68T05

1 Introduction

Regularisation is often described as a process of adding additional information or using previous knowledge about the solution to solve an ill-posed problem or to prevent an algorithm from overfitting to the given data. This makes it a very important method for learning a function from empirical data from very large classes of functions. Intuitively its purpose is to pick from all the functions that may explain the data the function which is the simplest in some suitable sense. Hence regularisation appears in various disciplines wherever empirical data is produced and has to be explained by a function. This has motivated to study regularisation problems in mathematics, statistics and computer science and in particular in machine

✉ Kevin Schlegel
schlegel@maths.ox.ac.uk

[1] Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter Woodstock Road, Oxford OX2 6GG, UK

learning theory (Cucker and Smale [4], Shawe-Taylor and Cristianini [16], Micchelli and Pontil [13]).

In particular regularisation in Hilbert spaces has been studied in the literature for various reasons. First of all the existence of inner products allows for the design of algorithms with very clear geometric intuitions often based on orthogonal projections or the fact that the inner product can be seen as a kind of similarity measure.

But in fact crucial for the success of regularisation methods in Hilbert spaces is the well known *representer theorem* which states that for certain regularisers there is always a solution in the linear span of the data points (Kimeldorf and Wahba [8], Cox and O'Sullivan [3], Schölkopf and Smola [14,17]). This means that the problem reduces to finding a function in a finite dimensional subspace of the original function space which is often infinite dimensional. It is this dimension reduction that makes the problem computationally tractable.

Another reason for Hilbert space regularisation finding a variety of applications is the *kernel trick* which allows for any algorithm which is formulated in terms of inner products to be modified to yield a new algorithm based on a different symmetric, positive semidefinite kernel leading to learning in *reproducing kernel Hilbert spaces* (Schölkopf and Smola [15], Shawe-Taylor and Cristianini [16]). This way nonlinearities can be introduced in the otherwise linear setup. Furthermore kernels can be defined on input sets which a priori do not have a mathematical structure by embeddings into a Hilbert space.

When we are speaking of regularisation we are referring to *Tikhonov regularisation*, i.e. an optimisation problem of the form

$$\min\left\{\mathcal{E}((\langle f, x_i\rangle_{\mathcal{H}}, y_i)_{i=1}^m) + \lambda\Omega(f) \,:\, f \in \mathcal{H}\right\}$$

where $\mathcal{H}$ is a Hilbert space, $\{(x_i, y_i) : i \in \mathbb{N}_m\} \subset \mathcal{H} \times Y$ is a set of given input/output data with $Y \subseteq \mathbb{R}$, $\mathcal{E}\colon \mathbb{R}^m \times Y^m \to \mathbb{R}$ is an *error function*, $\Omega : \mathcal{H} \to \mathbb{R}$ a *regulariser* and $\lambda > 0$ is a *regularisation parameter*. Argyriou et al. [1] show that under very mild conditions this regularisation problem admits a linear representer theorem if and only if the regularised interpolation problem

$$\min\left\{\Omega(f) \,:\, f \in \mathcal{H}, \langle f, x_i\rangle_{\mathcal{H}} = y_i \,\forall i = 1, \ldots, m\right\} \tag{1}$$

admits a linear representer theorem. They argue that we can thus focus on the regularised interpolation problem which is more convenient to study. It is easy to see that their argument holds for the more general setting of the problem which we are going to introduce in this paper so we are going to take the same viewpoint in this paper and consider regularised interpolation.

We will be interested in regularisation not only in Hilbert spaces as stated above but *extend the theory to uniformly convex, uniformly smooth Banach spaces*, allowing for learning in a much larger variety of spaces. While any two Hilbert spaces of the same dimension are linearly isometrically isomorphic this is far from true for Banach spaces so they exhibit much richer geometric variety which may be exploited in learning algorithms. Furthermore we may encounter applications where the data has some intrinsic structure so that it cannot be embedded into a Hilbert space. Having a large amount of Banach spaces for potential embeddings may help to overcome this problem. Analogous to learning in reproducing kernel Hilbert spaces the generalisation to Banach spaces allows for learning in *reproducing kernel Banach spaces* which have been introduced by Zhang et al. [18]. Our results regarding the existence of representer theorems are in line with Zhang and Zhang's work on representer theorems for reproducing kernel Banach spaces [19].

But as we will show at the end of this paper the variety of spaces to pose the problem in is of even greater importance. It is often said that the regulariser favours solutions with a

certain desirable property. We will show that in fact for regularised interpolation when we rely on the linear representer theorem it is essentially *the choice of the space*, and only the choice of the space not the choice of the regulariser, which *determines the solution*.

It is well known that non-decreasing functions of the Hilbert space norm admit a linear representer theorem. Argyriou et al. [1] showed that this condition is not just necessary but for differentiable regularisers also sufficient. In this paper we *remove the differentiablity condition* and show that any regulariser on a uniformly convex and uniformly smooth Banach space that admits a linear representer theorem is in fact very close to being radially symmetric, thus giving a (more) complete answer to the question when there is a representer theorem. Before presenting those results we present the necessary theory of semi-inner products to generalise the Hilbert space setting considered by Argyriou, Micchelli and Pontil to Banach spaces.

In Sect. 2 we will introduce the notion of *semi-inner products* as defined by Lumer [11] and later extended by Giles [6]. We will state the results without proofs as they mostly are not difficult and can be found in the original papers. Another extensive reference about semi-inner products and their properties is the work by Dragomir [5].

After introducing the relevant theory we will present the generalised regularised interpolation problem in Sect. 3, replacing the inner product in Eq. (1) by a semi-inner product. We then state one of the main results of the paper that regularisers that admit a representer theorem are almost radially symmetric in a way that will be made precise in the statement. Before giving the proof of the theorem we state and prove two essential lemmas capturing most of the important structure of the problem to prove the theorem. We finish the section by giving the proof of the main result.

Finally in Sect. 4 we prove that in fact for admissible regularisers there is a unique solution of the regularised interpolation problem in the linear span of the data and it is independent of the regulariser. This in particular means that we may choose the regulariser which is most suitable for our task at hand without changing the solution.

1.1 Notation

Before the main sections we briefly introduce some notation used throughout the paper. We use $\mathbb{N}_m$ as a shorthand notation for the set $\{1, \dots, m\} \subset \mathbb{N}$. We will assume we have m data points $\{(x_i, y_i) : i \in \mathbb{N}_m\} \subset \mathcal{B} \times Y$, where $\mathcal{B}$ will always denote a uniformly convex, uniformly smooth real Banach space and $Y \subseteq \mathbb{R}$. Typical examples of Y are finite sets of integers for classification problems, e.g. $\{-1, 1\}$ for binary classification, or the whole of $\mathbb{R}$ for regression.

We briefly recall the definitions of a Banach space being uniformly convex and uniformly smooth, further details can be found in [2,9,10].

Definition 1 (*Uniformly convex Banach space*) A normed vector space V is said to be uniformly convex if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $x, y \in V$ with $\|x\|_V = \|y\|_V = 1$ and $\|x - y\|_V > \varepsilon$ then $\|\frac{x+y}{2}\|_V < 1 - \delta$.

Definition 2 (*Uniformly smooth Banach space*) A normed vector space V is said to be uniformly smooth if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $x, y \in V$ with $\|x\|_V = 1$, $\|y\|_V \leq \delta$ then $\|x + y\|_V + \|x - y\|_V \leq 2 + \varepsilon\|y\|_V$.

Remark 1 There are two equivalent conditions of uniform smoothness which we will make use of in this paper.

(i) The modulus of smoothness of the space V is defined as

$$\rho_V(\delta) = \sup\left\{\frac{\|x+y\|_V + \|x-y\|_V}{2} - 1 \,:\, \|x\|_V = 1, \|y\|_V = \delta\right\} \tag{2}$$

Now V is uniformly smooth if and only if $\frac{\rho_V(\delta)}{\delta} \underset{\delta\to 0}{\longrightarrow} 0$.

(ii) The norm on V is said to be uniformly Fréchet differentiable if the limit

$$\lim_{t\to 0} \frac{\|x + t\cdot y\|_V - \|x\|_V}{t}$$

exists uniformly for all real t and $x, y \in V$ with $\|x\|_V = \|y\|_V = 1$. The space V is uniformly smooth if its norm is uniformly Fréchet differentiable.

We always write $\mathcal{H}$ to denote a Hilbert space and for the first part of Sect. 2 we will be speaking of general normed linear spaces denoted by V. Once we have seen the reasons to require the space to be a uniformly convex and uniformly smooth Banach space the remainder of Sect. 2 and the paper will consider such spaces denoted by $\mathcal{B}$. When only the norm $\|\cdot\|_{\mathcal{B}}$ on $\mathcal{B}$ is considered the subscript will often be omitted for simplicity. Throughout we will denote the inner product on a Hilbert space by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and a semi-inner product on a normed linear space by $[\cdot,\cdot]_V$.

2 Semi-inner product spaces

There are various definitions of semi-inner products aiming to generalise Hilbert space methods to more general cases. The notion of semi-inner products we are going to use was first introduced by Lumer [11] and further developed by Giles [6]. In comparison to inner products the assumption of symmetry, or equivalently additivity in the second argument, is dropped. Following Giles we assume homogeneity in the second argument. Moreover we need to assume that the Cauchy–Schwarz inequality holds, as it is crucial for the semi-inner products to have inner-product like behaviour.

In this section we will give a brief summary of the most important results. More details can be found in the two papers referenced above. Moreover an extensive overview of the theory of this and other notions of semi-inner products can be found in Dragomir [5]. While all results in this section can be given for complex vector spaces we will only present them for the real case as our results only apply to real vector spaces.

Definition 3 (*Semi-inner product*) A semi-inner product (s.i.p.) on a real vector space V is a map $[\cdot,\cdot]_V : V \times V \to \mathbb{R}$ which is linear in the first argument, homogeneous in the second argument, positive definite, and satisfies the Cauchy–Schwarz inequality.

With these properties a semi-inner product $[\cdot,\cdot]_V$ induces a norm $[x,x]_V = \|x\|_V$ on V. Conversely every norm $\|\cdot\|_V$ on a linear space V is induced by at least one semi-inner product, i.e. there exists at least one semi-inner product $[\cdot,\cdot]_V$ such that $\|x\|_V = [x,x]_V$. This means that every normed linear space is a s.i.p. space, but the semi-inner product inducing the norm is in general not unique. We do have uniqueness for uniformly smooth spaces though.

Proposition 1 *If the norm $\|\cdot\|_V$ on V is uniformly Fréchet differentiable as defined in* Remark 1*, then the differential of the norm for $x \neq 0$ is given by*

$$\lim_{t\to 0} \frac{\|x + ty\|_V - \|x\|_V}{t} = \frac{\mathrm{Re}\,[y,x]_V}{\|x\|_V}$$

The existence of a semi-inner product allows us to define a notion of orthogonality analogous to orthogonality in Hilbert spaces by requiring the semi-inner product to be zero. It turns out that this is equivalent to a generalisation of orthogonality to normed linear spaces introduced by James [7]. James in particular points out that his definition is closely related to linear functionals and hyperplanes. This is essential for our applications as we will see in the main part of the paper.

The lack of symmetry of the semi-inner product means that our notion of orthogonality is not symmetric in general and x being normal to y does not imply that y is normal to x.

Definition 4 (*Orthogonality*) Let V be a s.i.p. space. For $x, y \in V$ we say x is normal to y if $[y, x]_V = 0$.

This is equivalent to James orthogonality, namely for $x, y \in V$

$$[y, x]_V = 0 \Leftrightarrow \|x + \lambda y\|_V \geq \|x\|_V \quad \text{for all } \lambda \in \mathbb{R}$$

A vector $x \in V$ is normal to a subspace $U \subset V$ if x is normal to all $y \in U$. Thus the orthogonal complement of U is defined as

$$U^{\perp} = \{x_{\perp} \in V : [x, x_{\perp}]_V = 0 \, \forall x \in U\}$$

The relation to James orthogonality also helps to get a geometric understanding of what orthogonality means in a s.i.p. space. From Definition 4 it is immediately clear that x being normal to y means that the vector y is tangent to the ball $B(0, \|x\|)$ at the point x, where $B(0, \|x\|)$ is the ball of radius $\|x\|$ centred at the origin.

If the space is a uniformly convex Banach space there exists a unique orthogonal decomposition for every $x \in V$. This is because in a uniformly convex space there is a unique closest point in a closed linear subspace and one easily checks that this immediately leads to a unique orthogonal decomposition.

For uniformly convex, uniformly smooth Banach spaces we are also able to establish a Riesz representation theorem using the semi-inner product.

Theorem 1 *(Riesz representation theorem) Let V be a uniformly convex, uniformly smooth s.i.p. space. Then for every $f \in V^*$ there exists a unique vector $y \in V$ such that $f(x) = [x, y]_V$ for all $x \in V$ and $\|y\|_V = \|f\|_{V^*}$.*

Summarising the above results we see that a necessary structure to have a unique semi-inner product inducing the norm and allowing for a Riesz representation theorem is that the space is a uniformly convex and uniformly Fréchet differentiable Banach space. For simplicity we will be calling such spaces uniform.

Definition 5 (*Uniform Banach space*) We say a space V is uniform if it is a uniformly convex and uniformly Fréchet differentiable Banach space.

It is well known that the dual space of a uniform Banach space is itself uniform. Lastly we note that the duality mapping of a uniform Banach space is norm-to-norm continuous.The proof for this is standard and can be found in the appendix.

Proposition 2 *The duality map $* : \mathcal{B} \to \mathcal{B}^*, x \mapsto x^*$ is norm-to-norm continuous. In particular $[z, x + ty]_{\mathcal{B}} \to [z, x]_{\mathcal{B}}$ for all $x, y, z \in \mathcal{B}$ and $t \in \mathbb{R}$.*

Thus the dual map is a homeomorphism from $\mathcal{B}$ to $\mathcal{B}^*$ with the norm topologies. It is linear if and only if $\mathcal{B}$ is a Hilbert space.

3 Existence of representer theorems

The definitions and results of the previous section allow us to consider the regularised interpolation problem

$$\min \left\{ \Omega(f) \,:\, f \in \mathcal{B}, [f, x_i]_{\mathcal{B}} = y_i \,\forall i \in \mathbb{N}_m \right\} \tag{3}$$

where the domain $\mathcal{B}$ of the interpolation problem is a real uniform Banach space. This generalises the setting considered by Argyriou et al. in [1] where the case of a Hilbert space domain is considered. In that setting the linear representer theorem states that there exists a solution to the interpolation problem which is in the linear span of the data points. Our work, similarly as [12], hints that in its essence the representer theorem is a result about the dual space rather than the space itself. Since in a Hilbert space the dual element is the element itself this doesn't become apparent in this setting and we obtain a result in the space itself. As the duality map is nonlinear for any Banach space which is not Hilbert we need to adjust the formulation of the representer theorem. Namely the linear representer theorem in a uniform Banach space states that there exists a solution such that its dual element is in the linear span of the dual elements of the data points. This is made precise in the following definition which is the analogue of Argyriou, Micchelli and Pontil calling regularisers which always admit a linear representer theorem admissible.

Definition 6 (*Admissible regulariser*) We say a function $\Omega : \mathcal{B} \to \mathbb{R}$ is admissible if for any $m \in \mathbb{N}$ and any given data $\{(x_i, y_i) \,:\, i \in \mathbb{N}_m\} \subset \mathcal{B} \times Y$ such that the interpolation constraints can be satisfied the regularised interpolation problem Eq. (3) admits a solution f_0 such that its dual element is of the form

$$f_0^* = \sum_{i=1}^{m} c_i x_i^*$$

With this definition at hand it is now our goal to classify all admissible regularisers. It is well known that being a non-decreasing function of the norm on a Hilbert space is a sufficient condition for the regulariser to be admissible. By a Hahn–Banach argument similar as e.g. in Zhang and Zhang [19] this generalises to our case of uniform Banach spaces. Below we show that this condition is already almost necessary in the sense that admissible regularisers cannot be very far from being radially symmetric.

Theorem 2 *A function Ω is admissible if and only if it is of the form*

$$\Omega(f) = h([f, f]_{\mathcal{B}})$$

for some non-decreasing h whenever $\|f\| \neq r$ for $r \in \mathcal{R}$. Here $\mathcal{R}$ is an at most countable set of radii where h has a jump discontinuity. For any f with $\|f\| = r \in \mathcal{R}$ the value $\Omega(f)$ is only constrained by the monotonicity property, i.e. it has to lie in between $\lim_{t \nearrow r} h(t)$ and $\lim_{t \searrow r} h(t)$.

In other words, Ω is radially non-decreasing and radially symmetric except for at most countably many circular jump discontinuities. In those discontinuities the function value is only limited by its monotonicity property.

In [1] Argyriou et al. show that any admissible regulariser on a Hilbert space is non-decreasing in orthogonal directions. An analogous result is true for uniform Banach spaces but with orthogonality not being symmetric and our intuition gained from the equivalence with James orthogonality we see that in fact it is tangential directions in which the regulariser

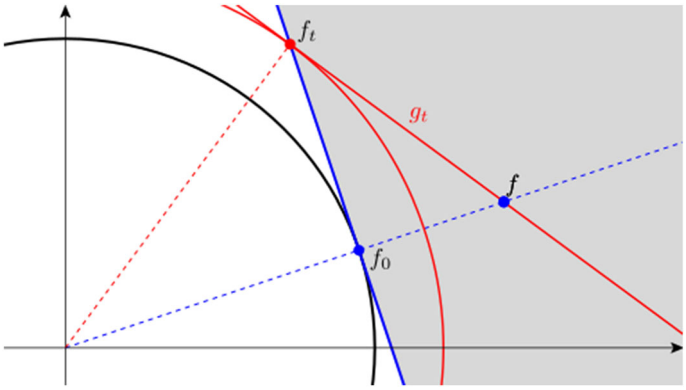

Fig. 1 We can extend the tangential bound to the ray $\lambda \cdot f_0$ by finding the point f_t along the tangent from where the tangent to f_t hits the desired point on the ray. Via the tangents to points along the ray the bound then extends to the shaded half space

is non-decreasing. This also becomes clear from the proofs in [1], in particular when proving radial symmetry.

Before we can prove the analogous result for uniform Banach spaces we need to show that we can extend this tangential bound considerably and a function that is non-decreasing in tangential directions is in fact non-decreasing in norm as is made precise in the following Lemma 1.

Lemma 1 *If $\Omega(f) \leq \Omega(f + f_T)$ for all $f, f_T \in \mathcal{B}$ such that $[f_T, f]_{\mathcal{B}} = 0$ then for any fixed $\hat{f}$ we have that $\Omega(\hat{f}) \leq \Omega(f)$ for all f such that $\|\hat{f}\| < \|f\|$.*

Proof *Part 1: (Bound Ω on the half space given by the tangent through $\hat{f}$)*

We start by showing that Ω is radially non-decreasing. Since it is non-decreasing along tangential directions this immediately gives the claimed bound for the entire half space given by the tangent through $\hat{f}$. The idea of the proof is to move out along a tangent until we can move back along another tangent to hit a given point along the ray $\lambda \cdot \hat{f}$ as shown in Fig. 1.

Fix some $\hat{f} \in \mathcal{B}$ and $1 < \lambda \in \mathbb{R}$ and set $f = \lambda \cdot \hat{f}$. We need to show that $\Omega(f) \geq \Omega(\hat{f})$. Let $f_T \in \mathcal{B}$ be such that $\left[f_T, \hat{f}\right]_{\mathcal{B}} = 0$ or equivalently $\|\hat{f} + t \cdot f_T\| > \|\hat{f}\|$ for all $t \neq 0$. Now let

$$f_t = \hat{f} + t \cdot f_T$$
$$g_t = f - f_t = (\lambda - 1) \cdot \hat{f} - t \cdot f_T$$

so that $f_t + g_t = f$. Note that by strict convexity and continuity of the norm $\|f_t\| = \|\hat{f} + t \cdot f_T\|$ is continuous and strictly increasing in t.

Now since $t \cdot f_T$ is the tangent through $\hat{f}$ and g_t points from f_t to f, for small t for which $\|f_t\| < \|f\|$ we must have that

$$\|f_t + s \cdot g_t\| > \|f_t\| \text{ for all } s \in (0, 1) \tag{4}$$

On the other hand for t big enough so that $\|f_t\| > \|f\|$ we thus must have

$$\|f_t + s \cdot g_t\| < \|f_t\| \text{ for } s \text{ small enough} \tag{5}$$

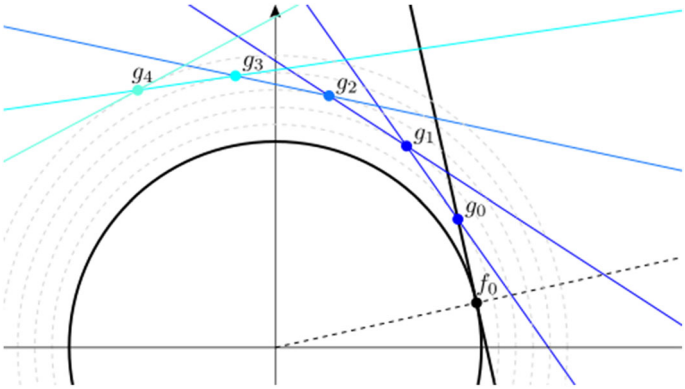

Fig. 2 By repeatedly taking steps along tangents we can move all the way around the circle

But we know that

$$\lim_{s\to 0}\frac{\|f_t + s\cdot g_t\| - \|f_t\|}{s} = \frac{[g_t, f_t]_{\mathcal{B}}}{\|f_t\|} = \frac{f_t^*(g_t)}{\|f_t\|}$$

and since the dual map is norm-to-norm continuous $\frac{f_t^*(g_t)}{\|f_t\|}$ is clearly continuous in t. By above discussion the expression is positive for small t and negative for large t so by the intermediate value theorem there exists t_0 such that

$$\frac{f_{t_0}^*(g_{t_0})}{\|f_{t_0}\|} = \frac{\left[g_{t_0}, f_{t_0}\right]_{\mathcal{B}}}{\|f_{t_0}\|} = 0$$

so that indeed $\left[g_{t_0}, f_{t_0}\right]_{\mathcal{B}} = 0$ and thus g_{t_0} is tangential to f_{t_0}. But this means that $\Omega(f) \geq \Omega(f_{t_0}) \geq \Omega(\hat{f})$ as claimed.

Hence we have the bound along the entire ray $\lambda \cdot \hat{f}$ for $1 < \lambda \in \mathbb{R}$ which extends along all tangents through those points to the half space given by the tangent through $\hat{f}$, i.e. the shaded region in Fig. 1.

Part 2: (Extend the bound around the circle)

Next we note that we can actually extend the bound further to apply all the way around the circle, namely $\Omega(f) \geq \Omega(\hat{f})$ for all f such that $\|f\| > \|\hat{f}\|$. This is done by considering $f_t = \hat{f} + t \cdot f_T$ as before but then instead of following the tangent into the half space just considered we follow the tangent in the opposite direction around the circle, as shown in Fig. 2. We fix another point along that tangent and repeat the process, moving around the circle. We claim that by making the step size along each tangent small enough we can this way move around the circle while staying arbitrarily close to it.

More precisely we need to show that the distance a step along a tangent takes us away from the circle decreases faster than the step along the tangent so that we move considerably further around the circle than away from it with each step, as shown in Fig. 3.

As stated in Eq. (2) let

$$\rho_{\mathcal{B}}(\delta) = \sup\left\{\frac{\|f+g\| + \|f-g\|}{2} - 1 \,:\, \|f\| = 1, \|g\| = \delta\right\}$$

be the modulus of smoothness of the space $\mathcal{B}$. For $f, f_T \in \mathcal{B}$ such that $[f_T, f]_{\mathcal{B}} = 0$, $\|f\| = 1$, $\|f_T\| = \delta$ we have that $\|f + t\cdot f_T\| > \|f\|$ for all $t \neq 0$ so in particular

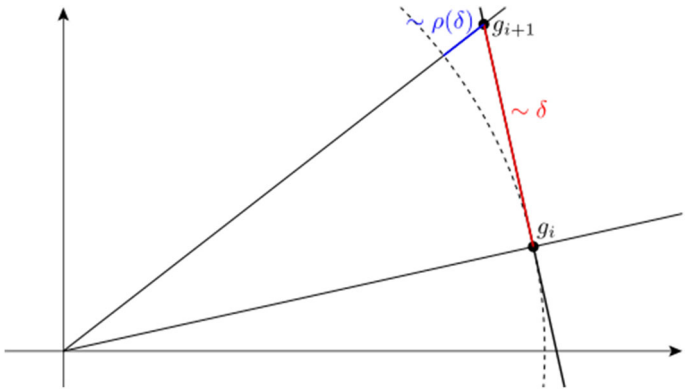

Fig. 3 When decreasing the step size along a tangent the step size away from the circle decreases significantly faster so that by making the steps along tangents small enough we can reach any point arbitrarily close to the circle

$\|f - f_T\| > \|f\|$. We thus easily see that

$$\begin{aligned}\|f + f_T\| &\leq 2 + 2\rho_{\mathcal{B}}(\delta) - \|f - f_T\| \\ &< 2 + 2\rho_{\mathcal{B}}(\delta) - \|f\| \\ &= 1 + 2\rho_{\mathcal{B}}(\delta)\end{aligned}$$

This means that for a step of order δ along a tangent, i.e. f_T of length δ, we take a step of order $\rho_{\mathcal{B}}(\delta)$ away from the circle. But since $\mathcal{B}$ is uniformly smooth we have that $\frac{\rho_{\mathcal{B}}(\delta)}{\delta} \to 0$ as $\delta \to 0$ proving that for small enough δ indeed the step away from the circle is significantly smaller than the step along the tangent as shown in Fig. 3.

Combining both arguments this proves that we can reach any point with norm greater than $\|\hat{f}\|$ from $\hat{f}$ only by moving along tangents giving the claimed bound. □

Having proved this lemma we are now in the position to prove that indeed any admissible regulariser on a uniform Banach space is non-decreasing in tangential directions. Note that the previous lemma will also play a crucial role in removing the differentiability assumption when establishing the closed form representation of the regulariser in Theorem 2.

Lemma 2 *A function Ω is admissible if and only if for every $f, f_T \in \mathcal{B}$ such that $[f_T, f]_{\mathcal{B}} = 0$ we have*

$$\Omega(f) \leq \Omega(f + f_T)$$

if and only if for any fixed $\hat{f}$ and all f such that $\|\hat{f}\| < \|f\|$ we have

$$\Omega(\hat{f}) \leq \Omega(f)$$

Proof *Part 1: (Ω admissible $\Rightarrow$ nondecreasing along tangential directions)*
Fix any $f \in \mathcal{B}$ and consider the regularised interpolation problem

$$\min\left\{\Omega(g) \,:\, g \in \mathcal{B}, [f, g]_{\mathcal{B}} = [f, f]_{\mathcal{B}}\right\}$$

As Ω is assumed to be admissible there exists a solution with dual element in $\text{span}\{f^*\}$ which by homogeneity of the dual map clearly is f itself. But if f_T is such that $[f_T, f]_{\mathcal{B}} = 0$

then $[f + f_T, f]_{\mathcal{B}} = [f, f]_{\mathcal{B}}$ so $f + f_T$ also satisfies the constraints and hence necessarily $\Omega(f + f_T) \geq \Omega(f)$ as claimed. The second claim follows immediately from Lemma 1.

Part 2: (Nondecreasing along tangential directions $\Rightarrow$ Ω admissible)

Conversely fix any data $\{(x_i, y_i) : i \in \mathbb{N}_m\} \subset \mathcal{B} \times Y$ such that the interpolation constraints can be satisfied. Let f_0 be a solution to the regularised interpolation problem. If $f_0^* \in \text{span}\{x_i^*\}$ we are done so assume it is not. We let

$$X^* = \text{span}\{x_i^*\} \subset \mathcal{B}^* \qquad X = \{x \in \mathcal{B} : x^* \in X^*\}$$

Further denote by $Z \subset \mathcal{B}$ the space corresponding to the orthogonal complement of X^* i.e.

$$Z = \{f_T \in \mathcal{B} : f_T^* \in (X^*)^{\perp}\} = \{f_T \in \mathcal{B} : [f_T, x_i]_{\mathcal{B}} = 0 \, \forall i \in \mathbb{N}_m\}$$

Thus $Z^* \cap X^* = \{0\}$ and by assumption $f_0^* \notin X^*$ and so also $\text{span}\{f_0^*\} \cap X^* = \{0\}$.

Now by definition we have that

$$Z = \bigcap_{i \in \mathbb{N}_m} \ker(x_i^*)$$

so the codimension of Z is m. Without loss of generality we can assume that not all y_i are zero as otherwise $f_0 = f_0^* = 0$ is a trivial solution in the span of the data points. Since not all y_i are zero $f_0 \notin Z$ and thus

$$\text{codim}(\text{span}\{f_0, Z\}) = m - 1$$

But since $X^* = \text{span}\{x_i^*\}$ and the dual map is a homeomorphism X is homeomorphic to a linear space of dimension m. This means that that $X \cap \text{span}\{f_0, Z\}$ is homeomorphic to a one-dimensional space and hence in particular contains a nonzero element.

Now fix such $0 \neq f \in X \cap \text{span}\{f_0, Z\}$. As we noted earlier f being nonzero means that $f \notin \text{span}\{f_0\}$ and $f \notin Z$. Thus $f = \lambda f_0 + \mu g$ for $\lambda, \mu \neq 0$, $g \in Z$. By homogeneity of the dual map $\lambda \cdot X = X$ and so

$$f \in X \cap \text{span}\{f_0, Z\} \Leftrightarrow \frac{1}{\lambda} f \in X \cap \text{span}\{f_0, Z\}$$

and thus

$$\frac{1}{\lambda} f = f_0 + \frac{\mu}{\lambda} g = f_0 + \widetilde{g} \in X \cap \text{span}\{f_0, Z\} \tag{6}$$

with $\widetilde{g} = \frac{\mu}{\lambda} g \in Z$.

This means we have constructed an $\overline{f_0} = f_0 + f_T$ with dual element in the span of the data points and $f_T \in Z$ which means by definition of Z that $\overline{f_0}$ satisfies the interpolation constraints. It remains to show that in fact $\overline{f_0}$ is in norm at most as large as f_0.

To this end note that for all $f_T \in Z$ by definition $\left[x^*, f_T^*\right]_{\mathcal{B}^*} = 0$ for all $x^* \in X^*$ and hence we see that for $\overline{f_0} = f_0 + f_T \in X$ we get that

$$\left[(f_0 + f_T)^*, f_T^*\right]_{\mathcal{B}^*} = [f_T, f_0 + f_T]_{\mathcal{B}} = 0$$

But by the equivalence with James orthogonality this means that $\|(f_0 + f_T) + t \cdot f_T\| > \|f_0 + f_T\|$ for all $t \neq 0$ or equivalently

$$\|f_0 + f_T\| = \min_{t \in \mathbb{R}} \|f_0 + t \cdot f_T\|$$

In particular $\|\overline{f_0}\| = \|f_0 + f_T\| < \|f_0 + 0 \cdot f_T\| = \|f_0\|$.

But by Lemma 1 we know that for a function which is non-decreasing along tangential directions is non-decreasing in norm so $\|\overline{f_0}\| < \|f_0\|$ implies that $\Omega(\overline{f_0}) \leq \Omega(f_0)$ and so we have found a solution with dual element in the span of the data points as claimed. □

Using those two results we can now give the proof that admissible regularisers are almost radially symmetric in the sense of Theorem 2.

Proof of Theorem 2 *Part 1: (Ω continuous in radial direction implies Ω radially symmetric)*
We now show that instead of differentiability, the assumption that Ω is continuous in radial direction is sufficient to conclude that it has to be radially symmetric. We prove this by contradiction. Assume Ω is admissible but not radially symmetric. Then there exists a radius r so that Ω is not constant on the circle with radius r and hence there are two points f and g so that, without loss of generality, $\Omega(f) > \Omega(g)$.

But then by Lemma 1 for all $1 < \lambda \in \mathbb{R}$ we have $\Omega(\lambda g) \geq \Omega(f)$ and thus as Ω non-negative and non-decreasing $|\Omega(\lambda g) - \Omega(g)| \geq |\Omega(f) - \Omega(g)| > 0$ contradicting radial continuity of Ω. Hence Ω has to be constant along every circle as claimed.

Part 2: (Radial mollification preserves being nondecreasing in tangential directions)
The observation in part 1 is useful as we can easily radially mollify a given Ω so that the property of being non-decreasing along tangential directions is preserved.

Indeed let ρ be a mollifier such that $\rho : \mathbb{R} \to [0, \infty)$ with support in $[-1, 0]$ and for each ray given by some $f_0 \in \mathcal{B}$ of unit norm, define the mollified regulariser by

$$\widetilde{\Omega}(sf_0) = \int_{\mathbb{R}} \rho(t)\Omega\left((s-t)f_0\right) \mathrm{d}t$$

We thus obtain a radially mollified regulariser on $\mathcal{B}$ given by

$$\begin{aligned}\widetilde{\Omega}(f) = \widetilde{\Omega}\left(\|f\|\frac{f}{\|f\|}\right) &= \int_{\mathbb{R}} \rho(t)\Omega\left((\|f\| - t)\frac{f}{\|f\|}\right) \mathrm{d}t \\ &= \int_{-1}^{0} \rho(t)\Omega\left((\|f\| - t)\frac{f}{\|f\|}\right) \mathrm{d}t\end{aligned}$$

We check that this function is still non-decreasing along tangential directions, i.e. we need to show that for f_T s.t. $[f_T, f]_{\mathcal{B}} = 0$ we still have

$$\begin{aligned}\widetilde{\Omega}(f + f_T) &= \int_{-1}^{0} \rho(t)\Omega\left((\|f + f_T\| - t)\frac{f + f_T}{\|f + f_T\|}\right) \mathrm{d}t \\ &\geq \int_{-1}^{0} \rho(t)\Omega\left((\|f\| - t)\frac{f}{\|f\|}\right) \mathrm{d}t = \widetilde{\Omega}(f) \qquad (7)\end{aligned}$$

Note that by Lemma 1 we have that $\Omega((\|f + f_T\| - t)\frac{f+f_T}{\|f+f_T\|}) \geq \Omega((\|f\| - t)\frac{f}{\|f\|})$ for all $t \in [-1, 0]$ if $\|(\|f + f_T\| - t)\frac{f+f_T}{\|f+f_T\|}\| \geq \|(\|f\| - t)\frac{f}{\|f\|}\|$ for all $t \in [-1, 0]$. But this is clear as it is equivalent to $|\|f + f_T\| - t\| \geq |\|f\| - t\|$. As t is non-positive we can drop the modulus to obtain that this happens if $\|f + f_T\| \geq \|f\|$ which is just James orthogonality and thus follows from the fact that $[f_T, f]_{\mathcal{B}} = 0$. This proves that the integral estimate Eq. (7) indeed holds and hence the radially mollified $\widetilde{\Omega}$ is indeed non-decreasing in tangential directions.

Part 3: (Ω is as claimed)

Putting these two observations together we obtain the result. By part 2 $\widetilde{\Omega}$ is of the form $\widetilde{\Omega}(f) = h\left([f, f]_{\mathcal{B}}\right)$ for some continuous, non-decreasing h. But if we consider Ω along any two distinct, fixed directions given by $f_1, f_2 \in \mathcal{B}$, $f_1 \neq f_2$, $\|f_1\| = \|f_2\| = 1$ as $\Omega(t \cdot f_i) = h_{f_i}\left([t \cdot f_i, t \cdot f_i]_{\mathcal{B}}\right)$ then the mollifications of both h_{f_1} and h_{f_2} must equal h so $h_{f_1} = h_{f_2}$ almost everywhere. Further by continuity of h they can only differ in points of discontinuity of h_{f_1} and h_{f_2}. As each h_{f_i} is a monotone function on the positive real line it can only have countably many points of discontinuity. Clearly as above bounds are only making statements about values outside a given circle and h is itself monotone, each h_{f_i} is free to attain any value within the monotonicity constraint in those points of discontinuity. This shows that Ω is of the claimed form. □

Remark 2 We see that everything we say about Ω in this section relies crucially on the observation that it being admissible is a statement about its behaviour along tangents as stated in Lemma 2. But there is in fact no tangent into the complex plane, i.e. for fixed $\hat{f}$ there is no tangent that intersects the ray $\{t \cdot e^{i\theta} \cdot \hat{f} \ : \ t \in \mathbb{R}\}$ for any θ. Likewise it is not possible to reach any point along said ray via an "out and back" argument as in part 1 of the proof of Lemma 1. For this reason it is currently not clear whether one can say anything about the situation in complex vector spaces.

4 The solution is determined by the space

First of all, while it has been known that for regularisers which are a strictly increasing function of the norm every solution is within the linear span of the data, the proofs in Sect. 3 show immediately that something stronger can be said. For a regularised interpolation problem with an admissible regulariser to have a solution which is not in the linear span of the data the regulariser must have a flat region and the solution then has to lie within the flat region.

But there is more to be said, in fact it turns out that for admissible regularisers the set of solutions in the linear span is independent of the regulariser.

In [12] Micchelli and Pontil consider the minimal norm interpolation problem

$$\inf\{\|x\|_X \ : \ x \in X, L_i(x) = y_i \ \forall i \in \mathbb{N}_m\}$$

where X is a Banach space and L_i are continuous linear functionals on X. Hence this agrees with Eq. (3) for $h(t) = \sqrt{t}$ i.e. $\Omega(f) = ([x, x]_{\mathcal{B}})^{\frac{1}{2}}$ and $X = \mathcal{B}$ a uniformly convex, uniformly smooth Banach space, giving the minimal norm interpolation problem

$$\min\{\|f\|_{\mathcal{B}} \ : \ f \in \mathcal{B}, x_i^*(f) = [f, x_i]_{\mathcal{B}} = y_i \ \forall i \in \mathbb{N}_m\} \tag{8}$$

This leads to the following result.

Theorem 3 *Let Ω be admissible. Then any f_0 which is such that $f_0^* = \sum_{i=1}^m c_i x_i^*$ is a solution of* Eq. (3) *if and only if it is a solution of* Eq. (8).

The proof of this result relies on the following result which was proved by Micchelli and Pontil in [12].

Proposition 3 (Theorem 1 in [12]) *f_0 is a solution of* Eq. (8) *if and only if it satisfies the constraints $x_i^*(f_0) = y_i$ and there is a linear combination of the continuous linear functionals defining the problem which peaks at f_0, i.e. there exists $(c_1, \ldots, c_m) \in \mathbb{R}^m$ such that*

$$\sum_{i=1}^{m} c_i x_i^*(f_0) = \left\| \sum_{i=1}^{m} c_i x_i^* \right\|_{\mathcal{B}^*} \cdot \|f_0\|_{\mathcal{B}}$$

Using this result it is easy to proof Theorem 3.

Proof of Theorem 3 *Part 1: (A solution of Eq. (3) is a solution of Eq. (8))*

Assume that f_0 is a solution of Eq. (3) such that $f_0^* = \sum_{i=1}^m c_i x_i^*$. Then trivially f_0 satisfies the interpolation constraints and by definition

$$f_0^*(f_0) = [f_0, f_0]_{\mathcal{B}} = \|f_0\|_{\mathcal{B}}^2 = \|f_0^*\|_{\mathcal{B}^*} \cdot \|f_0\|_{\mathcal{B}}$$

so f_0^*, which is a linear combination of the continuous linear problems defining the problem, peaks at f_0. Thus by Proposition 3 f_0 is a solution of Eq. (8).

Part 2: (A solution of Eq. (8) is a solution of Eq. (3))

Assume f_0 is a solution of Eq. (8). Then by Eq. (3) there exists $(c_1, \ldots, c_m) \in \mathbb{R}^m$ such that the functional $\sum_{i=1}^m c_i x_i^*$ peaks at f_0, i.e.

$$\sum_{i=1}^{m} c_i x_i^*(f_0) = \left\| \sum_{i=1}^{m} c_i x_i^* \right\|_{\mathcal{B}^*} \cdot \|f_0\|_{\mathcal{B}}$$

But then for any $g \in Z = \{f \in \mathcal{B} : x_i^*(f) = [f, x_i]_{\mathcal{B}} = 0, \forall i = 1, \ldots, m\}$ we have that

$$\left\| \sum_{i=1}^{m} c_i x_i^* \right\|_{\mathcal{B}^*} \cdot \|f_0\|_{\mathcal{B}} = \sum_{i=1}^{m} c_i x_i^*(f_0) = \sum_{i=1}^{m} c_i x_i^*(f_0 + g) < \left\| \sum_{i=1}^{m} c_i x_i^* \right\|_{\mathcal{B}^*} \cdot \|f_0 + g\|_{\mathcal{B}}$$

where the last inequality is strict because $\sum_{i=1}^m c_i x_i^*$ peaks at f_0 and by strict convexity it peaks at a unique point. But this inequality shows that

$$\|f_0\|_{\mathcal{B}} < \|f_0 + g\|_{\mathcal{B}}$$

for all $g \in Z$ and thus as Ω is admissible also

$$\Omega(f_0) \leq \Omega(f_0 + g)$$

and f_0 is a solution of Eq. (3). □

This result shows that any admissible regulariser on a uniformly convex and uniformly smooth Banach space has a unique solution in the linear span of the data and the solution is the same for every admissible regulariser. This in particular means that it is the choice of the function space, and only the choice of the space, which determines the solution of the problem. We are thus free to work with whichever regulariser is most convenient in application. Computationally in many cases this is likely going to be $\frac{1}{2}\|\cdot\|^2$, for theoretical results other regularisers may be more suitable, such as in the afore mentioned paper [12] which heavily relies on a duality between the norm of the space and its continuous linear functionals.

A Appendix

Proof of Proposition 2 We begin by showing norm-to-weak continuity and subsequently extend it to norm-to-norm continuity.

Since $\mathcal{B}$ is reflexive the weak and weak$*$ topologies on $\mathcal{B}^*$ coincide, so we need to show that if $x_n \to x$ in norm then $x_n^*(y) \to x^*(y)$ for all $y \in \mathcal{B}$.

Now as $\|x_n^*\|_{\mathcal{B}^*} = \|x_n\|_{\mathcal{B}}$ the sequence (x_n^*) is bounded so it has a weakly$*$ convergent subsequence $x_{n_k}^* \overset{*}{\rightharpoonup} \overline{x}^*$. By [2] proposition 3.13 (iv) we then have

$$x_{n_k}^*\left(x_{n_k}\right) \underset{k\to\infty}{\longrightarrow} \overline{x}^*(x)$$

But $x_{n_k}^*(x_{n_k}) = \|x_{n_k}\|^2 \to \|x\|^2$ and so $\overline{x}^*(x) = \|x\|^2$. By [2] proposition 3.13 (iii) we further know that $\|\overline{x}^*\| \leq \liminf \|x_{n_k}^*\| = \|x\|$. By strict convexity there is a unique element with those two properties and hence $\overline{x}^* = x^*$.

Note that this means that for any subsequence there exists a further subsequence converging to a unique limit. This means that in fact the entire sequence converges to this unique limit. Hence indeed $x_n^* \rightharpoonup x^*$ as claimed.

Having established norm-to-weak continuity one can easily extend it to norm-to-norm continuity using [2] proposition 3.32. Since $\limsup \|x_n^*\|_{\mathcal{B}^*} = \|x\|_{\mathcal{B}} = \|x^*\|_{\mathcal{B}^*}$ all the assumptions of proposition 3.32 in [2] are satisfied and so indeed $x_n^* \to x^*$ in norm. □

References

1. Argyriou, A., Micchelli, C.A., Pontil, M.: When is there a representer theorem? vector versus matrix regularizers. J. Mach. Learn. Res. **10**, 2507–2529 (2009)
2. Brezis, H.: Functional Analysis, Sobolev Spaces and Partial Differential Equations. Universitext, 1st edn. Springer, New York (2011)
3. Cox, D.D., O'Sullivan, F.: Asymptotic analysis of penalized likelihood and related estimators. Ann. Stat. **18**(4), 1676–1695 (1990). https://doi.org/10.1214/aos/1176347872
4. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Am. Math. Soc. **39**(1), 1–49 (2001)
5. Dragomir, S.S.: Semi-inner Products and Applications. Nova Science Publishers, Hauppauge (2004)
6. Giles, J.R.: Classes of semi-inner-product spaces. Trans. Am. Math. Soc. **129**(3), 436–446 (1967)
7. James, R.C.: Orthogonality and linear functionals in normed linear spaces. Trans. Am. Math. Soc. **61**(2), 265–292 (1947)
8. Kimeldorf, G., Wahba, G.: Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. **33**(1), 82–95 (1971). https://doi.org/10.1016/0022-247X(71)90184-3
9. Köthe, G.: Topological Vectorspaces I, Grundlehren der Mathematischen Wissenschaften, vol. 159. Springer, Berlin (1983)
10. Lindenstrauss, J., Tzafriri, L.: Classical Banach Spaces II: Function Spaces, Ergebnisse der Mathematik und ihrer Grenzgebiete, vol. 97. Springer, Berlin (1979)
11. Lumer, G.: Semi-inner-product spaces. Trans. Am. Math. Soc. **100**(1), 29–43 (1961)
12. Micchelli, C.A., Pontil, M.: A function representation for learning in banach spaces. In: Shawe-Taylor, J., Singer, Y. (eds.) Learning Theory, pp. 255–269. Springer, Berlin (2004)
13. Micchelli, C.A., Pontil, M.: Learning the kernel function via regularization. J. Mach. Lear. Res. **6**, 1099–1125 (2005)
14. Schölkopf, B., Herbrich, R., Smola, A.J.: A generalized representer theorem. In: Helmbold, D., Williamson, B. (eds.) Computational Learning Theory, pp. 416–426. Springer, Berlin (2001)
15. Schölkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)
16. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004). https://doi.org/10.1017/CBO9780511809682
17. Smola, J.A., Schölkopf, B.: On a kernel-based method for pattern recognition, regression, approximation, and operator inversion. Algorithmica **22**(1), 211–231 (1998). https://doi.org/10.1007/PL00013831

18. Zhang, H., Xu, Y., Zhang, J.: reproducing kernel banach spaces for machine learning. J. Mach. Learn. Res. **10**, 2741–2775 (2009)
19. Zhang, H., Zhang, J.: Regularized learning in banach spaces as an optimization problem: representer theorems. J. Global Optim. **54**(2), 235–250 (2012). https://doi.org/10.1007/s10898-010-9575-z