Approximate Representer Theorems in Non-reflexive Banach Spaces

Kevin Schlegel SCHLEGEL@MATHS.OX.AC.UK
Mathematical Institute
University of Oxford
Andrew Wiles Building, Radcliffe Observatory Quarter
Woodstock Road, Oxford, OX2 6GG, UK

Editors: Aryeh Kontorovich and Gergely Neu

Abstract

The representer theorem is one of the most important mathematical foundations for regularised learning and kernel methods. Classical formulations of the theorem state sufficient conditions under which a regularisation problem on a Hilbert space admits a solution in the subspace spanned by the representers of the data points. This turns the problem into an equivalent optimisation problem in a finite dimensional space, making it computationally tractable. Moreover, Banach space methods for learning have been receiving more and more attention. Considering the representer theorem in Banach spaces is hence of increasing importance. Recently the question of the necessary condition for a representer theorem to hold in Hilbert spaces and certain Banach spaces has been considered. It has been shown that a classical representer theorem cannot exist in general in non-reflexive Banach spaces. In this paper we propose a notion of approximate solutions and approximate representer theorem to overcome this problem. We show that for these notions we can indeed extend the previous results to obtain a unified theory for the existence of representer theorems in any general Banach spaces, in particular including l^1-type spaces. We give a precise characterisation when a regulariser admits a classical representer theorem and when only an approximate representer theorem is possible.

Keywords: representer theorem, approximate representer theorem, regularised interpolation, regularisation

1. Introduction

It is a common approach in learning theory to formulate a problem of estimating functions from input and output data as an optimisation problem. Most commonly used is regularisation, in particular *Tikhonov regularisation* where we consider an optimisation problem of the form

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

where $\mathcal{H}$ is a Hilbert space with inner product $\langle\cdot,\cdot\rangle$, $\{(x_i, y_i) \,:\, i = 1, \dots, m\} \subset \mathcal{H} \times Y$ 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*. The representer theorem is one of the most important mathematical foundations for such regularised learning problems. It states that under certain conditions on the regulariser the optimisation problem has a solution in the finite dimensional subspace spanned by the data points $x_i \in \mathcal{H}$, making it computationally tractable.
While these problems are well understood in Hilbert spaces, Banach space methods have been receiving more and more attention in machine learning for various reasons, such as e.g. the richer

geometric variety in comparison to Hilbert spaces, and certain desirable properties of Banach space norms such as the l^1 norm inducing sparsity of the solution vector. We are thus going to consider the more general regularisation problem

$$\inf \{\mathcal{E}((L_i(f), y_i)_{i=1}^m) + \lambda\Omega(f) \ : \ f \in \mathcal{B}\} \tag{1}$$

where $\mathcal{B}$ is a Banach space and the L_i are continuous linear functionals on $\mathcal{B}$. This framework is general enough to include all classical Hilbert space techniques such as least squares, SVMs and Kernel PCA but also their counterparts in reproducing Kernel Banach spaces introduced by Zhang et al. (2009); Zhang and Zhang (2012). Furthermore it includes popular regularisation frameworks such as lasso (Tibshirani, 1996) and its variants, e.g. square-root lasso (Belloni et al., 2011).
Moreover, while the L_i could be simple point evaluations $L_i(f) = f(x_i)$, phrasing the problem using general linear functionals has the advantage of including other interesting cases such as local averages of the form $L(f) = \int_{\mathcal{B}} f(x) \, \mathrm{d}P(x)$ where P is a probability measure on $\mathcal{B}$.
With the data given as functionals in the dual space $\mathcal{B}^*$ it is clear that the representer theorem in Banach spaces in fact has to be rooted in the dual space rather than the space itself, as can also be seen in the work by Micchelli and Pontil (2004); Zhang et al. (2009); Zhang and Zhang (2012) and our earlier work (Schlegel, 2019a,b). It turns out that the representer theorem is closely related to the properties of the duality mapping

$$J : \mathcal{B} \to 2^{\mathcal{B}^*} \qquad J(f) = \{L \in \mathcal{B}^* \ : \ L(f) = \|L\| \cdot \|f\|, \|L\| = \|f\|\} \tag{2}$$

This does not become apparent in Hilbert spaces as the duality mapping is the identity. Before we discuss this in more detail we introduce another common assumption to simplify the problem. While in applications we are often interested in regularisation problems of the form (1), Argyriou et al. (2009) and our earlier work (Schlegel, 2019a,b) show that in Hilbert spaces and reflexive Banach spaces under very mild conditions (1) admits a representer theorem if and only if the regularised interpolation problem

$$\inf \{\Omega(f) \ : \ f \in \mathcal{B}, L_i(f) = y_i \, \forall i = 1, \dots, m\} \tag{3}$$

admits a representer theorem. Here by admitting a representer theorem we mean that a solution determined by a linear combination of the data always exists whenever the constraints can be satisfied. In this case we will call Ω admissible. The connection between regularisation and regularised interpolation is not surprising as the regularisation problem is more general and one obtains a regularised interpolation problem in the limit as the regularisation parameter goes to zero. Thus we can, and will, focus our attention on the regularised interpolation problem which is more convenient to study. The precise statement of this fact with the required conditions and its proof for general Banach spaces are presented in appendix C, as the proof only requires a few technical modifications from the one presented in our previous work (Schlegel, 2019b). Note that in fact any representer theorem for regularised interpolation holds for any regularisation problem with the same regulariser without any further assumptions. Thus any representer theorem for regularised interpolation proved below is immediately valid for regularisation problems of the form (1).

It is well known that a regulariser is admissible if it is a nondecreasing function of the Hilbert space norm. By a Hahn-Banach argument as e.g. by Zhang and Zhang (2012) the same is true for reflexive Banach spaces. Argyriou et al. (2009) showed that this condition is also necessary for

differentiable regularisers on Hilbert spaces. Dinuzzo and Schölkopf (2012) extend this result to lower semicontinuous regularisers on Hilbert spaces. Recently we removed the regularity assumptions on the regulariser (Schlegel, 2019a), proving that an admissible regulariser cannot be very far from being a nondecreasing function of the norm, in a sense made precise in the paper. Moreover the results apply to uniformly convex, uniformly smooth Banach spaces, extending the theory to a wide range of Banach spaces. More recently we further showed that in fact the same necessary and sufficient condition holds for reflexive Banach spaces (Schlegel, 2019b). It is interesting, and instructive for this work, to note that our previous work clearly highlights the relationship between the properties of the duality mapping (2) and the formulation of the representer theorem. To account for the nonlinearity of the duality mapping in uniform Banach spaces (Schlegel, 2019a) we defined a regulariser to be admissible if there exists a solution f_0 to (3) with dual element in the linear span of the linear functionals defining the interpolation problem, i.e. $\sum c_i L_i = J(f_0)$. To account for the duality mapping not being univocal in Banach spaces which are not smooth (Schlegel, 2019b) this equality turns into an inclusion, i.e. $\sum c_i L_i \in J(f_0)$.
Moreover, by giving a counterexample (Schlegel, 2019b) we showed that it is not possible in general to obtain a representer theorem in this sense if the space is not reflexive. This is unfortunate since l^1, which is frequently used in applications, is not reflexive. Only the finite dimensional l^1_n is reflexive.

To overcome this issue we propose to follow the approach of reflecting the properties of the duality mapping in the formulation of the representer theorem. The reason why a representer theorem in the above sense cannot exist in a non-reflexive Banach space is that the duality mapping is not surjective. This means that we cannot expect to find a solution with dual element in the linear span of the linear functionals defining the optimisation problem as described above. But Bishop and Phelps (1961) prove that every Banach space is subreflexive, i.e. the image of the duality mapping J is norm-dense in $\mathcal{B}^*$. Thus we can hope to be able to get arbitrarily close to $\operatorname{span}\{L_i\}$, i.e. $\operatorname{dist}(J(f_0), \operatorname{span}\{L_i\}) < \varepsilon$. This leads to a notion of *approximate solution* and *approximate representer theorem* which we are going to introduce in this paper. We are going to show that for this weaker concept of solutions we can indeed obtain the immediate generalisations of the results of Argyriou et al. (2009) and our earlier work(Schlegel, 2019a,b). This provides a unified theory for the existence of representer theorems in arbitrary Banach spaces, in particular including l^1-type spaces which are very frequently used in applications. More precisely we are going to prove the following theorem.
Theorem A function $\Omega : \mathcal{B} \to \mathbb{R}$ is admissible if and only if viewed as a function $\overline{\Omega}$ of the faces F of the norm ball in $\mathcal{B}$, $\overline{\Omega}(F) = \min_{f \in F} \Omega(f)$ it is of the form

$$\overline{\Omega}(F) = h(\|f\|_{\mathcal{B}} \,:\, f \in F)$$

for some nondecreasing $h : [0, \infty) \to \mathbb{R}$ whenever $\|f\|_{\mathcal{B}} \neq r$ 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 $\|f\|_{\mathcal{B}} = r \in \mathcal{R}$ the value $\overline{\Omega}(F)$ is only constrained by the monotonicity property., i.e. it has to lie in between $\lim_{t \nearrow r} h(t)$ and $\lim_{t \searrow r} h(t)$.

2. Approximate representer theorems

We let $\mathcal{B}$ be an arbitrary Banach space with duality mapping (2) and consider the regularised interpolation problem (3). There are two main differences to the setting of reflexive Banach spaces that need to be overcome.
Firstly, Argyriou et al. (2009) and our earlier work (Schlegel, 2019a,b) assume that a minimiser of (3) always exists, whenever the constraints can be satisfied. But in a non-reflexive Banach space we cannot expect the minimum of (3) to always be attained. More precisely, if we denote by Z the subspace

$$Z = \bigcap_{i=1}^{m} \ker(L_i)$$

it is easy to see that solving the minimal norm interpolation problem, i.e. $\Omega(f) = \|f\|_{\mathcal{B}}$ in (3), is equivalent to minimising $\inf\{\|\overline{f} + f_T\|_{\mathcal{B}} \,:\, f_T \in Z\}$ where $\overline{f} \in \mathcal{B}$ is any function satisfying the interpolation constraints. In other words the infimum of the minimal norm interpolation is attained at f_0 if and only if the distance of 0 to the affine space $\overline{f} + Z$ is attained at $f_0 \in \overline{f} + Z$. It is well known that such f_0 does not always exist if $\mathcal{B}$ is not reflexive. Now different values of the y_i correspond to different shifts $\overline{f}$ of Z so that if the distance is attained, it happens at different points. Thus a solution to the minimal norm interpolation always exist for any given data exactly when Z is proximinal.

Definition 1 (Proximinal Subspace) *Let V be a real normed vector space and $W \subset V$ a closed subspace of V. We say W is proximinal if the distance from any point in V to W is attained, i.e. for every $x \in V$ there is a $y \in W$ such that $\|x - y\|_V = \operatorname{dist}(x, W)$.*

Following this intuition, instead of assuming a solution to the regularised interpolation always exists when the constraints can be satisfied, we will assume that a solution to eq. (3) always exists if Z is proximinal. While in a reflexive space every closed linear subspace is proximinal the question becomes a lot more delicate in non-reflexive spaces and there are spaces which contain in a sense very few proximinal subspaces, e.g. no proximinal subspace of finite codimension greater than one (Read, 2018; Kadets et al., 2018). Conditions for when a subspace is proximinal are still an active area of research. Some good references for what is known include Singer (1970); Holmes (1975); Conway (1994). We state two results which are of particular relevance to our work in appendix D.

Secondly the duality mapping J is surjective if and only if the space is reflexive. Thus $\operatorname{span} L_i$ may not be entirely contained in the image of J, or as we illustrate in our earlier work (Schlegel, 2019b), possibly even $J(\mathcal{B}) \cap \operatorname{span}\{L_i\} = \emptyset$. We thus cannot hope for a solution with a dual element in the linear span of the functionals, i.e. $J(f_0) \cap \operatorname{span}\{L_i\} \neq \emptyset$. But since every Banach space is subreflexive (Bishop and Phelps, 1961), which means the image of the duality mapping is norm dense in the dual space, we might expect to be able to get arbitrarily close to the linear span, i.e. $\operatorname{dist}(J(f_0), \operatorname{span}\{L_i\}) < \varepsilon$.
Combining both, approximation of the infimum in (3) and norm-closeness to the span of the L_i leads to the afore mentioned notion of *approximate solution* and *approximate representer theorem* and hence a new definition of admissibility of regularisers.

Definition 2 (Admissible Regularizer) *We say a function $\Omega : \mathcal{B} \to \mathbb{R}$ is admissible if for any $m \in \mathbb{N}$ and any given data $\{L_1, \ldots, L_m\} \subset \mathcal{B}^*$ and $\{y_1, \ldots, y_m\} \subset Y$ such that the interpolation constraints can be satisfied the regularised interpolation problem eq. (3) either*

(i) Admits a solution f_0 *such that there exist coefficients* $\{c_1, \dots, c_m\} \subset \mathbb{R}$ *such that*

$$\hat{L} = \sum_{i=1}^{m} c_i L_i \in J(f_0) \qquad \text{if } Z = \bigcap_{i \in \mathbb{N}_m} \ker(L_i) \text{ is proximinal}$$

(ii) Or otherwise admits for every $\varepsilon > 0$ *an approximate solution, which satisfies the interpolation constraints and approximates the infimum. More precisely there exists an* f_0^ε *such that*

$$\Omega(f_0^\varepsilon) \leq \inf \{\Omega(f) \,:\, f \in \mathcal{B},\, L_i(f) = y_i \,\forall i = 1, \dots, m\} + \varepsilon$$

and there exist $\hat{L} \in J(f_0^\varepsilon)$ *and coefficients* $\{c_1, \dots, c_m\} \subset \mathbb{R}$ *such that*

$$\|\hat{L} - \sum_{i=1}^{m} c_i L_i\|_{\mathcal{B}^*} < \varepsilon$$

2.1. Existence of approximate representer theorems

We now show that with this notion of admissibility we can indeed obtain the analogue of the results of Argyriou et al. (2009) and our previous work (Schlegel, 2019a,b) that being in a sense nondecreasing along tangents is a necessary and sufficient condition for admissibility. As became apparent in the case of reflexive Banach spaces (Schlegel, 2019b), when the space is not strictly convex we can only hope to characterise the regulariser as a function of the faces of the norm ball. Recall that an exposed face F of the norm ball $B_r \subset \mathcal{B}$ is a non-empty subset of B_r such that $F = \left\{x \in B_r \,:\, L(x) = \sup_{y \in B_r} L(y)\right\}$ for some $L \in \mathcal{B}^*$ (for more details see e.g. Hiriart-Urruty and Lemaréchal (2001); Aizpuru and Garca-Pacheco (2005)).

Lemma 3 *A function* $\Omega : \mathcal{B} \to \mathbb{R}$ *is admissible if and only if for every exposed face of the norm ball,* Ω *attains its minimum in at least one point and for every* f *in the face where the minimum is attained and every* $L \in J(f)$ *exposing the face and every* $f_T \in \ker(L)$ *we have*

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

Definition 4 *We are going to refer to the points lemma 3 applies to as admissible points.*

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

$$\min \left\{\Omega(g) \,:\, g \in \mathcal{B}, L(g) = L(f) = \|f\|^2\right\}$$

Conway (1994, Prop. 4.7) proves that $\ker(L)$ is proximinal if and only if L is in the image of the duality mapping. As Ω is assumed to be admissible we thus are in the case (i) of definition 2 and there exists a solution f_0 such that $c \cdot L \in J(f_0)$. We can thus argue exactly as in the case of a reflexive space, we include the short proof for completeness.
If there does not exist $g \in \mathcal{B}$ such that $g \neq f$ and $L \in J(g)$ then the solution can only be f itself. Then for any $f_T \in \ker(L)$ also $L(f + f_T) = L(f) = \|f\|^2$ and $f + f_T$ also satisfies the constraints and hence necessarily $\Omega(f + f_T) \geq \Omega(f)$.

But if there exists $f \neq g \in \mathcal{B}$ such that $L \in J(g)$ then f is contained in an exposed face. We have no way of making a statement about how $\Omega(f)$ and $\Omega(g)$ compare, all we can say is that Ω attains its minimum in at least one point within this face. It is clear that for any of those minimal points the above discussion is true for L exposing the face so that we obtain the tangential bound.

Part 2: *Nondecreasing along tangential directions* $\Rightarrow$ Ω *admissible*
Fix any data $(L_i, y_i) \in \mathcal{B}^* \times Y$ for $i = 1, \dots, m$ such that the constraints can be satisfied. We now have the two cases of definition 2 to consider.
Case 1: If Z is proximinal then by assumption there exists a solution f_0 of the regularised interpolation problem and we are looking for a solution in the sense of definition 2 (i). We need to show that if f_0 is not a solution in this sense then there exists $f_T \in Z$ such that $\operatorname{span}\{L_i\} \cap J(f_0 + f_T) \neq \emptyset$. It turns out that the proof for reflexive Banach spaces (Schlegel, 2019b) remains valid, and understanding its main ideas is instructive for dealing with the second case. The proof is based on minimising the functional

$$F_{f_0} \colon \mathcal{B} \to \mathbb{R}, \quad F_{f_0}(f) = \int_0^{\|f - f_0\|} t \, \mathrm{d}t = \frac{\|f - f_0\|^2}{2} \tag{4}$$

over the subspace Z. Reflexivity of $\mathcal{B}$ is only used to ensure reflexivity of Z and thus the existence of a minimiser on Z of the continuous, convex and coercive functional F_{f_0}. But this minimiser clearly exists exactly when the metric projection of f_0 onto Z exists, thus by definition when Z is proximinal. One can check that with the existence of a minimiser of F_{f_0} on Z the rest of the proof for reflexive spaces remains valid. Again we include the remaining short argument for completeness.
For the minimiser $f_T \in Z$ of F_{f_0} we have that there exists $L \in J(f_0 + f_T)$ such that $L\big|_Z \equiv 0$. Since $\operatorname{span}\{L_i\} = Z^\perp$ this in turn means that $L \in \operatorname{span}\{L_i\}$. It remains to show that $\hat{f}$ indeed minimises Ω. But for $L \in J(f_0 + f_T) \cap Z^\perp$ we have $-f_T \in \ker(L)$. If $f_0 + f_T$ is exposed by L then the tangential bound applies and

$$\Omega(f_0 + f_T) \leq \Omega((f_0 + f_T) + (-f_T)) = \Omega(f_0)$$

so $f_0 + f_T$ is a solution of the regularised interpolation problem.
If on the other hand $f_0 + f_T$ is not exposed by L, then it is contained in a face exposed by L. But then for any $\overline{f_T} \in \mathcal{B}$ such that $(f_0 + f_T) + \overline{f_T}$ is still contained in this face we have that $L \in J(f_0 + f_T + \overline{f_T})$ and $\overline{f_T} \in \ker(L)$ so that $f_0 + f_T + \overline{f_T}$ satisfies the interpolation constraints. We can thus choose $\overline{f_T}$ such that $f_0 + f_T + \overline{f_T}$ is a minimum of Ω in the face and the tangential bound applies to it. Thus similarly to before

$$\Omega(f_0 + f_T + \overline{f_T}) \leq \Omega((f_0 + f_T + \overline{f_T}) + (-f_T - \overline{f_T})) = \Omega(f_0)$$

and $f_0 + f_T + \overline{f_T}$ is a solution of the regularised interpolation problem of the desired form.

Case 2: If Z is not proximinal the existence of a minimiser of (3) is not guaranteed. But for every $\varepsilon > 0$ there exists f_0^ε which ε-almost attains the infimum. We need to show that if any such f_0^ε is not a solution in the sense of definition 2 (ii) then there exists $f_T^\varepsilon \in Z$ such that $f_0^\varepsilon + f_T^\varepsilon$ is, i.e. $\operatorname{dist}(J(f_0^\varepsilon + f_T^\varepsilon), \operatorname{span}\{L_i\}) < \varepsilon$.
Following the approach from case 1 this means we are looking for f_T^ε with $L \in J(f_0^\varepsilon + f_T^\varepsilon)$ such

that $\|L\big|_Z\| < \varepsilon$. We are again going to consider the functional $F_{f_0^\varepsilon}$ as defined in (4), for simplicity denoted by F below. With Z not proximinal we do not get a minimiser of $F\big|_Z$ anymore. But by Ekelands variational principle (Ekeland, 1974) for every $\varepsilon > 0$ there exists an approximate minimiser $f_T^\varepsilon \in Z$ such that

$$F(f_T^\varepsilon) \leq \inf_{f \in Z} F(f) + \varepsilon \quad \text{and} \quad F(f_T^\varepsilon) - F(g) < \varepsilon \cdot \|f_T^\varepsilon - g\| \quad \forall f_T^\varepsilon \neq g \in Z \tag{5}$$

Choosing $g = f_T^\varepsilon + th$ for $h \in Z$ in eq. (5) we obtain a bound on the directional derivative of F

$$F'(f_T^\varepsilon, h) = \lim_{t \searrow 0} \frac{F(f_T^\varepsilon + th) - F(f_T^\varepsilon)}{t} > -\varepsilon \cdot \|h\| \tag{6}$$

By a corollary of the Sandwich theorem by Simons (Appendix A theorem 9) there exists $L \in Z^*$ such that $L \in \partial F\big|_Z(f_T^\varepsilon)$ which is necessary to extend it to $L \in J(f_0^\varepsilon + f_T^\varepsilon)$. Moreover

$$\inf_{h \in B} L(h) = \inf_{h \in B} F'(f_T^\varepsilon, h) \overset{(6)}{>} -\varepsilon \cdot \|h\|$$

which implies that $\|L\|_{Z^*} < \varepsilon$. By a Hahn-Banach argument this functional can be extended to an $L \in J(f_0^\varepsilon + f_T^\varepsilon)$ such that $\operatorname{dist}(L, \operatorname{span}\{L_i\}) < \varepsilon$. The construction is not difficult but technical and given in appendix B. Thus $f_0^\varepsilon + f_T^\varepsilon$ satisfies the assumptions of definition 2 (ii).
The fact that $f_0^\varepsilon + f_T^\varepsilon$ indeed minimises Ω follows in the same way as in case 1. If $f_0^\varepsilon + f_T^\varepsilon$ is an exposed point it satisfies the tangential bound and thus

$$\Omega(f_0^\varepsilon + f_T^\varepsilon) \leq \Omega((f_0^\varepsilon + f_T^\varepsilon) + (-f_T^\varepsilon)) = \Omega(f_0^\varepsilon)$$

If $f_0^\varepsilon + f_T^\varepsilon$ is not exposed it is contained in a face and just as before we can add another $\overline{f_T} \in Z$ so that the sum is within the face and

$$\Omega(f_0^\varepsilon + f_T^\varepsilon + \overline{f_T}) \leq \Omega((f_0^\varepsilon + f_T^\varepsilon + \overline{f_T}) + (-f_T^\varepsilon - \overline{f_T})) = \Omega(f_0^\varepsilon)$$

Since this new point is in the same face it has the same L as a dual element and is thus an admissible solution. ■

2.2. Uniformly non-rotund spaces

Argyriou et al. (2009) and our earlier work (Schlegel, 2019a,b) prove that being tangentially non-decreasing is equivalent to being (almost) radially symmetric. We now want to prove the corresponding geometric interpretation of lemma 3. As we argued in the case of reflexive Banach spaces (Schlegel, 2019b), the geometric variety of arbitrary Banach spaces does not allow for a general, closed form result of this kind. It is clear that our arguments for strictly convex spaces remain true even without reflexivity, but with the most important examples of non-reflexive spaces being l^1 and L^1 we are going to introduce and consider a class of function spaces which in particular contains those spaces. The results we obtain are closely related to the ones for l_n^1 (Schlegel, 2019b). Recall that a point $x \in \mathcal{B}$ is rotund if for any $y \in \mathcal{B}$ such that $\|y\| = \|x\|$ we have $\|y\| = \|\frac{x+y}{2}\|$ implies $x = y$.

Definition 5 (Uniformly non-rotundness) *We say a point $0 \neq f \in \mathcal{B}$ is uniformly non-rotund if it is not rotund for any two dimensional subspace of $\mathcal{B}$ containing it. In other words, f is not rotund in any direction. We say the space $\mathcal{B}$ is uniformly non-rotund if every $0 \neq f \in \mathcal{B}$ is uniformly non-rotund.*

The main reason for uniform non-rotundness to be useful is because it means that there cannot exist faces with a smooth boundary. If any part of the boundary of a face was smooth one would be able to find a two dimensional subspace containing the smooth boundary point and a rotund point in its neighbourhood. If no point in the boundary of a face is smooth then the boundary consists of faces of a lower dimension. As the faces are closed convex sets forming the surface of the norm ball this means that the boundary of a face is given by the intersections with its neighbouring faces. These lower dimensional faces are exposed by another functional and contain their own minimum of Ω. This provides us with a way of running a similar argument as in the cases of uniform and reflexive Banach spaces (Schlegel, 2019a,b). From any admissible point we can reach a minimum on the boundary of its face and from there either go back for a radial bound or move further around the ball for a circular bound.

Lemma 6 *If for every exposed face of the norm ball Ω attains its minimum in at least one point, and for every f in the face where the minimum is attained and every $L \in J(f)$ exposing the face and every $f_T \in \ker(L)$ we have $\Omega(f + f_T) \geq \Omega(f)$, then for any fixed admissible $\hat{f} \in \mathcal{B}$ we have that*

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

for all $f \in \mathcal{B}$ such that $\|\hat{f}\| < \|f\|$.

Proof Once again we follow the proof ideas as for reflexive Banach spaces (Schlegel, 2019b). In particular the proof for l_n^1 is instructive. More precisely, the tangential bound from lemma 3 can be extended to a radial bound by moving "out and back" along tangents. But since the minimum can occur anywhere within the face we actually view Ω as a function $\overline{\Omega}$ of the faces F of the norm ball in $\mathcal{B}$

$$\overline{\Omega}(F) = \min_{f \in F} \Omega(f)$$

We are going to prove that $\overline{\Omega}$ is monotone along the ray λF, $\lambda > 1$, i.e. the minimum of Ω within a face is nondecreasing as a function of the norm. Since each minimum satisfies the tangential bound this gives the half space bound for all half spaces defined by a tangent plane through the minimum $\hat{f}$, given by some $\hat{L} \in J(\hat{f})$, as illustrated in figs. 1 and 2. Moreover by repeatedly moving along tangents we can extend the tangential bound all the way around the circle as can be seen in fig. 3. But since a general Banach space may not contain any exposed points we need to be more careful than in the cases of strictly convex Banach spaces and l_n^1. The difficulties lie in the fact that we need to prove for both arguments that we can always find *admissible points* at which to consider the tangents.

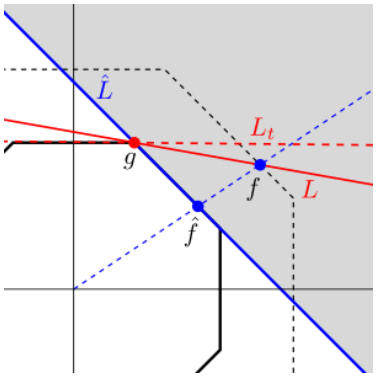

Figure 1: If $\hat{f}$ was the minimum in the face $\hat{F}$, then it has the tangential bound from $\hat{L}$ to reach $\overline{g}$. From $\overline{g}$ we have the tangential bound from L_t to reach any point within $\lambda\hat{F}$ for $1 < \lambda < 1 + \varepsilon$, in particular the minimum within the face.

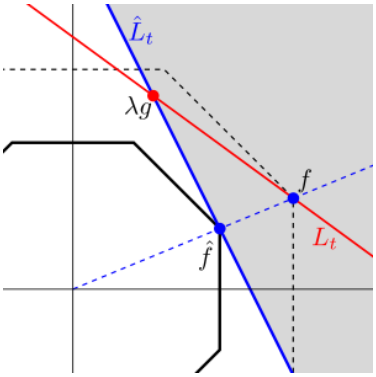

Figure 2: If $\hat{f}$ was an exposed point, then we can construct a set of functionals $\hat{L}_t$ which expose $\hat{f}$ and hit $\lambda\overline{g}$, the minimum in the face λF_t. For $\lambda\overline{g}$ we then get a tangential bound back to the face containing $\mu\hat{f}$ in the same way as above.

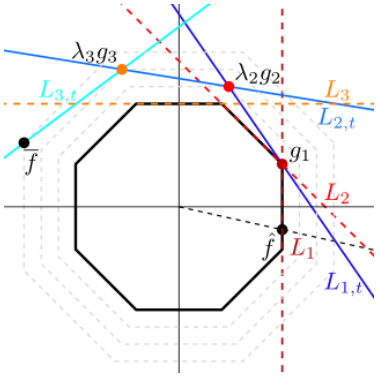

Figure 3: We can move around the circle along points which are extremal in the two dimensional subspace, while staying arbitrarily close to $\overline{g}$ we have the tangential bound from the the circle.

Part 1: *Bound Ω on the half spaces given by the tangent planes through $\hat{f}$*

We start by proving that $\overline{\Omega}$ is radially nondecreasing. Note that we don't need to show monotonicity for the entire ray λF for $1 < \lambda$. It is sufficient to consider $1 < \lambda < 1 + \varepsilon$ as long as the ε is at least nondecreasing as a function of the norm along the ray.

Fix an admissible $\hat{f} \in \mathcal{B}$ and let X be any 2-dimensional subspace containing $\hat{f}$. As $\mathcal{B}$ is uniformly non-rotund no point in X is rotund so its unit ball consists of straight line sections and corners as shown in figs. 1 to 3. In particular there exists $g \neq \hat{f}$ in the same straight section as $\hat{f}$ and exposed in X. It is also clear that there are linear functionals $\hat{L}, L \in X^*$, where $\hat{L}$ exposes the straight segment containing $\hat{f}$ and g, and L exposes only the point g. By Hahn-Banach there are extensions of these functionals to $\mathcal{B}$, also denoted by $\hat{L}$ and L, exposing faces $\hat{F}$ and F respectively.

We now let $L_t = t\hat{L} + (1-t)L$, $t \in (0,1)$ so that L_t exposes the face $F_t = \hat{F} \bigcap F$ which is strictly smaller than $\hat{F}$. Thus Ω has a minimum in F_t, $\overline{g}$ say. Since $\overline{g} \in F_t \subset \hat{F}$ it is clear that $\hat{L}$ attains its norm at $\overline{g}$ which means that there is a tangent from $\hat{f}$ to $\overline{g}$. Being the minimum in F_t we have that $\overline{g}$ has the tangential bound for all L_t.

Putting those observations together we obtain the claimed bound. If $\hat{f}$ was the minimum in the face $\hat{F}$, then it has the tangential bound from $\hat{L}$ to reach $\overline{g}$. From $\overline{g}$ we have the tangential bound from L_t to reach any point within $\lambda\hat{F}$ for $1 < \lambda < 1 + \varepsilon$, in particular the minimum within the face. This is illustrated in fig. 1.

If on the other hand $\hat{f}$ was an exposed point, then it is clear that using an argument similar to the one above we can construct a set of functionals $\hat{L}_t$ which expose $\hat{f}$ and hit $\lambda\overline{g}$, the minimum in the face λF_t. For $\lambda\overline{g}$ we then get a tangential bound back to the face containing $\mu\hat{f}$ in the same way as above. This is illustrated in fig. 2.

This shows that the minimum of Ω for any fixed face F is indeed monotone, which in turn means that any admissible point bounds every point in the open half spaces spanned by a tangent plane at the point.

Part 2: *Extend the bound around the circle*
Next we show that from any fixed admissible point $\hat{f}$ we can reach every other admissible point of norm strictly bigger than $\|\hat{f}\|$. This combined with the half space bound gives the claimed bound for all points outside the circle.
Fix an admissible point $\hat{f} \in \mathcal{B}$ and the admissible point $\overline{f} \neq \hat{f}$ with $\|\overline{f}\| > \|\hat{f}\|$ to be reached. Then $\hat{f}$ and $\overline{f}$ span a two dimensional subspace X. As before X only consists of straight line sections and corners. Clearly we can construct a sequence of points g_i and linear functionals L_i exposing the straight line section from g_{i-1} to g_i as illustrated in fig. 3. As in part 1 by Hahn-Banach we can extend the L_i to $\mathcal{B}$, exposing faces F_i. Moreover by a similar construction as in part 1 we obtain functionals $L_{i,t} = L_i + (1-t)L_{i+1}$, $t \in (0,1)$ exposing the face $F_{i,t} = F_i \bigcap F_{i+1}$ which in particular contains g_i and has a minimum $\overline{g_i}$. This provides us with a tangent from either g_i or $\overline{g_i}$ to g_{i+1} or if necessary $\overline{g_{i+1}}$ so that we can indeed get from $\hat{f}$ to $\overline{f}$ along tangents to points which are minima of a face and hence admissible. Each step includes a step away from the circle but it is clear that it can always be made arbitrarily small by varying t.
With this process we can reach any admissible $\overline{f}$ with $\|\overline{f}\| > \|\hat{f}\|$, which combined with the half space bound from part 1 proves the claim. ■

The proof makes clear that, just as for l_n^1, we are only able to make statements about the minima of faces but not about their location within a face or the remaining points within the face. We thus can only obtain a result about radial symmetry in the spirit of Argyriou et al. (2009) and our previous work (Schlegel, 2019a,b) by viewing Ω as a function of the faces of the norm ball as in the proof of lemma 6. In other words we are thinking of the faces as being collapsed to one point where Ω is minimised. If we think of Ω in this way then the same intuition of almost radial symmetry as in the afore mentioned papers applies.

Theorem 7 *A function $\Omega : \mathcal{B} \to \mathbb{R}$ is admissible if and only if viewed as a function $\overline{\Omega}$ of the faces F of the norm ball in $\mathcal{B}$, $\overline{\Omega}(F) = \min_{f \in F} \Omega(f)$ it is of the form*

$$\overline{\Omega}(F) = h(\|f\|_{\mathcal{B}} \,:\, f \in F)$$

for some nondecreasing $h : [0, \infty) \to \mathbb{R}$ whenever $\|f\|_{\mathcal{B}} \neq r$ 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 $\|f\|_{\mathcal{B}} = r \in \mathcal{R}$ the value $\overline{\Omega}(F)$ is only constrained by the monotonicity property., i.e. it has to lie in between $\lim_{t \nearrow r} h(t)$ and $\lim_{t \searrow r} h(t)$.

Moreover if a face F contains an exposed point then in points of continuity of h the function Ω attains its minimum in every exposed point in F.

Proof *(Sketch)* It turns out that the proof of this result for uniform Banach spaces (Schlegel, 2019a) with the small adjustments for l_n^1 (Schlegel, 2019b) is also valid for non-reflexive Banach spaces. We are going to sketch the arguments below for completeness, more detail can be found in the afore mentioned papers.

Firstly it is easy to show that if Ω is continuous in radial direction then $\overline{\Omega}$ has to be radially symmetric. It is clear that we can only obtain radial symmetry for admissible points but since these bound all other points from below this is sufficient. If f and g are admissible points of the same norm and $\Omega(f) > \Omega(g)$ say, then by lemma 6 for all $1 < \lambda \in \mathbb{R}$ we have $\Omega(\lambda g) \geq \Omega(f)$, which implies that

$|\Omega(\lambda g) - \Omega(g)| \geq |\Omega(f) - \Omega(g)| > 0$ contradicting radial continuity of Ω.

Moreover by the same arguments as for uniform Banach spaces and l_n^1 we can define the radially mollified regulariser

$$\widetilde{\Omega}(f) = \int_{-1}^{0} \rho(t)\Omega\left((\|f\| - t)\frac{f}{\|f\|}\right) \mathrm{d}t$$

and check by direct calculations that $\widetilde{\Omega}(f + f_T) \geq \widetilde{\Omega}(f)$ so $\widetilde{\Omega}$ is tangentially nondecreasing and hence admissible if Ω was admissible. This means that we can mollify in radial direction while preserving admissibility.

Putting these two observations together we obtain the result. We know that $\overline{\Omega}$ is a monotone function of the norm, so a monotone function on the real line and after mollification it is in fact radially symmetric. Thus the same considerations as for uniform and reflexive Banach spaces (Schlegel, 2019a,b) say that $\overline{\Omega}$ must have been of the claimed form.
The converse is clear, since the value of $\overline{\Omega}$ is defined to be the minimum across each face, so minima exist and clearly satisfy the tangential bound.

For the moreover part assume f is an exposed point in a face F which contains a minimum $g \neq f$ of Ω. Assume further that h is continuous in $\|f\|$. Then there are tangents from λf to g for $1 - \varepsilon < \lambda < 1$. This is essentially the same situation as we saw before in fig. 1, from the exposed point we can hit a point in the face above. Thus $\Omega(\lambda f) \leq \Omega(g)$. But since g is a minimum for Ω and is in the same face as f

$$\Omega(\lambda f) \leq \Omega(g) \leq \Omega(f)$$

By continuity of h in $\|f\|$ we have $\Omega(\lambda f) \underset{\lambda \to 1}{\longrightarrow} \Omega(f)$ and so $\Omega(f) = \Omega(g)$ ■

This shows that for any Banach space which is either strictly convex or uniformly non-rotund an admissible regulariser has to be essentially radially symmetric in the appropriate sense. This includes every space we can think of which is commonly used in applications. One should expect that similar arguments are possible for any Banach space once the space has been fixed to remove the issue of geometric variety. More precisely, if a space is relevant for an application it should be an easy check that the same proof strategy of moving between admissible points along tangents can be applied to obtain the analogous result of lemma 6 and thus also of theorem 7. This conjecture is reasonable because with l^1, l^∞, c_{00} and L^1 we cover some examples of spaces often thought of as "as bad as it can get". Many of the spaces one would think of as giving the geometric variety to make a general statement impossible can likely be seen as "nicer" than some of the examples covered here. Once one fixes the space it is usually not difficult to find admissible points to prove the required results.

3. Conclusions

The above results conclude the work by Argyriou et al. (2009); Dinuzzo and Schölkopf (2012) and our earlier work (Schlegel, 2019a,b), providing a unified framework for the existence of representer theorems in general Banach spaces. Most notably this framework now includes non-reflexive Banach spaces, in particular l^1 and L^1-type spaces. It thus includes common methods such as

lasso (Tibshirani, 1996) and variations of it such as square-root lasso (Belloni et al., 2011). Moreover it contains other spaces which may be very interesting for applications, but which are currently not used due to a lack in mathematical and computational theory. As an example consider c_0, the space of sequences converging to zero equipped with the maximum norm. Sequences in this space can for applications be ε-approximated by vectors in c_{00}, i.e. sequences of finitely many non-zero bounded coefficients. Our framework may provide a basis for the development of a theory for regularised learning in such spaces.

3.1. Applicability

The representer theorem for Hilbert spaces provides a basis for a finite dimensional subspace in which a solution to the regularisation problem can be found. In the Banach space case as presented above and in our previous work (Schlegel, 2019a,b) this basis is only for a finite dimensional subspace of the dual space. The statement remains useful since once a Banach space has been fixed we know its duality mapping. We can thus first work with the linear system in the dual space and then pull back the solution via the inverse duality mapping. This can also be seen in the earlier work by Micchelli and Pontil (2004) who propose to include function composition in the representation of the solution and use functionals of the form

$$\phi\left(\sum c_i L_i\right) \tag{7}$$

Now with ϕ an inverse of the duality mapping we obtain a solution to the regularisation problem in the original space. They present an example of this in Besov spaces which also include Lebesgue spaces as a special case. Micchelli and Pontil further remark in their paper that by taking linear combinations of functionals of the form (7) this setting is even general enough to include single hidden layer neural networks.

Another example illustrating the applicability of a representer theorem in Banach spaces is given by Zhang and Zhang (2012). Using a differentiable regulariser they obtain characterisation equations and combining these these with the linear representer theorem in the dual space leads to a system of usually nonlinear equations which describes the solution to the regularisation problem. Solving this system of equations will in general require more advanced approaches than the linear equations in the Hilbert space setting, but other algorithms may be practical. As an example they note that for L^p spaces one obtains a system of polynomial equations for which one then has to find a common zero, allowing for approaches from computational commutative algebra.

3.2. Optimality

It is clear from the proof of lemma 3 that proximinality of the subspace Z is by definition the property that determines whether we can have an exact representer theorem for any given data y_i. We note further that definition 2 (ii) is the best we can hope for when Z is not proximinal. Firstly the infimum is not always attained so we can only find a sequence of approximate minimisers. But moreover we also cannot achieve $\operatorname{dist}(J(f_0), \operatorname{span} L_i) < \varepsilon$ for all $\varepsilon > 0$ with a single $f_0 \in \mathcal{B}$. To see this consider the case $\mathcal{B} = l^1$, $\mathcal{B}^* = l^\infty$. Let $L = (n/n+1)_{n\in\mathbb{N}} = (1/2, 2/3, 3/4, \ldots)$ and consider the regularised interpolation problem

$$\min\{\Omega(f) \,:\, f \in l^1, L(f) = \|L\|_{l^\infty}^2 = 1\}$$

First of all $\|L\|_{l^\infty} = 1$ and there does not exist $f \in l^1$ such that $\|f\|_{l^1} = 1$ and $L(f) = 1$ so $\operatorname{span} L \cap J(l^1) = \{0\}$ and there cannot be a solution in the sense of definition 2 (i). Furthermore any solution f_0 has to be of norm bigger than 1. This means that also any $\hat{L} \in J(f_0)$ would be of norm bigger than 1, $1 + \delta$ for some $\delta > 0$ say. But as $\hat{L} \in l^\infty$ is in the image of the duality mapping, there exists an element in the sequence where the norm is attained, $\hat{L}_i = 1 + \delta$. But then $\|\hat{L} - L\| \geq \hat{L}_i - L_i > (1 + \delta) - 1 = \delta > 0$ and so f_0 could not be a valid solution for any $\varepsilon < \delta$. This shows that the best we could hope for is finding a distinct solution for any $\varepsilon > 0$.

3.3. Future work

Using the characterisation of admissible regularisers we showed (Schlegel, 2019a,b) that in fact the solution in the sense of the exact representer theorem (definition 2 (i)) is independent of the regulariser but only depends on the function space the optimisation problem is posed in. This is a very interesting result which highlights the importance of extending common learning frameworks to a variety of Banach spaces. Moreover it means that one is free to choose whichever regulariser Ω is most suitable for a given application, whether this is numerical computation or mathematical proofs.
The proof of this is based on Theorem 1 in Micchelli and Pontil (2004) which characterises solutions to the regularised interpolation problem as points where the distance of 0 to the subspace $\overline{f} + Z$ is attained, as discussed at the beginning of section 2. It is thus plausible to expect a similar result to hold for the approximate representer theorem (definition 2 (ii)) by characterising approximate solutions as points where the distance of 0 to $\overline{f} + Z$ is almost attained.

Furthermore, even when an exact representer theorem exists, in numerical implementations we are often not going to compute the exact solution but only an approximation to a given ε accuracy. It would be interesting to explore whether the notion of an approximate representer theorem can lead to the design of new algorithms which may improve the computation of approximate solutions even in cases when an exact version of the theorem exists.

Appendix A. The sandwich theorem

Using the Hahn-Banach-Lagrange theorem, a stronger version of the Hahn-Banach theorem, Simons (2008) proves the following Sandwich theorem.

Theorem 8 (Sandwich Theorem) *Let V be a nonzero, real vector space and $P : V \to \mathbb{R}$ sublinear. Define a vector ordering $\leq_P$ on V by*

$$u \leq_P v \text{ if } P(u - v) \leq 0$$

Further assume X is a nonempty set, $k : X \to (-\infty, \infty]$ not identically ∞ and $j : X \to V$. Suppose that for all $x_1, x_2 \in \operatorname{dom}(k)$ there exists $u \in \operatorname{dom}(k)$ such that

$$j(u) \leq_P \frac{1}{2} j(x_1) + \frac{1}{2} j(x_2) \qquad k(u) \leq \frac{1}{2} k(x_1) + \frac{1}{2} k(x_2)$$

Then there exists a linear functional L on V such that $L \leq P$ and

$$\inf_{x \in X} [L(j(x)) + k(x)] = \inf_{x \in X} [P(j(x)) + k(x)]$$

Using this theorem we can easily deduce a corollary that allows us to construct a continuous linear functional of small norm which is in the subdifferential of a given convex function. For a real valued, convex function $F : V \to \mathbb{R}$ on a Banach space V define the directional derivative of F at $\overline{f} \in V$ in direction $h \in V$ as the limit

$$F'(\overline{f}, h) = \lim_{t \searrow 0} \frac{F(\overline{f} + th) - F(\overline{f})}{t}$$

Then F' is everywhere finite and sublinear (Borwein and Lewis, 2006). We choose $P = F'(\overline{f}, \cdot)$ for some fixed $\overline{f}$ in the Sandwich theorem. For simplicity we denote the order relation by $\leq_F$. We let $X = B_V$ the unit ball in V and $j(f) = f$ be the canonical embedding of B_V into V. Lastly define k to be identically 0.
With j being the identity map we get

$$j(h) \leq_F \frac{1}{2} j(h_1) + \frac{1}{2} j(h_2) \Leftrightarrow F'(\overline{f}, h - \frac{1}{2}h_1 - \frac{1}{2}h_2) \leq 0$$

But for any $h_1, h_2 \in B_V$ also $1/2 h_1 + 1/2 h_2 \in B_V$ and $F'(\overline{f}, 0) = 0$ trivially. Further the condition on k is trivially satisfied since k is identically 0. Thus we obtain the following corollary of the sandwich theorem which yields a linear map in the subdifferential of F at $\overline{f}$ with some control over its behaviour on the unit ball which will allow us to bound its norm.

Corollary 9 (Sandwich theorem for subdifferentials) *Let V be a nonzero, real vector space, $F : V \to \mathbb{R}$ a convex, everywhere continuous function and $\overline{f} \in V$. Then there exists a linear functional L on V such that $L(\cdot) \leq F'(\overline{f}, \cdot)$, i.e. $L \in \partial F(\overline{f})$, and*

$$\inf_{h \in B_V} L(h) = \inf_{h \in B_V} F'(\overline{f}, h)$$

Appendix B. Extension of the linear functional in the proof of lemma 3

In the proof of lemma 3 we obtain a functional $L \in Z^*$ such that $\|L\|_{Z^*} < \varepsilon$. We want to extend this functional to $L \in \mathcal{B}^*$ such that $L \in J(f_0^\varepsilon + f_T^\varepsilon)$. We proceed similarly to the proof of the Beurling-Livingston theorem (Blazek, 1982; Schlegel, 2019b). Let $\overline{Z}$ be the vector space generated by Z and f_0^ε and extend L to $\overline{Z}$ by setting

$$L(f_0^\varepsilon) = L(f_T^\varepsilon) - \|f_T^\varepsilon - f_0^\varepsilon\|_{\mathcal{B}}^2$$

Then $L(f_T^\varepsilon - f_0^\varepsilon) = \|f_T^\varepsilon - f_0^\varepsilon\|_{\mathcal{B}}^2$ so $\|L\|_{\overline{Z}^*} \geq \|f_T^\varepsilon - f_0^\varepsilon\|$. Since the norm of L on Z is bounded by ε, and we can without loss of generality assume $\varepsilon \leq \|f_T^\varepsilon - f_0^\varepsilon\|$, we have that the norm of L on $\overline{Z}$ can only be strictly bigger than $\|f_T^\varepsilon - f_0^\varepsilon\|$ if there is a point $\lambda f_T + \nu f_0^\varepsilon$ for $f_T \in Z$ and $\nu \neq 0$ where L has a value strictly bigger than $\|f_T^\varepsilon - f_0^\varepsilon\| \cdot \|\lambda f_T + \nu f_0^\varepsilon\|$. Since ν is nonzero we can divide through by ν and absorb the constant into the subspace Z to equivalently look at points of the form $f_T + f_0^\varepsilon$. But for those points we find that

$$\begin{aligned} L(f_T + f_0^\varepsilon) &= L(f_T + f_T^\varepsilon) - \|f_T^\varepsilon - f_0^\varepsilon\|^2 \\ &\leq \varepsilon \cdot \|f_T + f_T^\varepsilon\| - \|f_T^\varepsilon - f_0^\varepsilon\|^2 \\ &\leq \|f_T^\varepsilon - f_0^\varepsilon\| \cdot |\|f_T + f_T^\varepsilon\| - \|f_T^\varepsilon - f_0^\varepsilon\|| \\ &\leq \|f_T^\varepsilon - f_0^\varepsilon\| \cdot \|f_T + f_0^\varepsilon\| \end{aligned}$$

Thus indeed

$$\|L\| = \|f_T^\varepsilon - f_0^\varepsilon\|$$

Now extend L by Hahn-Banach to a linear functional on $\mathcal{B}$ of the same norm. Then since $L(f_T^\varepsilon - f_0^\varepsilon) = \|f_T^\varepsilon - f_0^\varepsilon\|^2$ by construction $L \in J(f_T^\varepsilon - f_0^\varepsilon)$. But then $-L \in J(f_0^\varepsilon + (-f_T^\varepsilon))$. This completes the proof.

Appendix C. Regularisation and interpolation

Theorem 10 *Let $\mathcal{E}$ be a lower semicontinuous error functional which is bounded from below. Assume further that for some $\nu \in \mathbb{R}^m \setminus \{0\}, y \in Y^m$ there exists a unique minimiser $0 \neq a_0 \in \mathbb{R}$ of $\min\{\mathcal{E}\left((a\nu_i, y_i)_{i\in\mathbb{N}_m}\right) : a \in \mathbb{R}\}$. Assume the regulariser Ω is lower semicontinuous and has bounded sublevel sets.*
Then Ω is admissible for the regularised interpolation problem (3) if the pair $(\mathcal{E}, \Omega)$ is admissible for the regularisation problem (1).

The proof is very similar to the case of reflexive Banach spaces (Schlegel, 2019b), which generalises the proof for Hilbert spaces given by Argyriou et al. (2009). We are going to sketch the overall argument, which can be found in detail in the afore mentioned papers, and only go into detail where ever the proof differs for non-reflexive Banach spaces.
Proof We are going to show that Ω is tangentially nondecreasing in the sense of lemma 3.
For every $\lambda > 0$ consider the regularisation problem

$$\min\left\{\mathcal{E}\left(\left(\frac{a_0}{\|L\|^2}L(f)\nu_i, y_i\right)_{i=1}^m\right) + \lambda\Omega(f) \;:\; f \in \mathcal{B}\right\}$$

Since $\ker(L)$ is proximinal (Conway, 1994, Prop. 4.7) we are in the situation of definition 2 (i) and by admissibility of the pair $(\mathcal{E}, \Omega)$ there exist solutions $f_\lambda \in \mathcal{B}$ such that

$$J(f_\lambda) \cap \operatorname{span}\{L\} \neq \emptyset$$

Using the boundedness of sublevel sets we obtain a weakly* convergent subsequence $(f_{\lambda_l})_{l\in\mathbb{N}}$ such that $\lambda_l \underset{l\to\infty}{\longrightarrow} 0$ and $f_{\lambda_l} \overset{*}{\rightharpoonup} \overline{f}^{**}$ as $l \to \infty$. Since $\mathcal{B}$ is not reflexive we do not get weak convergence as in the cases of Hilbert spaces and reflexive Banach spaces (Argyriou et al., 2009; Schlegel, 2019b). But by lower semicontinuity of $\mathcal{E}$ we still have that

$$\mathcal{E}\left(\left(\frac{a_0}{\|L\|^2}\overline{f}^{**}(L)\nu_i, y_i\right)_{i=1}^m\right) \leq \mathcal{E}((a_0\nu_i, y_i)_{i=1}^m)$$

which as before implies that $\overline{f}^{**}(L) = \|L\|^2$.
Just as before we obtain $\|\overline{f}^{**}\| = \|L\|$ so that $\overline{f}^{**} \in J(L)$. This means that $\overline{f}^{**}$ and $\hat{f}$, where $\hat{f}(L) = L(f)$, both are in the same face of the norm ball in $\mathcal{B}^{**}$.
Considering the lower semicontinuous extension $\overline{\Omega} : \mathcal{B}^{**} \to \mathbb{R}$ of Ω as before we find that $\overline{f}^{**}$ is the minimiser of

$$\min\{\overline{\Omega}(f^{**}) \;:\; f^{**} \in \mathcal{B}^{**}, f^{**}(L) = \|L\|^2\}$$

But by Conway (1994, Prop. 4.7) $\ker(L)$ is proximinal and thus by assumption the interpolation problem

$$\min\{\Omega(f) \;:\; f \in \mathcal{B}, L(f) = \|L\|^2\}$$

has a solution. When the original function attains its minimum then the minimum of the lower semicontinuous extension is not less than the minimum of the original function. Thus $\overline{\Omega}$ attains its minimum on $\hat{\mathcal{B}}$. Thus there exists a $g \in \mathcal{B}$ such that $\hat{g}$ is in the same face as $\overline{f}^{**}$ and $\overline{\Omega}(\overline{f}^{**}) = \overline{\Omega}(\hat{g})$. By the same arguments as for reflexive Banach spaces (Schlegel, 2019b) either $g = f$ or f is an equivalent minimum or f is not admissible.

Finally note that the claim is trivially true for $L = 0$ as in that case $\mathcal{E}$ is independent of f and for every λ the minimiser f_λ has to be zero to satisfy $J(f_\lambda) \cap \{0\} \neq \emptyset$. This means Ω is minimised at 0.
■

Theorem 11 *Let $\mathcal{E}, \Omega$ be an arbitrary error functional and regulariser satisfying the general assumption that minimisers always exist. Then the pair $(\mathcal{E}, \Omega)$ is admissible for the regularisation problem (1) if Ω is admissible for the regularised interpolation problem (3).*

Proof Let f_0 be a solution of the regularisation problem (1). Consider the associated regularised interpolation problem

$$\min\{\Omega(f) \,:\, f \in \mathcal{B}, L_i(f) = L_i(f_0)\, \forall i \in \mathbb{N}_m\}$$

Since Ω is admissible for regularised interpolation, for this interpolation problem there exists a solution $\overline{f_0}$ (or $\overline{f_0^\varepsilon}$) in the sense of definition 2. But then $\Omega(\overline{f_0}) \leq \Omega(f_0)$ and they have the same error as they agree on the data. Thus $\overline{f_0}$ is a solution of (1) in the sense of the representer theorem and the pair $(\mathcal{E}, \Omega)$ is admissible.
■

In conclusion under the assumptions of theorem 10 we have that the pair $(\mathcal{E}, \Omega)$ is admissible for the regularisation problem (1) if and only if Ω is admissible for the regularised interpolation problem (3).

Appendix D. Proximinal subspaces

The following corollary of Godini's theorem gives a criterium for a subspace to be proximinal which is of particular relevance to our work. Godini's theorem and the corollary, including their proofs, can be found in Holmes (1975).

Corollary 12 *Let V be a real normed vector space with unit ball B_V and $W \subset V$ a closed subspace of V.*

(i) If W is finite dimensional it is proximinal.

(ii) If $\operatorname{codim}(W) = m < \infty$ then for any basis $L_1, \ldots, L_m$ of $W^\perp$ define a map S by

$$S : V \to \mathbb{R}^m$$
$$S(x) = (L_1(x), \ldots, L_m(x))$$

Then W is proximinal if and only if $S(B_V)$, the image of the unit ball of V under the map S, is closed in $\mathbb{R}^m$.

Condition (ii) gives a condition for proximinality of the subspace Z in our work, based on the linear functionals defining the regularised interpolation problem.
Singer (1970) addresses the question when every closed subspace of finite codimension, i.e. every possible Z above, is proximinal. He proves the following result.

Proposition 13 *Let $\mathcal{B}$ be a Banach space. Then all closed linear subspaces W of a fixed, finite codimension m, where $1 \leq m \leq \dim(\mathcal{B}) - 1$ are proximinal if and only if $\mathcal{B}$ is reflexive.*

This means that our result is optimal in the sense that for every non-reflexive Banach space $\mathcal{B}$ there exists a combination of linear functionals L_i such that $Z = \cap L_i$ is not proximinal and we cannot obtain an exact representer theorem.

References

Antonio Aizpuru and Francisco J Garca-Pacheco. Some questions about rotundity and renormings in banach spaces. *Journal of the Australian Mathematical Society*, 79(01):131–140, 2005.

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

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Errett Bishop and R. R. Phelps. A proof that every banach space is subreflexive. *Bull. Amer. Math. Soc.*, 67(1):97–98, 1961.

Jaroslav Blazek. Some remarks on the duality mapping. *Acta Universitatis Carolinae. Mathematica et Physica*, 23(2):15–19, 1982.

Jonathan Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer-Verlag New York, second edition, 2006.

J.B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 1994.

Francesco Dinuzzo and Bernhard Schölkopf. The representer theorem for hilbert spaces: a necessary and sufficient condition. In *Advances in neural information processing systems*, pages 189–196, 2012.

Ivar Ekeland. On the variational principle. *Journal of Mathematical Analysis and Applications*, 47: 324–353, 1974.

J.B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer-Verlag Berlin Heidelberg, 2001.

R.B. Holmes. *Geometric Functional Analysis and its Applications*. Graduate Texts in Mathematics. Springer-Verlag, 1975.

Vladimir Kadets, Gins Lpez, Miguel Martn, and Dirk Werner. Equivalent norms with an extremely nonlineable set of norm attaining functionals. *Journal of the Institute of Mathematics of Jussieu*, page 121, 2018. doi: 10.1017/S1474748018000087.

Charles A. Micchelli and Massimiliano Pontil. A function representation for learning in banach spaces. In *Learning Theory. COLT 2004*, pages 255–269. Springer Berlin Heidelberg, 2004.

Charles J Read. Banach spaces with no proximinal subspaces of codimension 2. *Israel Journal of Mathematics*, 223(1):493–504, 2018.

Kevin Schlegel. When is there a representer theorem? nondifferentiable regularisers and banach spaces. *Journal of Global Optimization*, Apr 2019a. doi: 10.1007/s10898-019-00767-0.

Kevin Schlegel. When is there a representer theorem? reflexive banach spaces. *arXiv*, 1809.10284v2, May 2019b.

S. Simons. *From Hahn-Banach to Monotonicity*. Lecture Notes in Mathematics. Springer Netherlands, 2008.

Ivan Singer. *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*. Grundlehren der Mathematischen Wissenschaften. Springer Berlin Heidelberg, 1970.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Haizhang Zhang and Jun Zhang. Regularized learning in banach spaces as an optimization problem: Representer theorems. *Journal of Global Optimization*, 54(2):235–250, 2012. doi: 10.1007/s10898-010-9575-z.

Haizhang Zhang, Yuesheng Xu, and Jun Zhang. Reproducing kernel banach spaces for machine learning. *Journal of Machine Learning Research*, 10:2741–2775, 2009.