On the Existence of Representer Theorems in Banach Spaces

Kevin Schlegel
Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

January 2019

Acknowledgements

First of all I would like to thank Professor Jan Kristensen for his support and guidance as my supervisor and Professor Charles Batty for providing valuable teaching opportunities. I would also like to thank the Foundation of German Business and the Engineering and Physical Sciences Research Council for their support. Moreover I want to thank my CDT cohort, and in particular my officemates Miles and Bogdan, for many discussions at the whiteboard and Sandra for proofreading a substential part of this thesis. Finally I would like to thank Nicky, Sophie and Beatrice for all the therapy sculling.

Abstract

We consider general regularisation and regularised interpolation problems for learning parameter vectors from data. In particular in Hilbert spaces regularisation methods have been applied very successfully, largely due to the well known *representer theorem*. Classical formulations of the theorem state that under certain conditions on the regulariser there exists a solution of the optimisation problem which is contained in a linear subspace spanned by the data points. This is at the core of kernel methods in machine learning as it significantly reduces the dimensionality and thus makes the problem computationally tractable. Most of the literature only deals with sufficient conditions on the regulariser for a representer theorem to hold, mostly in Hilbert spaces with some generalisations to certain classes of Banach spaces. In this work we give an essentially complete answer to the question of existence of representer theorems in general Banach spaces. This question had previously been answered for Hilbert spaces with an intuitive characterisation for differentiable regularisers. We show how the necessary and sufficient conditions extend to arbitrary Banach spaces and give the more intuitive geometric characterisation for a variety of classes of Banach spaces, which contain all spaces we know of which are commonly used in applications. We conjecture that the same characterisation can also be given for any other Banach space not currently covered by those classes. We further show that, if the learning relies on the linear representer theorem, in most cases the solution is actually independent of the regulariser and determined by the function space alone. This is interesting for two reasons. Firstly it means one is free to choose whichever regulariser is most suitable for the application at hand, whether this is computational efficiency or ease of calculations. Moreover it shows the importance of extending classical elements of learning theory such as kernel methods from Hilbert spaces to Banach spaces.

Contents

1 Introduction

A common problem in learning theory, and many other scientific fields, is to find the best function within a class of functions V to explain some given empirical data $(x_i, y_i)_{i=1}^m$. This class of functions may not contain the function which actually generated the data, but does represent some belief we might have about the nature of the data. As an example consider noisy measurement data from physics, meteorology or any other scientific field. We will be looking for the function that explains the law by which the data was generated, without the noise. The noisy function may not be contained in the class V if we have a good enough idea of the properties of the true function. In a way we can view this problem as function approximation. Based on only a finite amount of data in general we can not hope to be able to find the exact true function. But in applications a *good enough* approximation, in an appropriate sense, is often sufficient. Thus the aim is to approximate the true function as well as possible, based on the given, finite amount of information about its behaviour.

Possibly the most common approach across disciplines, supervised and semi-supervised learning but also any other discipline where empirical data is generated, is to formulate the estimation as an optimisation problem. One defines an *error functional* $\mathcal{E}$ to measure the error suffered from using a function f, constrained in a set of functions V, as an approximation of the true function to predict the output for x_i when y_i is the true value. For large, expressive classes of functions minimising the error alone is an ill-posed problem as there may be a large number of functions with the same minimal error. One thus defines a map Ω from the function space into $\mathbb{R}$ which is generally thought of as favouring a certain desirable property of the function such as its regularity. Adding this *regulariser* Ω to the error with a parameter $\lambda > 0$ to trade off between accuracy and regularity one arrives at the regularisation problem

$$\inf \{ \mathcal{E}((f(x_i), y_i)_{i=1}^m) + \lambda \Omega(f) : f \in V \} .$$

This has motivated the study of regularisation problems in mathematics, statistics and computer science, in particular machine learning (Cucker and

Smale [CS01], Shawe-Taylor and Cristianini [STC04], Micchelli and Pontil [MP05a]).

In the generality stated above the regularisation problem is a very hard problem and we have little hope of finding a solution. Thus one commonly will make extra assumptions, in particular on the function space V. Due to the nice geometric structure of Hilbert spaces and the intuition gained from it, regularisation in Hilbert spaces has been studied widely. There are various ways one can phrase the above regularisation problem in Hilbert spaces, the most common one is likely *Tikhonov regularisation*, where we consider an optimisation problem of the form

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

where $\mathcal{H}$ is a Hilbert space and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denotes the inner product on $\mathcal{H}$. There are numerous reasons why regularisation in Hilbert spaces has been studied in great detail and been applied very successfully. Firstly the existence of an inner product and thus the ability to measure angles and the existence of orthogonality and orthogonal projections are very useful tools in the design of algorithms to solve the optimisation problem.
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 [KW71], Cox and O'Sullivan [CO90], Schölkopf and Smola [SS98, SHS01]). This means that the problem reduces to finding a function in a finite-dimensional subspace of the original function space, which is often very high-dimensional or infinite-dimensional. It is this dimension reduction that makes the problem computationally tractable. Since this reduction is so crucial for the success of regularised learning in Hilbert spaces the representer theorem has been extensively studied ([SS02, MP04, AMP09, AD14]) and it will be the main element of study of this thesis.
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 (RKHS)* (Schölkopf and Smola [SS02], Shawe-Taylor and Cristianini [STC04]). 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 embedding the set into a Hilbert space.

It is commonly stated that the regulariser favours certain desirable properties of the solution and can thus intuitively be thought of as picking the function that explains the data and is, in some suitable sense, the simplest. This is in analogy with how a human would pick a function when seeing a plot of the data. One of the main results of this work will clarify this view. We will show that if the learning relies on the linear representer theorem, in most cases the set of solutions in the sense of the representer theorem is the same for every suitable regulariser. This means that the solution is in fact *independent of the regulariser* and it is solely *determined by the function space* we chose to work in. This is in particular always true for Hilbert spaces. The solution being determined by the space alone, and not by the regulariser, strongly motivates extending the successful kernel methods from Hilbert spaces to a greater variety of spaces.

A very important step in extending kernel methods to Banach spaces has been taken by Zhang, Xu and Zhang [ZXZ09] who defined *reproducing kernel Banach spaces (RKBS)*, in part making use of *semi-inner products* which had been introduced by Lumer [Lum61] and Giles [Gil67] exactly for the purpose of extending Hilbert space like methods to Banach spaces.
Zhang and Zhang [ZZ12] extended the theory further and proved representer theorems for a wide variety of regularisation problems of the form

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

where the L_i are continuous linear functionals on $\mathcal{B}$, a Banach space which is at least reflexive, often also uniformly convex and uniformly smooth. The representer theorems proved are different from the ones known for Hilbert spaces though. The representer theorems for RKBS turns out to be naturally rooted in the dual space. They characterise the solution as having a dual element which is contained in the linear span of the dual elements of the data points. This is not in contradiction to Hilbert spaces, but rather may allow deeper insight into the essence of the representer theorem to fully understand it. Since a Hilbert space is self-dual and the duality mapping is the identity it simply does not become apparent that the representer theorem is a statement about the dual space. Our work presented in this thesis will illustrate this fact further.

Reproducing kernel Banach spaces have been extended further by Song, Zhang and Hickernell [SZH13] and Georgiev, Sánchez-Gonzáles and Pardalos [GSGP14] to construct spaces which are not reflexive, in particular of l^1-type. Since l^1 regularisation induces sparsity of the solution, these are very commonly used in applications.

The classical statements of the representer theorem give sufficient conditions on the regulariser for the existence of a solution either in a finite dimensional subspace given by the data points or with dual element in the linear span of their dual elements. Argyriou, Micchelli and Pontil in [AMP09] started to address the question of proving necessary conditions to classify all regularisers which admit a linear representer theorem. They prove a necessary and sufficient condition for Hilbert spaces and give a geometric interpretation of their result for differentiable regularisers. This geometric interpretation is very useful as it makes the result very easily applicable.

In this thesis we are going to expand on their work and provide an essentially complete answer to the question of existence of representer theorems. We show how to remove the differentiability assumption in the arguments of Argyriou, Micchelli and Pontil and extend both, the result about necessary and sufficient conditions and its geometric interpretation, step by step to more general classes of functions. Starting from the Hilbert space case we extend the results to Banach spaces which are uniformly convex and uniformly smooth. In a sense such Banach spaces can be considered almost Hilbert and they form a good starting point to get an idea how the ideas of the previous proofs generalise. We then show how to weaken the assumptions on the function space to reflexivity. Finally we remove even the assumption of reflexivity and prove analogous results for non-reflexive Banach spaces. This is done in particular to include l^1-type Banach spaces, which are commonly used in application. Due to the large geometric variety of general Banach spaces it is not possible to give a closed form geometric interpretation of the necessary and sufficient condition as before anymore. We prove this result for certain classes of function spaces though, which include all cases we know of that are commonly used in application. Since the difficulties arise from the geometric variety, we conjecture that the same sort of arguments can be used to prove the geometric intuition for any Banach space once a particular space has been fixed and its geometry is known. These results have been or will be made available as journal publication in [Sch19b, Sch21] and in conference proceedings [Sch20] as well as arXiv preprints [Sch18, Sch19c, Sch19a].

A complete characterisation of regularisers which admit a linear representer theorem in an arbitrary Banach space and a geometric interpretation for those spaces, which are commonly used in applications is a satisfying result. But for most function spaces there is an important consequence to these results which we already mentioned above. It turns out that in the majority of cases using the characterisation of regularisers that admit a representer theorem one can prove that in fact the solution does not depend on the regulariser. More precisely, for a fixed Banach space $\mathcal{B}$ different regularisers Ω may have a different solution set, but the solution which is determined by the linear subspace generated by the data points is the same for all regularisers which admit a linear representer theorem. This means the solution in the sense of the linear representer theorem is determined by the function space alone.

This is interesting for two reasons. Firstly it means that we can always pick the regulariser best suited for the application at hand. This may be computational efficiency in a concrete application or mathematical properties for use in a proof, such as e.g. the duality of the norm with linear functionals which is exploited in the proofs in [MP04]. Secondly it further illustrates the importance of being able to learn in a larger variety of spaces, i.e. of extending classical elements of learning theory to a variety of Banach spaces.

We will begin this thesis with a general introduction to learning theory in Chapter 2. The aim will not be to be exhaustive, which would not be possible in such a diverse field. The purpose of the chapter is to introduce a reader with mathematical background but no knowledge of learning theory to the most important concepts. The material presented should provide the reader with some context for the results presented later on. We will introduce a general framework for regularisation problems as the ones stated above and give an overview of the most important results for reproducing kernel Hilbert spaces. As this is very well established and known theory, which has been covered extensively in the literature, we will mostly only sketch or omit the proofs unless they provide useful insight for later results.

To be able to use Hilbert space like arguments in Banach spaces we are going to need the semi-inner product theory introduced by Lumer [Lum61] and Giles [Gil67]. Semi-inner products were not only used by Zhang, Xu and Zhang in [ZXZ09, ZZ12] to define reproducing kernel Banach spaces, but will also be used in some of our results presented in this thesis in Chapter 6. Since semi-inner products are not very well known, we are going to discuss them in detail in Chapter 3. We are not presenting any new results in this section but aim to cover what is needed for the reader to get a good impression of the structure semi-inner products provide. The proofs for all results that have relevance to our results, in the sense that they are crucial in proving the necessary structure of the space, will be given. We are going to include some further results which are not strictly required for our results but provide some further context and illustrate the usefulness of semi-inner products. Will we be a bit more brief when covering those results.

After introducing semi-inner products in Chapter 3 we will show in Chapter 4 how to use them to construct reproducing kernel Banach spaces, as done by Zhang, Xu and Zhang [ZXZ09] and Zhang and Zhang [ZZ12]. This chapter aims to provide context and examples for our results similarly to Chapter 2.

But since RKBS are much less well known than RKHS we are going to provide a great deal more detail than we did when covering RKHS and proofs for many of the results will be given. We will see later that semi-inner products do exist for any normed space, but for our purposes they are less useful when we consider spaces which are not reflexive. In the second part of Chapter 4 we will thus show how to construct non-reflexive RKBS using the duality pairing rather than semi-inner products. These constructions include in particular spaces of l^1-type for their relevance in applications.

The discussion of non-reflexive RKBS in Section 4.2 already illustrates the need for more general tools than semi-inner products to develop a unified theory of representer theorems for arbitrary Banach spaces. In Chapter 5 we are going to introduce all further mathematical tools needed for our results. This includes results about the duality pairing and its connections with the geometry of the space, annihilators as a generalisation of orthogonal complements, and discussions of other geometric properties of the space which have significance for our results, such as exposed faces and the attainment of distances of points to linear subspaces.

In Chapter 6 we are going to present most of the main results of this thesis. We begin with a brief discussion of the work by Argyriou, Micchelli and Pontil, which is the starting point of our work. They proved a necessary and sufficient condition on the regulariser for a representer theorem to hold and gave a geometric interpretation of the result for differentiable regularisers. We will then present our own work expanding on these results. We will first show how to remove the differentiability condition in the geometric interpretation. Subsequently we show how to generalise the proofs of both, the necessary and sufficient condition and the geometric interpretation, to uniformly convex and uniformly smooth Banach spaces using semi-inner products.

In a further step we extend the results to any reflexive Banach space. It turns out though, that the results one can obtain in this case may be weaker than the ones from previous sections. The property that determines the exact form of the results obtained will turn out to be strict convexity. While the necessary and sufficient condition can still be given in a closed form for an arbitrary reflexive Banach space, for the geometric interpretation this will not be possible anymore at this stage. This is due to the large geometric variety one may encounter within the class of reflexive Banach spaces. For the time being we are thus going to restrict ourselves to the case of strictly convex spaces when giving the geometric interpretation to reduce this variety and obtain a nice, easily interpretable result. A similar result can be proved for other classes of reflexive Banach spaces but the discussion of spaces which are not strictly convex is postponed, since this case has much in common with the results for non-reflexive Banach spaces.

Following the discussion of reflexive Banach spaces we then give an example why reflexivity is in some sense necessary for a representer theorem of the type obtained throughout the earlier sections of this chapter to hold. To circumnavigate this issue we are finally going to propose a notion of *approximate solution* and thus *approximate representer theorem*. We will show that for this new notion we can indeed obtain the same necessary and sufficient condition on the regulariser to admit a representer theorem. Finally we will show that also the same geometric interpretation holds for a certain class of function spaces, which in particular contains l^1 as the most common example of a non-reflexive Banach space used in applications. We conjecture that the same result could be proved for any Banach space, as long as it has been fixed to remove the geometric variety.

In summary, throughout Section 6 we present a sequence of progressive generalisations of three results. The necessary and sufficient condition on the regulariser is proved for Hilbert spaces, uniformly convex and uniformly smooth Banach spaces, reflexive Banach spaces, and finally arbitrary Banach spaces. Both the lemma which allows to remove the differentiability condition from [AMP09] and the resulting geometric interpretation are proved for Hilbert spaces, uniformly convex and uniformly smooth Banach spaces, strictly convex and reflexive Banach spaces, and uniformly non-rotund Banach spaces.

Finally in Chapter 7 we are going to present the last of our main results. This is the previously mentioned fact that in most cases the solution in the sense of the representer theorem is actually independent of the regulariser but determined by the space alone. Again strict convexity is the crucial property here. In a space which is not strictly convex the regulariser has some effect on the solution, but only to the extent that it determines which point within a given exposed face of the norm ball is optimal. This result is currently only valid for reflexive Banach spaces. It seems reasonable to expect a generalisation to non-reflexive Banach spaces but we have not yet been able to prove it. We then illustrate a few other open questions and connections with other results that could provide starting points for interesting further work.

1.1 Notation

We briefly fix some notation used throughout this thesis. Firstly we are going to use $\mathbb{N}_m$ to denote the set $\{1, \ldots, m\} \subset \mathbb{N}$. We sometimes will consider constructions which are valid for real or complex vector spaces. We will then use $\mathbb{F}$ to denote the base field $\mathbb{R}$ or $\mathbb{C}$.

We will generally denote by X a non-empty set which might not have any further mathematical structure. A normed vector space will usually be called V with dual space V^*. Continuous linear functionals on V will most of the time denoted by L and variants of it. Occasionally, when there is a bijective duality mapping and the correspondence between a point x and the continuous linear functional attaining its maximum at the point is relevant, we may also use x^*. Elements in the second dual will generally be denoted as x^{**} and the respective variants, with $\hat{x}$ denoting the natural embedding of $x \in V$ into the bidual V^{**}.
Subspaces of a vector space V will usually be called U and W. The unit ball of V will be denoted by B_V and the unit sphere by S_V. Balls and spheres of a fixed radius r will be denoted by B_r and S_r respectively.

By $\mathcal{H}$ we will always mean a Hilbert space and $\mathcal{B}$ will be a Banach space. The inner product on a Hilbert space $\mathcal{H}$ is going to be denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a semi-inner product on a Banach space $\mathcal{B}$ by $[\cdot, \cdot]_{\mathcal{B}}$. By $(\cdot, \cdot)_V$ we will denote

either a bilinear form on V or the duality pairing on $V \times V^*$. Which of the two meanings is used should always be clear from the context and not cause any confusion.
We will say a space is a Hilbert space of functions if it is a Hilbert space for which the norm is zero if and only if the function is zero everywhere. Equally a Banach space of functions is a Banach space for which the norm is zero if and only if the function is zero everywhere.

The domain of a function $f : V \to [-\infty, \infty]$ is defined as the set

$$\mathrm{dom}(f) = \{x \in V \,:\, |f(x)| < \infty\}.$$

The core of a set $S \subset V$ is the set

$$\mathrm{core}(S) = \{x \in S \,:\, \forall y \in X \; \exists t_y > 0 \text{ s.t. } x + ty \in S \; \forall t \in [0, t_y]\}.$$

2 An Introduction to Learning Theory

In this chapter we give an introduction into learning theory. The aim is not to be exhaustive. Learning theory is a very broad term covering a diverse range of subdisciplines, making use of a wide variety of branches of mathematics. The aim of this section will be to give an introduction into a part of learning theory that leads to applications of classical functional analysis. It is aimed at a reader with a mathematical background but no knowledge of learning theory or machine learning. The material presented should provide the reader with some context for the new results presented in this thesis. As such the focus will be on results that aid understanding of the mathematical concepts and tools rather than results of direct practical use.

At the same time this section provides an opportunity to fix notation and nomenclature which can in general differ across the common literature.

The general view taken in this work is that *learning* essentially is function approximation. Given some data we are looking for a function f in a given prescribed class which explains the data. With only a finite amount of data we can in general not expect to be able to find the exact function underlying the data. By *learning* we mean trying to find a sufficiently good approximation of this function. How we measure how *good* an approximation is can vary significantly, the common theme though is that it is measured on data points which were not previously known.

The material presented in this chapter will mainly be following the work of Cucker and Smale [CS01] and Schölkopf and Smola [SS02]. Both are excellent references for an introduction to the theory. The paper by Cucker and Smale is more on the theory side, focusing largely on statistical learning theory. The book by Schölkopf and Smola is much broader and also concerned about the applications. Many concrete examples are given. The work by Evgeniou, Pontil and Poggio [EPP00] aims to bridge between a functional analysis perspective and methods from probability theory and statistics and is as such also closely related to the material presented in this chapter.

2.1 Mathematical Foundations

This section will introduce a common mathematical framework for learning. At first we will mainly be following the work of Cucker and Smale [CS01] and later adopt the viewpoint of Schölkopf and Smola [SS02].

We will consider a set of inputs X, which a priori does not need to have any mathematical structure. Further let $Y \subseteq \mathbb{R}$ denote the set of outputs. This could be all of $\mathbb{R}$ for regression tasks or a finite subset for classification tasks. We assume empirical data $z = (z_1, \dots, z_m) = ((x_1, y_1), \dots, (x_m, y_m)) \subset X \times Y$ is drawn according to a fixed but unknown Borel probability distribution ρ on $Z = X \times Y$. The task now is to find a function f_0 in some function space V, the hypothesis space, such that f_0 gives a good prediction of the output $y \in Y$ on points $x \in X$ which were not part of the empirical data z. For this purpose we need to introduce a framework, which allows to measure in various ways what we mean by a good prediction.
We start by defining what we mean by the loss incurred by using a function f to predict the output y at a single data point x.

Definition 2.1 *(Loss Function)*

A loss function is a map $\mathcal{C} : X \times Y \times Y \to [0, \infty)$ such that $\mathcal{C}(x, y, y) = 0$ for all $x \in X$ and all $y \in Y$. The error of using a function f to predict the output y at a single point x is then given by $\mathcal{C}(x, f(x), y)$.

This in particular means that an exact prediction $f(x) = y$ never incurs any loss.

Remark 2.2

Note that by this definition the loss function may depend on the point x. While many common loss functions are independent of x this is an important feature for some applications. Sometimes having data fall within or outside a certain range may mean that an error made in this case is much more or less severe than in other cases. As an example consider e.g. medical data where errors can have very different significance.

When considering a loss function which measures the error independently of x we may just write $\mathcal{C}(f(x), y)$ for simplicity.

Examples of loss functions include the misclassification loss

$$\mathcal{C}(f(x), y) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{otherwise} \end{cases}$$

and the most commonly used squared loss

$$\mathcal{C}(f(x), y) = (f(x) - y)^2.$$

Now that we have defined how to measure how well a function is doing on a single data point we need to extend this to the full space to be able to assess any given candidate function.

Definition 2.3 *(Error Function)*
The error of a function f with respect to the loss function $\mathcal{C}$ and the probability measure ρ is defined as

$$\mathcal{E}(f) = \int_Z \mathcal{C}(x, f(x), y)\,\mathrm{d}\rho(x, y). \tag{1}$$

The task is to find a function which attains or comes sufficiently close to the infimum

$$\inf\{\mathcal{E}(f) : f \in V\}$$

i.e. to minimise the error over a function space V, the hypothesis space. Note that we are not assuming that the true function which generated the data is in V. The minimal error might thus be strictly positive.
For computational reasons the least squares error

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2\,\mathrm{d}\rho(x, y)$$

is widely used and can be thought of any time we are talking about an error functional in this thesis.

Since the probability distribution generating the data is unknown we cannot compute the integral (1). The quantity we can compute is the empirical error

$$\mathcal{E}_z(f) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{C}(x_i, f(x_i), y_i).$$

We are often going to be dealing with bounded linear functionals on a vector space V which may or may not be point evaluations. In this case we will consider the empirical error

$$\mathcal{E}_z(f) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{C}(L_i, L_i(f), y_i).$$

where we make the obvious generalisation of the loss function $\mathcal{C}$ in its first argument to be a map on the dual space V^* of a function space V.

The empirical error $\mathcal{E}_z$ obviously strongly depends on the sample and really should be thought of as a function of f and the m data points $(x_i, y_i) \in X \times Y$ or $(L_i, y_i) \in V^* \times Y$ respectively. This dependence on the sample is commonly hidden in the subscript z.

When we do want to make the sample dependence explicit we may write

$$\mathcal{E}_z\left((x_i, f(x_i), y_i)_{i\in\mathbb{N}_m}\right) \quad \text{and} \quad \mathcal{E}_z\left((L_i, L_i(f), y_i)_{i\in\mathbb{N}_m}\right).$$

Note that sometimes also empirical error functions which are not additively separable as presented above are used. These are much less common though. The results presented in this thesis are in general also valid for such empirical error functions.

Without further assumptions minimising the empirical error does not guarantee good generalisation of the minimiser to unseen data. Consider the case when we make no assumption at all about the function to be learned, i.e. V contains all possible functions. Then for any fixed m training data points $((x_1, y_1), \dots, (x_m, y_m)) \subset X \times Y$ and k additional data points not used for training, $((x_{m+1}, y_{m+1}) \dots, (x_{m+k}, y_{m+k})) \subset X \times Y$ say, there are functions f_1 and f_2 such that

$$f_1(x_i) = f_2(x_i) \text{ for } i \in \mathbb{N}_m$$
$$f_1(x_i) \neq f_2(x_i) \text{ for } i = m+1, \dots, m+k$$

We would have no way of telling which of the two functions to consider as the solution by just minimising the empirical error $\mathcal{E}_z$ on the m training points. Cucker and Smale in [CS01] suggest to use compact subsets of $C(X)$, the space of continuous functions on X, or closed balls in finite-dimensional subspaces of $C(X)$ as hypothesis space. The compactness or finite-dimensionality allow to relate the actual error to the empirical error and hence to give some learning guarantees.

We are going to take a different approach, which is very common in applications and well covered in the book by Schölkopf and Smola [SS02]. We will be adding a regularisation term to the empirical error for better conditioning of the problem. More precisely we seek to solve the minimisation problem

$$\inf\{\mathcal{E}_z(f) + \lambda\Omega(f) \,:\, f \in V\},$$

where the map $\Omega : V \to \mathbb{R}$ is a regulariser and $\lambda > 0$ is a regularisation parameter to balance the trade-off between minimisation of $\mathcal{E}_z$ and regularity of the solution enforced by Ω. This optimisation problem appears very commonly used in applications, in particular supervised and semi-supervised learning but also many other disciplines, wherever empirical data is produced and has to be explained by a function.

2.2 Example: Support Vector Machines

Let us illustrate the above setting by considering a toy problem in binary classification from [SS02], where more details can be found. In binary classification we want to assign a label to a given input, putting it into one of two categories. This means we are setting $Y = \{\pm 1\}$ and we are looking for a function $f : X \to \{\pm 1\}$.

For the toy example we will assume $X = \mathbb{R}^d$ and particularly think of $\mathbb{R}^2$ to be able to visualise the example. Assume we are given some data

$$z = (z_1, \ldots, z_m) = ((x_1, y_1), \ldots, (x_m, y_m)),$$

with $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ to learn from and assume further that the data points of the two categories are clearly separated by a hyperplane

$$\{x \in \mathbb{R}^d \,:\, \langle w, x\rangle_{\mathbb{R}^d} + b = 0\},$$

where $w \in \mathbb{R}^d, b \in \mathbb{R}$. This is illustrated in Fig. 1.

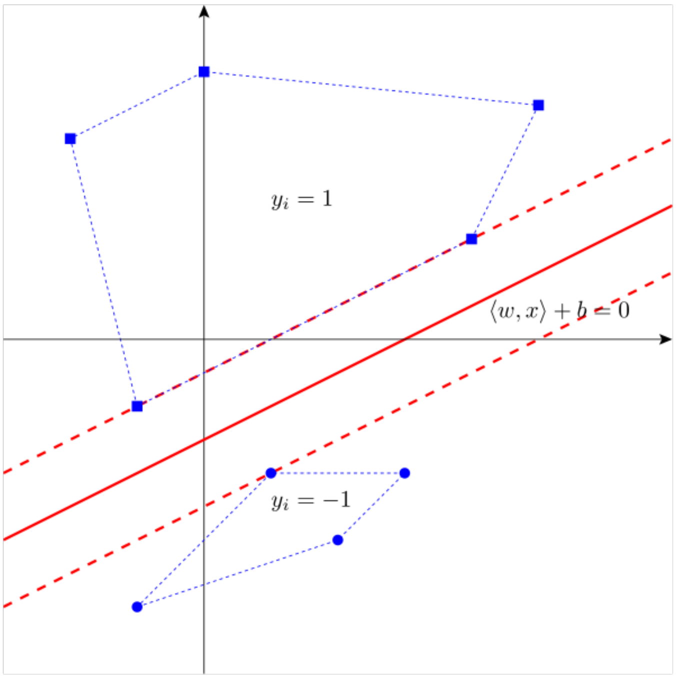


Figure 1: Maximal margin hyperplane classification

We want to use such a hyperplane as decision boundary for the classification. More precisely we are looking for a decision function of the form

$$f(x) = \operatorname{sgn}\left(\langle w, x \rangle_{\mathbb{R}^d} + b\right).$$

It is clear that under the given assumptions there are in general infinitely many hyperplanes separating the data. As stated above we aim to find the hyperplane which performs best on classifying previously unseen data points. There are theoretical arguments, some of which can be found in [SS02], that this hyperplane is the one with the maximal margin, i.e. the unique separating hyperplane which has the maximal distance to any of the training points, as pictured in Fig. 1.

To find the maximal margin hyperplane we have to solve

$$\begin{aligned} &\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2}\|w\|^2, \\ \text{subject to} \quad &y_i \left(\langle w, x_i \rangle_{\mathbb{R}^d} + b\right) \geq 1 \text{ for all } i \in \mathbb{N}_m. \end{aligned} \tag{2}$$

Using Lagrange multipliers we obtain the dual optimisation problem

$$\begin{gathered}\max_{l \in \mathbb{R}^m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i l_j y_i y_j \left\langle x_i, x_j \right\rangle_{\mathbb{R}^d}, \\ \text{subject to } \alpha_i \geq 0 \text{ for all } i \in \mathbb{N}_m \\ \text{and } \sum_{i=1}^{m} \alpha_i y_i = 0, \end{gathered} \tag{3}$$

which leads to the decision function

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i \left\langle x, x_i \right\rangle_{\mathbb{R}^d} + b \right), \tag{4}$$

where the b can also be computed explicitly using the constraints. This puts us into a good position because we have obtained a quadratic program (3) which we can solve.

When introducing the Lagrange multipliers one can also notice that all α_i for which the constraints are not an equality vanish, i.e. if $y_i \left(\left\langle w, x_i \right\rangle_{\mathbb{R}^d} + b \right) > 1$ for some $i \in \mathbb{N}_m$ then $\alpha_i = 0$. This means the solution will in general only depend on a very small subset of the training data. Those data points (x_i, y_i) with $\alpha_i \neq 0$ are called support vectors owing to the similarity to usual Lagrange formulations in classical mechanics and hence this classification method is often referred to as *support vector classification* or *support vector machine (SVM)*. Since we require all the data points to be classified correctly without any margin for error this case is also referred to as *hard margin classification*.

In practise we often encounter situations where the data is not perfectly separable by a hyperplane. This could happen e.g. due to noise corrupting the data. To account for a few examples violating the constraints in (2) one introduces slack variables $s_i \geq 0$ for all $i \in \mathbb{N}_m$ and relaxes the constraints to

$$y_i \left(\left\langle w, x_i \right\rangle_{\mathbb{R}^d} + b \right) \geq 1 - s_i \text{ for all } i \in \mathbb{N}_m.$$

One then penalises the slack variables in the objective function to obtain a trade-off between maximising the margin and minimising the error made by misclassifying training data due to the slack variables. This setting is called *soft margin classification.* We can e.g. penalise the l_1 norm of the slack to obtain the new optimisation problem

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{m} s_i,$$
$$\text{subject to } s_i \geq 0 \text{ for all } i \in \mathbb{N}_m$$
$$\text{and } y_i\left(\langle w, x_i\rangle_{\mathbb{R}^d} + b\right) \geq 1 - s_i \text{ for all } i \in \mathbb{N}_m.$$

This can be handled as sketched above to arrive at the very similar dual problem

$$\max_{l \in \mathbb{R}^m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i l_j y_i y_j \left\langle x_i, x_j\right\rangle_{\mathbb{R}^d},$$
$$\text{subject to } 0 \leq \alpha_i \leq C \text{ for all } i \in \mathbb{N}_m$$
$$\text{and } \sum_{i=1}^{m} \alpha_i y_i = 0.$$

In fact the only difference to (3) is the upper bound $C \in \mathbb{R}$ on the α_i which is limiting the influence an individual data point can have. Clearly we again obtain a decision function of the form (4) with the same properties as discussed earlier.

Support vector machines (SVMs) fit into the setting introduced in Section 2.1 by using the hypothesis space

$$V = \left\{f(x) = \operatorname{sgn}\left(\langle w, x\rangle_{\mathbb{R}^d} + b\right) \,:\, w \in \mathbb{R}^d, b \in \mathbb{R}\right\}.$$

For the soft margin classification we use the misclassification loss function

$$\mathcal{C}(\langle w, x\rangle_{\mathbb{R}^d} + b, y) = \begin{cases} 0 & \text{if } \operatorname{sgn}(\langle w, x\rangle_{\mathbb{R}^d} + b) = y \\ 1 & \text{otherwise} \end{cases} = \max\{1 - \operatorname{sgn}(\langle w, x\rangle_{\mathbb{R}^d} + b)\cdot y, 0\}$$

and the regulariser

$$\Omega(\operatorname{sgn}(\langle w, x\rangle_{\mathbb{R}^d} + b)) = \frac{1}{2}\|w\|^2$$

to obtain the regularisation problem

$$\inf\left\{\sum_{i=1}^{m} \max\{1 - \operatorname{sgn}(\langle w, x_i\rangle_{\mathbb{R}^d} + b)\cdot y_i, 0\} + \lambda \cdot \left(\frac{1}{2}\|w\|^2\right)\right\}. \tag{5}$$

The hard margin classification could be obtained by allowing the loss function to take the value $+\infty$ when a point is misclassified, forcing all points in the training data to be classified correctly. Another approach is to pose it as a regularised interpolation problem, which we will deal with in great detail in Chapter 6. In this case we again choose the regulariser

$$\Omega(\operatorname{sgn}(\langle w, x\rangle_{\mathbb{R}^d} + b)) = \frac{1}{2}\|w\|^2$$

again to obtain the optimisation problem

$$\inf\left\{\frac{1}{2}\|w\|^2 \,:\, \operatorname{sgn}(\langle w, x_i\rangle_{\mathbb{R}^d} + b) = y_i,\ w \in \mathbb{R}^d, b \in \mathbb{R}\right\}.$$

Note that this is also the limit of Eq. (5) as $\lambda \to 0$, a fact which will be illustrated further and exploited in Chapter 6.

2.3 The Representer Theorem

The example of SVMs in Section 2.2 leads us to the main result studied in many variations in this thesis, the well known representer theorem. We notice that in the examples presented in Section 2.2 we obtain a solution which is a linear combination of inner products with the training data points. This is no coincidence. The classical representer theorem states that for certain regularisers one can always find such a solution.

More precisely if our data X is embedded into a hypothesis space which is a Hilbert space $\mathcal{H}$, i.e. $x_i \in \mathcal{H}$, then we can always find a solution in the linear span of the data points if the regulariser is a nondecreasing function of the Hilbert space norm.

Proposition 2.4 *(Representer theorem)*

Let $\mathcal{H}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and consider the regularisation problem

$$\min\{\mathcal{E}_z(\langle f, \cdot\rangle_{\mathcal{H}}) + \lambda\Omega(\|f\|_{\mathcal{H}}) \,:\, f \in \mathcal{H}\} \tag{6}$$

for $\mathcal{E}_z$ the empirical error for an arbitrary loss function $\mathcal{C}$, and a nondecreasing function $\Omega\colon[0,\infty) \to \mathbb{R}$.

Then there always exists a minimiser $f_0 \in \mathcal{H}$ of Eq. (6) of the form

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

This theorem is at the core of the success of support vector machines and kernel methods which we will introduce in the next section. This is because it reduces the problem of finding a minimiser in the possibly infinite dimensional space $\mathcal{H}$ to the finite dimensional subspace $\operatorname{span}\{x_i : i \in \mathbb{N}_m\}$. The representer theorem thus makes the problem computationally tractable. The proof of the theorem is not difficult and will be omitted here. We will be giving a proof of the version for reproducing kernel Hilbert spaces later. The original proof of this theorem goes back to Kimeldorf and Wahba [KW71].

2.4 Reproducing Kernel Hilbert Spaces

We now give a quick introduction into the classical theory of reproducing kernel Hilbert spaces (RKHS), stating the most important definitions and properties. We will illustrate their relevance to machine learning, in particular we will introduce what is commonly known as the kernel trick. The kernel trick provides us with a way of generalising settings such as SVMs significantly, allowing us to introduce nonlinearities. Further detail on RKHS and their applications from a machine learning perspective can be found in the main source of this section [SS02].

We begin with the most abstract definition of a reproducing kernel Hilbert space.

Definition 2.5 *(Reproducing Kernel Hilbert Space I)*

Let X be an arbitrary, non-empty set. A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is a Hilbert space of functions $f : X \to \mathbb{F}$ such that all point evaluation functionals δ_x are continuous.

For most applications this definition is not very instructive. Using the Riesz representation theorem we can obtain a much more intuitive definition. Since for every $x \in X$ we have $\delta_x \in \mathcal{H}^*$, Riesz representation theorem implies that for every $x \in X$ there exists a unique $f_x \in \mathcal{H}$ such that for every $f \in \mathcal{H}$

$$f(x) = \delta_x(f) = \langle f, f_x \rangle_{\mathcal{H}} .$$

Thus, choosing $f = f_y$, in particular $f_y(x) = \langle f_y, f_x \rangle_{\mathcal{H}}$ and we can define a function $k : X \times X \to \mathbb{F}$ by

$$k(x,y) = \langle f_x, f_y \rangle_{\mathcal{H}} .$$

It is clear that this defines a positive definite kernel in the following sense.

Definition 2.6 *((Positive Definite) Kernel)*
Let X be a non-empty set. A symmetric function $k : X \times X \to \mathbb{F}$ is called a positive definite kernel if for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in X$ we have

$$\sum_{i,j=1}^{n} c_i \overline{c_j} k(x_i, x_j) \geq 0. \tag{7}$$

This means for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in X$ the kernel gives rise to a positive semi-definite Gram matrix $K_{i,j} = k(x_i, x_j)$.
Note that in the complex case the assumption that k is symmetric can be dropped as conjugate symmetry is already implied by the definition.

This leads to an alternative definition for reproducing kernel Hilbert spaces.

Definition 2.7 *(Reproducing Kernel Hilbert Space II)*
Let X be an arbitrary, non-empty set. A Hilbert space of functions $\mathcal{H} = \{f : X \to \mathbb{F}\}$ is called a reproducing kernel Hilbert space if there exists a function $k : X \times X \to \mathbb{F}$, the reproducing kernel, such that

(i) Reproducing property:

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}} \text{ for all } f \in \mathcal{H};$$

so in particular

$$k(x,y) = \langle k(x,\cdot), k(\cdot,y) \rangle_{\mathcal{H}}$$

(ii) k spans $\mathcal{H}$:

$$\mathcal{H} = \overline{\operatorname{span}}\{k(x,\cdot) : x \in X\}.$$

From the remarks preceding this definition it is clear that every reproducing kernel Hilbert space in the sense of Definition 2.5 defines a positive definite reproducing kernel with the property (i). Similarly from the reproducing property and continuity of the inner product it is immediately clear that point evaluations are continuous on a RKHS in the sense of Definition 2.7. It is also not difficult to see that the reproducing kernel k for a given RKHS is unique. This follows directly by combining its symmetry and the reproducing property.

As in most cases one is interested in positive definite kernels in the above sense it is common in the machine learning community to refer to those simply as kernels and use more cumbersome terminology for other cases such as e.g. strictly positive definite kernel if we want equality in Eq. (7) to only occur if all c_i are zero.

The spanning property (ii) in Definition 2.7 becomes clearer through the Moore-Aronszajn theorem, which is sort of the converse to the above construction.

Theorem 2.8 *(Moore-Aronszajn)*

Let k be a positive definite kernel on a non-empty set X. Then there is a unique Hilbert space of functions on X for which k is a reproducing kernel.

Proof *(sketch)*:

We set $\mathcal{H}_0 = \operatorname{span}\{k(x,\cdot) : x \in X\}$ and for $f = \sum_{i=1}^{n} s_i k(x_i,\cdot)$ and $g = \sum_{j=1}^{m} t_j k(y_j,\cdot)$ where $n, m \in \mathbb{N}, s_i, t_j \in \mathbb{F}, x_i, y_j \in X$ we define an inner product by

$$\langle f, g\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} s_i t_j k(x_i, y_j).$$

One can directly check that this is well defined, i.e. is independent of the representation of f and g, and indeed defines an inner product on $\mathcal{H}_0$. We set $\mathcal{H}$ to be the completion of $\mathcal{H}_0$ with respect to this inner product. The reproducing property (i) is satisfied by construction and thus in particular $k(x,y) = \langle k(x,\cdot), k(\cdot,y)\rangle_{\mathcal{H}_0}$. Thus any two inner products must agree on $\mathcal{H}_0$ and since the completion is unique we have uniqueness of $\mathcal{H}$.

❑

The Moore-Aronszajn theorem is interesting because it shows that there is a one-to-one correspondence between positive definite kernels and reproducing kernel Hilbert spaces on X.

The proof of the Moore-Aronszajn theorem gives rise to an interesting way of thinking about reproducing kernel Hilbert spaces which is of particular relevance for applications. In practice we may not know which feature space $\mathcal{H}$ to choose for learning but we may have some intuition about what it means for two inputs x and y to be similar. We can think of a kernel function as a measure of similarity, the value $k(x,y)$ saying how similar x and y are. For this case the proof of Theorem 2.8 is instructive. We simply constructed the feature space associated to our desired kernel k as $\mathcal{H} = \overline{\operatorname{span}}\{k(x,\cdot) : x \in X\}$. This means we obtain the feature map

$$\begin{aligned}\Phi : X &\to \mathcal{H},\\ x &\mapsto k(x,\cdot),\end{aligned}$$

thus identifying every point in the set X with a function in the RKHS $\mathcal{H}$. We can think of this as identifying the point with a measure of its similarity with all other points of the set.
We obtain the identity

$$k(x,y) = \langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}}.$$

This means we can always start with a given kernel as a similarity measure and construct a feature space $\mathcal{H}$ so that the kernel corresponds to the inner product of the feature space.

Another common way of constructing a feature space for a given kernel k is the Mercer l_2 feature space which is defined in terms of eigenvalues and eigenfunctions of an integral operator induced by k.

Proposition 2.9 *(Mercer's theorem)*

Let X be a set and $(X, \mathcal{A}, \mu)$ be a finite measure space and $k \in L^\infty(X \times X)$ be a symmetric, real-valued function such that the integral operator

$$\begin{aligned} T_k : L^2(X) &\to L^2(X), \\ (T_k f)(x) &= \int\limits_X k(x,y) f(y) \, \mathrm{d}\mu(y) \end{aligned}$$

is positive definite, i.e. for all $f \in L^2(X)$ we have

$$\int\limits_{X \times X} k(x,y) f(x) f(y) \, \mathrm{d}\mu(x,y) \geq 0.$$

Let $(\lambda_i, \psi_i) \in (0, \infty) \times L^2(X)$ be the eigenvalues and associated orthonormal eigenfunctions of T_k, sorted in non-increasing order. Then

- $(\lambda_i) \in l^1$;
- $k(x,y) = \sum\limits_{i=1}^{N_{\mathcal{H}}} \lambda_i \psi_i(x) \psi_i(y)$ holds for almost all $(x,y) \in X \times X$. Either $N_{\mathcal{H}} \in \mathbb{N}$ or $N_{\mathcal{H}} = \infty$ in which case the series converges absolutely and uniformly for almost all $(x,y) \in X \times X$.

This theorem allows us to define for a given kernel k a feature map Φ by

$$\begin{aligned} \Phi : X &\to l^2_{N_{\mathcal{H}}}, \\ x &\mapsto (\sqrt{\lambda_i} \psi_i(x))_{i=1,\dots,N_{\mathcal{H}}} \end{aligned}$$

for almost all $x \in X$. This map again satisfies

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle_{l^2_{N_{\mathcal{H}}}} .$$

Moreover this construction is particularly useful in applications as it allows for finite truncations while maintaining a given accuracy ε. This is due to the uniform convergence of the sequence. More precisely we have the following result.

Proposition 2.10 *(Approximate Mercer Feature Map)*
If k is a kernel satisfying the conditions of Proposition 2.9 then for any $\varepsilon > 0$ there exists an $n \in \mathbb{N}$ and a map Φ_n into l_n^2 such that

$$|k(x,y) - \langle \Phi_n(x), \Phi_n(y) \rangle_{l_n^2}| < \varepsilon \text{ for almost all } (x,y) \in X \times X.$$

2.4.1 The Kernel Trick

An important aspect to take away from the discussion above is that we can think about kernels as being inner products in some other space. It is this property that allows us to use kernels to extend support vector classification, and any other learning algorithm that solely depends on inner products, to a wider range of problems.

We have described in detail how it is possible to find a feature map for a given kernel so that the kernel corresponds to the inner product in the feature space. Conversely it is not hard to see that with a map

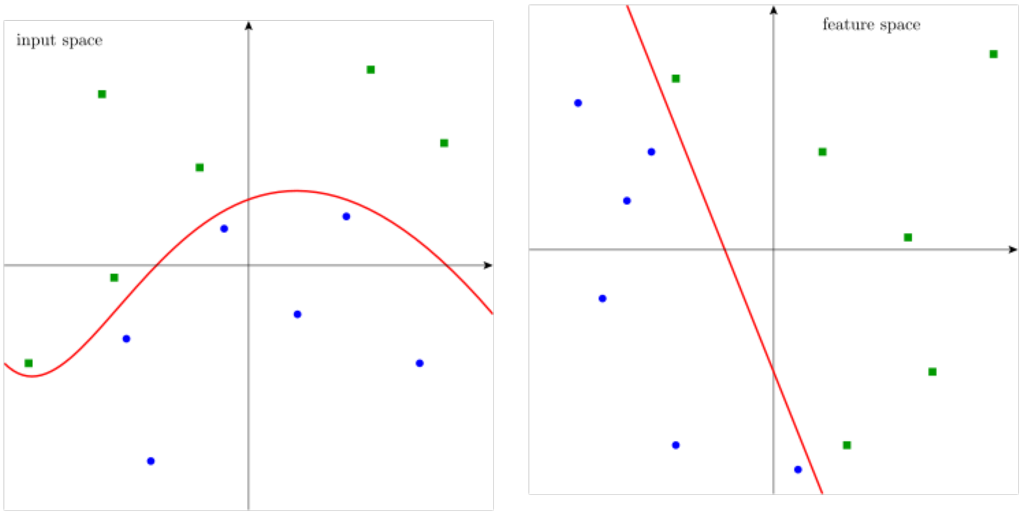
$$\begin{aligned} \Phi : X &\to \mathcal{H} \\ x &\mapsto \Phi(x) \end{aligned}$$

where $\mathcal{H}$ is a Hilbert space we obtain a positive definite kernel on X by setting

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} .$$

Now we are able to apply any algorithm that relies solely on inner products in the feature space $\mathcal{H}$, e.g. the SVM example from Section 2.2. The solution in terms of inner products in $\mathcal{H}$ we can then pull back to the RKHS on X generated by k. We obtain a solution in terms of kernels centred on points in X. One advantage of this construction is that it does not need any assumptions on X other than being a non-empty set. But it also leads to the so called kernel trick which is a fundamental reason for the success of kernel methods in learning. It states that for any algorithm which is stated purely in terms of a positive definite kernel k, another algorithm can be constructed by replacing the kernel k by a different kernel $\overline{k}$.

The most common application of the kernel trick is the one already described above, where the original algorithm is based on inner products. This is of particular relevance as inner products have a well understood geometric interpretation which makes them very suitable for the design of algorithms. Being able to replace the inner product by a positive definite kernel means that one can extend the algorithm past the linear case and use it to learn nonlinear functions of a type determined by the kernel. In the example of SVM classification above it means that we do not need the data to be separated by a hyperplane anymore but we want to find a feature space in which the data is separated by a hyperplane. This separating hyperplane in feature space then corresponds to a nonlinear decision boundary in the original space as illustrated in Fig. 2.



(a) The data might not be separated by a hyperplane in input space.

(b) One might be able to find a feature space in which the data is separated by a hyperplane.

Figure 2: Obtaining a non-linear decision boundary by mapping into a suitable feature space

But there are other advantages of the kernel trick. Sometimes it can be computationally more efficient to compute a kernel as a dot product in a different space. Moreover some feature maps induce certain additional structure on the data which can be used by an algorithm.

It is this variety in functions that can be learned and the variety in possible algorithms, combined with the simplicity and sometimes increased efficiency of implementations which have made kernel methods such an important tool in machine learning.

2.4.2 Examples

In this section we will briefly present some examples of common kernels. These examples are taken from [SS02] and we assume they are defined on a set $X \subseteq \mathbb{R}^d$.

- The homogeneous and inhomogeneous polynomial kernels

$$k(x,y) = \langle x,y\rangle^n\,,$$
$$k(x,y) = (\langle x,y\rangle + c)^n, \qquad c > 0.$$

- The Gaussian kernel

$$k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right), \qquad \sigma > 0.$$

- B-spline kernels of odd order

$$k(x,y) = B_{2p+1}(\|x-y\|_2), \qquad p \in \mathbb{N}.$$

We will see later that many of these kernels can also be used to generate reproducing kernel Banach spaces.

2.4.3 Representer Theorem

It is not surprising that a representer theorem as stated in Proposition 2.4 for general Hilbert spaces is true for RKHS. In this section we will state and prove the RKHS version of the representer theorem, once again following [SS02]. The proof is short and not difficult but it is important to note its essential ingredients, representation of point evaluations by inner products and orthogonal complements. This will motivate much of the theory developed in the following chapters in order to generalise the representer theorem to Banach spaces.

Theorem 2.11 *(Representer theorem)*

Let X be a non-empty set and $\mathcal{H}$ the RKHS on X associated to the kernel k. Consider the regularisation problem

$$\min\{\mathcal{E}_z(f) + \lambda\Omega(\|f\|_{\mathcal{H}}) \,:\, f \in \mathcal{H}\} \tag{8}$$

for $\mathcal{E}_z$ the empirical error for an arbitrary loss function $\mathcal{C}$, and a nondecreasing function $\Omega : [0, \infty) \to \mathbb{R}$. Then there always exists a minimiser $f_0 \in \mathcal{H}$ of Eq. (8) of the form

$$f_0(x) = \sum_{i=1}^{m} c_i k(x, x_i).$$

Proof:

Without loss of generality we can assume $\Omega(\|f\|_{\mathcal{H}}) = \widetilde{\Omega}(\|f\|_{\mathcal{H}}^2)$ because $\widetilde{\Omega}$ is nondecreasing if and only if Ω is.

Now decompose the RKHS into the span of the kernels functions centred at the training data and its orthogonal complement, i.e. we write $f \in \mathcal{H}$ as

$$f(x) = f_{\text{ker}}(x) + f_{\perp}(x) = \sum_{i=1}^{m} c_i k(x, x_i) + f_{\perp}(x),$$

with $c_i \in \mathbb{F}$ and $f_{\perp} \in \mathcal{H}$ such that $\langle f_{\perp}, k(\cdot, x_i)\rangle_{\mathcal{H}} = 0$ for all $i \in \mathbb{N}_m$.

By the reproducing property (i) we have for every data point $x_j \in X$

$$f(x_j) = \langle f(\cdot), k(\cdot, x_j)\rangle_{\mathcal{H}} = \sum_{i=1}^{m} c_i k(x_i, x_j) + \langle f_{\perp}(\cdot), k(\cdot, x_j)\rangle_{\mathcal{H}} = \sum_{i=1}^{m} c_i k(x_i, x_j).$$

Thus for any fixed $c_1, \ldots, c_m$ the loss function $\mathcal{C}(x_i, f(x_i), y_i)$ is constant and hence so is the empirical error $\mathcal{E}_z(f)$. But for any $f_{\perp}$ added to a fixed $f_{\text{ker}} \in \operatorname{span}\{k(\cdot, x_i) \,:\, i \in \mathbb{N}_m\}$ we see that

$$\Omega(\|f\|_{\mathcal{H}}) = \widetilde{\Omega}\left(\|\sum_{i=1}^{m} c_i k(\cdot, x_i)\|_{\mathcal{H}}^2 + \|f_{\perp}\|_{\mathcal{H}}^2\right) \geq \widetilde{\Omega}\left(\|\sum_{i=1}^{m} c_i k(\cdot, x_i)\|_{\mathcal{H}}^2\right),$$

which proves that Eq. (8) is minimised for $f_{\perp} = 0$.

❑

2.5 RKHS as Generalisation of Sobolev Spaces

We close this chapter by sketching how RKHS can in a way be seen as a generalisation of Sobolev spaces. We are following the discussion in [Wen04] where further details can be found. Throughout the section we will denote the Fourier transform of a function f by $\hat{f}$.

The Sobolev embedding theorem for $p > \frac{d}{2}$ says that $W^{p,2}(\mathbb{R}^d) \subseteq C(\mathbb{R}^d)$ in the sense that every equivalence class contains a continuous representer. This means, by always choosing the continuous representation, point evaluations make sense on $W^{p,2}(\mathbb{R}^d)$ for $p > \frac{d}{2}$ and it can be viewed as a Hilbert space of functions. We want to show that in this case $W^{p,2}(\mathbb{R}^d)$ can be made a RKHS. The essential result for this is the following proposition, which is theorem 10.12 from [Wen04].

Proposition 2.12

Let $k \in C(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ a real valued, positive definite function. Define the function space

$$\mathcal{H} = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \,:\, \frac{\hat{f}}{\sqrt{\hat{k}}} \in L^2(\mathbb{R}^d) \right\},$$

with the bilinear form

$$\langle f, g \rangle_{\mathcal{H}} = (2\pi)^{-\frac{d}{2}} \left\langle \frac{\hat{f}}{\sqrt{\hat{k}}}, \frac{\hat{g}}{\sqrt{\hat{k}}} \right\rangle_{L^2(\mathbb{R}^d)} = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \frac{\hat{f}(x)\overline{\hat{g}(x)}}{\hat{k}(x)} \,\mathrm{d}x.$$

Then $\mathcal{H}$ is a RKHS with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and reproducing kernel $k(\cdot - \cdot)$. In fact $\mathcal{H}$ is the RKHS for $k(\cdot - \cdot)$ as constructed in the Moore-Aronszajn theorem (Theorem 2.8).

Recall the Besov characterisation of Sobolev spaces, namely $W^{p,2}(\mathbb{R}^d)$ can be equivalently defined as

$$W^{p,2}(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d) \,:\, \hat{f}(\cdot)(1 + \|\cdot\|_2^2)^{\frac{p}{2}} \in L^2(\mathbb{R}^d) \right\}$$

with inner product

$$\langle f, g\rangle_{W^{p,2}(\mathbb{R}^d)} = (2\pi)^{-\frac{d}{2}} \int\limits_{\mathbb{R}^d} \hat{f}(x)\overline{\hat{g}(x)}(1 + \|x\|_2^2)^p \,\mathrm{d}x.$$

Now from Proposition 2.12 we see that for $p > \frac{d}{2}$ we obtain a positive definite function k by $\hat{k}(x) = (1 + \|x\|_2^2)^{-p}$ such that

$$\langle f, k(\cdot - y)\rangle_{W^{p,2}(\mathbb{R}^d)} = (2\pi)^{-\frac{d}{2}} \int\limits_{\mathbb{R}^d} \frac{\hat{f}(x)\overline{\hat{k}(x)e^{-iy\cdot x}}}{\hat{k}(x)} \,\mathrm{d}x = f(y)$$

and k generates a RKHS which coincides with $W^{p,2}(\mathbb{R}^d)$. In summary we obtain the following corollary, which is corollary 10.13 in [Wen04].

Corollary 2.13

If $k \in L^1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ is such that

$$c_1(1 + \|x\|_2^2)^{-p} \le \hat{k}(x) \le c_2(1 + \|x\|_2^2)^{-p}$$

for $x \in \mathbb{R}^d$, $p > \frac{d}{2}$ and two constants $0 \le c_1 \le c_2$ then the RKHS with reproducing kernel $k(\cdot - \cdot)$ coincides with the Sobolev space $W^{p,2}(\mathbb{R}^d)$ and the RKHS norm $\|\cdot\|_{\mathcal{H}}$ and the Sobolev norm are equivalent.

3 Semi-inner Product Spaces

We saw the usefulness of RKHS for learning in Chapter 2. The theory is very well developed and widely used in applications. In order to extend it to Banach spaces we need a tool to replace the structures of Hilbert spaces crucially used for RKHS. In particular we need a representation of point evaluations by elements of the space and a notion of orthogonality. Extending Hilbert space type arguments to Banach spaces is exactly what semi-inner products were introduced for. Lumer in [Lum61] was the first to define and discuss semi-inner products. Giles in [Gil67] then developed the theory further, obtaining a class of semi-inner product spaces which are "almost Hilbert". In this chapter we will mainly be following the paper by Giles, occasionally influenced by the presentations in [ZXZ09, Dra04]. The proofs are generally not difficult and often boil down to just a clever application of the Cauchy-Schwarz inequality but will be included here for completeness.

Definition 3.1 *(Semi-inner product)*

A semi-inner product (s.i.p.) on a vector space V is a map

$$[\cdot,\cdot]_V : V \times V \to \mathbb{F}$$

with the following properties:

(i) Linearity in the first argument:
$[\lambda x + \mu y, z]_V = \lambda\,[x,z]_V + \mu\,[y,z]_V$ for all $x,y,z \in V$ and $\lambda,\mu \in \mathbb{F}$.

(ii) Positive definiteness:
$[x,x]_V \geq 0$ and $[x,x]_V = 0 \Leftrightarrow x = 0$.

(iii) Cauchy-Schwarz inequality:
$|[x,y]_V|^2 \leq [x,x]_V\,[y,y]_V$.

In particular, in comparison to inner products, we have dropped conjugate symmetry which makes it necessary to assume a Cauchy-Schwarz inequality to be true.

We can make a further homogeneity assumption, which does not lead to any significant resitrictions as we will see shortly.

(iv) Homogeneity property:
$[x, \lambda y]_V = \overline{\lambda}\,[x, y]_V \quad$ for all $x, y \in V$ and $\lambda \in \mathbb{F}$.

A vector space V with a s.i.p. $[\cdot,\cdot]_V$ is called a semi-inner-product space (s.i.p. space).

It is worth noting that now the only difference to an inner product is the lack of additivity in the second argument. In fact this property exactly distinguishes semi-inner products from inner products.

Lemma 3.2
A semi-inner product is conjugate symmetric if and only if it is additive in its second argument.

Proof:
Note that assuming additivity in the second argument we get

$$[x + \lambda y, x + \lambda y]_V = [x, x]_V + [\lambda y, \lambda y]_V + \lambda\,[y, x]_V + \overline{\lambda}\,[x, y]_V\,.$$

But then by positive definiteness we have that the left hand side is real and so are $[x, x]_V$ and $[\lambda y, \lambda y]_V$. Thus we must also have $\lambda\,[y, x]_V + \overline{\lambda}\,[x, y]_V$ real. But setting $\lambda = 1$ we see this can only happen if $\operatorname{Im}[x, y]_V = -\operatorname{Im}[y, x]_V$ and setting $\lambda = i$ gives $\operatorname{Re}[x, y]_V = \operatorname{Re}[y, x]_V$, i.e. $[\cdot,\cdot]_V$ is symmetric and hence an inner product. ❑

As a first example consider lp for $p \in (1, \infty)$ with the semi-inner product

$$[x, y]_{lp} = \frac{\sum_i x_i y_i |y_i|^{p-2}}{\|y\|_p^{p-2}}$$

We will se further examples in Section 4. For Definition 3.1 to be useful we need it to be connected to the norm of the space in a similar manner as inner products are. This is indeed the case as can be seen from the following theorem.

Theorem 3.3

Any semi-inner product space V is a normed linear space with norm $\|x\|_V = [x,x]_V^{\frac{1}{2}}$. Conversely every normed linear space V can be made into a semi-inner-product space with the homogeneity property (iv). In general there may be infinitely many semi-inner products inducing the norm on V.

Proof:

It is easy to check that $\|x\|_V = [x,x]_V^{\frac{1}{2}}$ indeed defines a norm. On the other hand let V be a normed linear space with dual space V^*. For any fixed $x \in V$ by the Hahn-Banach theorem there exists at least one functional $L_x \in S_{V^*}$ such that $L_x(x) = \|x\|_V$. Fixing exactly one such functional L_x for each $x \in V$ would give us a semi-inner product satisfying properties (i) to (iii) by setting $[x,y]_V = \|y\|_V \cdot L_y(x)$.

To also satisfy property (iv) we instead only fix exactly one such functional for every $x \in S_V$ and for $x \in V$, writing $x = \lambda\widetilde{x}$ with $\widetilde{x} \in S_V, \lambda \in \mathbb{F}$, choose $L_x = \overline{\lambda}L_{\widetilde{x}}$. It is now again easy to check that $[x,y]_V = L_y(x)$ satisfies properties (i) to (iv) for a semi-inner product.

❑

The lack of uniqueness is rather unsatisfactory. It turns out that we can obtain uniqueness by imposing a continuity property in the second argument.

Definition 3.4 *(Continuous s.i.p. space)*

A semi-inner-product space is called a continuous s.i.p. space or uniformly continuous s.i.p. space if additionally to (i) to (iv) the following properties hold respectively

(v) (Uniform) continuity property in the second argument:

(v.a) Continuity:
$\operatorname{Re}[x,y+tx]_V \to \operatorname{Re}[x,y]_V$ for any real $t \to 0$ and for every $x,y \in S_V$;

(v.b) Uniform continuity:
The limit $\operatorname{Re}[x, y+tx]_V \to \operatorname{Re}[x,y]_V$ for any real $t \to 0$ is approached uniformly for every $x, y \in S_V$.

As the norm is induced by the semi-inner product the continuity property is closely linked to the regularity of the norm.

Theorem 3.5

A s.i.p. space V is continuous (uniformly continuous) if and only if the norm is Gâteaux (uniformly Fréchet) differentiable.
The differential for $x \neq 0$ is given by

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

Thus a Gâteaux differentiable normed vector space has a unique semi-inner product.

Proof:

Part 1: *(Continuity property ⇒ Differentiability of norm)*
For $x, y \in S_V$ and $t > 0$ by expanding $\frac{\|x+ty\|_V - \|x\|_V}{t}$ by $\|x\|_V$ and applying the Cauchy-Schwarz inequality we find

$$\begin{aligned} \frac{\|x+ty\|_V - \|x\|_V}{t} &\overset{(iii)}{\geq} \frac{|[x+ty,x]_V| - \|x\|_V^2}{t\|x\|_V} \\ &\geq \frac{\operatorname{Re}[x+ty,x]_V - \|x\|_V^2}{t\|x\|_V} \\ &= \frac{\operatorname{Re}[y,x]_V}{\|x\|_V}. \end{aligned} \tag{9}$$

But similarly as before, expanding by $\|x+ty\|_V$ and applying the Cauchy-

Schwarz inequality

$$\begin{aligned}\frac{\|x+ty\|_V - \|x\|_V}{t} &\overset{(iii)}{\leq} \frac{\|x+ty\|_V^2 - |[x,x+ty]_V|}{t\|x+ty\|_V} \\ &= \frac{\operatorname{Re}[x+ty,x+ty]_V - |[x,x+ty]_V|}{t\|x+ty\|_V} \\ &\leq \frac{|[x,x+ty]_V| + t\operatorname{Re}[y,x+ty]_V - |[x,x+ty]_V|}{t\|x+ty\|_V} \\ &= \frac{\operatorname{Re}[y,x+ty]_V}{\|x+ty\|_V}.\end{aligned} \tag{10}$$

By the continuity property we have

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

so we get that also

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

and the limit exists uniformly if it exists uniformly in the continuity property.

Part 2: *(Differentiability of norm ⇒ Continuity property)*
As we have seen in part 1 we have for $t > 0$

$$\frac{\|x+ty\|_V - \|x\|_V}{t} \geq \frac{\operatorname{Re}[y,x]_V}{\|x\|_V} \geq \frac{\|x-ty\|_V - \|x\|_V}{-t}$$

and hence

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

To relate this limit to the expression in the continuity property, we expand by $\|x+ty\|_V$ again to find in view of Eq. (9) for $x, y \in S_V$ and $t > 0$

$$\begin{aligned}\frac{\|x+ty\|_V - \|x\|_V}{t} &= \frac{\operatorname{Re}[x,x+ty]_V + t\operatorname{Re}[y,x+ty]_V - \|x\|_V\cdot\|x+ty\|_V}{t\|x+ty\|_V} &(11)\\ &= \frac{\operatorname{Re}[x,x+ty]_V - \|x\|_V\cdot\|x+ty\|_V}{t\|x+ty\|_V} + \frac{\operatorname{Re}[y,x+ty]_V}{\|x+ty\|_V} &(12)\\ &\geq \frac{\operatorname{Re}[y,x]_V}{\|x\|_V} &(13)\end{aligned}$$

and hence, denoting the two summands in (12) by a_t and b_t respectively we get

$$\lim_{t\searrow 0} \frac{\|x+ty\|_V - \|x\|_V}{t} = \limsup_{t\searrow 0} \left(a_t + b_t\right) \geq \liminf_{t\searrow 0} a_t + \limsup_{t\searrow 0} b_t.$$

We now bound $\liminf\limits_{t\searrow 0} a_t$. Rearranging (11) with the inequality (13) we obtain

$$\operatorname{Re}\left[x, x+ty\right]_V - \|x\|_V \cdot \|x+ty\|_V + t\left(\operatorname{Re}\left[y, x+ty\right]_V - \operatorname{Re}\left[y, x\right]_V \frac{\|x+ty\|_V}{\|x\|_V}\right) \geq 0.$$

In order to drop the last term we note that

$$\begin{aligned} |\operatorname{Re}\left[y, x+ty\right]_V - \operatorname{Re}\left[y, x\right]_V \frac{\|x+ty\|_V}{\|x\|_V}| &\leq \|y\|_V \cdot \|x+ty\|_V + \|y\|_V \cdot \|x\|_V \frac{\|x+ty\|_V}{\|x\|_V} \\ &\leq 2\|y\|_V \left(\|x\|_V + t\|y\|_V\right). \end{aligned}$$

As $x, y \in S_V$ this is bounded by $2(1+t)$. Since (9) and (10) imply that the term in parentheses is positive this means that as t goes to zero we can drop the last term to obtain

$$\liminf_{t\searrow 0} \left(\operatorname{Re}\left[x, x+ty\right]_V - \|x\|_V \cdot \|x+ty\|_V\right) \geq 0,$$

which immediately implies that

$$\liminf_{t\searrow 0} \left(\frac{\operatorname{Re}\left[x, x+ty\right]_V - \|x\|_V \cdot \|x+ty\|_V}{t\|x+ty\|_V}\right) \geq 0.$$

On the other hand (9) and (10) directly give that every b_t, and thus both $\limsup\limits_{t\searrow 0} b_t$ and $\liminf\limits_{t\searrow 0} b_t$, are bounded from below by $\frac{\operatorname{Re}[y,x]_V}{\|x\|_V}$.
Putting these together we obtain

$$\begin{aligned} \frac{\operatorname{Re}\left[y, x\right]_V}{\|x\|_V} = \lim_{t\searrow 0} \frac{\|x+ty\|_V - \|x\|_V}{t} &\geq \limsup_{t\searrow 0} \left(\frac{\operatorname{Re}\left[y, x+ty\right]_V}{\|x+ty\|_V}\right) \\ &\geq \frac{\operatorname{Re}\left[y, x\right]_V}{\|x\|_V} \end{aligned} \tag{14}$$

and hence equality all the way through.
Now for negative t the inequalities (9) and (10) get flipped and hence all the inequalities throughout the proof derived from them also flip. We thus get a negative but vanishing term when bounding the infimum and carrying

everything through we obtain the equality (14).
Putting both together we have indeed that

$$\lim_{t\to 0} \operatorname{Re}\,[y, x+ty]_V = \operatorname{Re}\,[y,x]_V$$

and the limit exists uniformly if it does in the differentiability property.

❑

3.1 Orthogonality

Having established uniqueness in Theorem 3.5 we are already in a good position, but many proofs in the kernel methods theory depend strongly on orthogonality. We are thus now going to establish a generalisation of orthogonality to s.i.p.-spaces.

Definition 3.6 *(Orthogonality)*

Let V be a continuous s.i.p. space. For $x, y \in V$ we say that x is normal to y and y is transversal to x if $[y,x]_V = 0$.
A vector $x \in V$ is normal to a subspace $W \subset V$ and W is transversal to x if x is normal to all $y \in W$.

James [Jam47] introduced a notion of orthogonality for general normed spaces which is equivalent to the inner product being zero in a Hilbert space. It turns out that the same equivalence holds true for the definition of orthogonality with respect to the semi-inner product of a continuous s.i.p. space.

Theorem 3.7

Let V be a continuous s.i.p. space and $x, y \in V$. Then x is normal to y if and only if

$$\|x+\lambda y\|_V \geq \|x\|_V \qquad \text{for all } \lambda \in \mathbb{F}. \tag{15}$$

Proof:

Part 1: *(x normal to y ⇒ Eq.* (15)*)*
If $[y,x]_V = 0$ then by the Cauchy-Schwarz property (iii)

$$\begin{aligned} \|x+\lambda y\|_V \cdot \|x\|_V &\overset{(iii)}{\geq} |[x+\lambda y, x]_V| \\ &= |\|x\|_V^2 + \lambda\,[y,x]_V| \\ &= \|x\|_V^2 \end{aligned}$$

so that indeed $\|x+\lambda y\|_V \geq \|x\|_V$.

Part 2: *(Eq.* (15) *⇒ x normal to y)*
If $\|x+\lambda y\|_V \geq \|x\|_V$ for all $\lambda \in \mathbb{F}$ then also

$$\|x+\lambda y\|_V^2 - \|x\|_V \cdot \|x+\lambda y\|_V \geq 0$$

and thus using the Cauchy-Schwarz property (iii) again we have

$$\operatorname{Re}\,[x, x+\lambda y]_V + \operatorname{Re}\lambda\,[y, x+\lambda y]_V - |[x, x+\lambda y]_V| \geq 0.$$

But since $|[x,x+\lambda y]_V| \geq \operatorname{Re}\,[x,x+\lambda y]_V$ this implies that

$$\operatorname{Re}\lambda\,[y, x+\lambda y]_V \geq 0 \qquad \text{for all } \lambda \in \mathbb{F}. \tag{16}$$

For real λ we see from Eq. (16)

$$\operatorname{Re}\,[y, x+\lambda y]_V \begin{cases} \geq 0 & \text{for } \lambda \geq 0, \\ \leq 0 & \text{for } \lambda \leq 0. \end{cases}$$

But in view of the continuity property (v.a) we must have

$$\begin{aligned} \operatorname{Re}\,[y, x+\lambda y]_V &\searrow \operatorname{Re}\,[y,x]_V \ \text{ as } \lambda \searrow 0, \\ \operatorname{Re}\,[y, x+\lambda y]_V &\nearrow \operatorname{Re}\,[y,x]_V \ \text{ as } \lambda \nearrow 0. \end{aligned}$$

Hence $\operatorname{Re}\,[y,x]_V = 0$.

For purely imaginary λ we write $\lambda = i\widetilde{\lambda}$ with $\widetilde{\lambda} \in \mathbb{R}$. Then

$$\operatorname{Re}\widetilde{\lambda}\,[iy, x+\lambda y]_V = \widetilde{\lambda}\operatorname{Re}\left[iy, x+\widetilde{\lambda} iy\right]_V \geq 0$$

and arguing as before using the continuity property (v.a) we get

$$\operatorname{Re}[iy,x]_V = -\operatorname{Im}[y,x]_V = 0.$$

Putting both together we have $[y,x]_V = 0$ as required.

❑

The orthogonality relation (15) was first studied by James in [Jam47] and he proved that it is additive if and only if the norm is Gâteaux differentiable. But by the linearity in the first argument of the semi-inner product our definition of orthogonality is clearly additive, i.e. if x is normal to both y and z then it is normal to all $\lambda y + \mu z$ for $\lambda, \mu \in \mathbb{F}$. Thus we immediately get that the continuity property (v.a) and Gâteaux differentiability of the norm are equivalent.

The equivalence with James orthogonality gives a good intuition what it means to be orthogonal with respect to the semi-inner product. The fact that adding any multiple of y to x does not decrease the norm means that the affine line $x + ty$ is a tangent to the ball of radius $\|x\|$ at x. Thus a vector x is normal to y if y is tangent at x as illustrated in Fig. 3. The figure also makes clear why the orthogonality is not symmetric anymore.

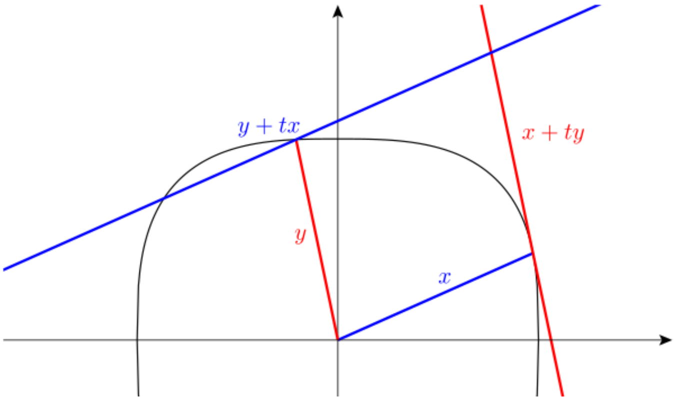


Figure 3: x is normal to y if y is a tangent at x

Having defined orthogonality between vectors and a vector and a subspace in Definition 3.6 we are now able to introduce a notion of an orthogonal complement which was a crucial ingredient of the proofs in the Hilbert space setting. Let $U, W \subset V$ be subspaces of V. Following Definition 3.6 we say U is orthogonal to W if every $x \in U$ is orthogonal to every $y \in W$. More precisely if $[y,x]_V = 0$ for all $x \in U$ and all $y \in W$. This leads to the following definition of an orthogonal complement.

Definition 3.8 *(Orthogonal Complement)*
Let W be a subspace of a continuous s.i.p. space V. Then the orthogonal complement of W in V is the set

$$W^{\perp} = \{y \in V \,:\, [x,y]_V = 0 \; \forall \, x \in W\}$$

i.e. all vectors $y \in V$ which are orthogonal to every vector $x \in W$.

But there is more one can say. The notion of orthogonality of subspaces generalises naturally to s.i.p spaces V which are not continuous via James orthogonality. More precisely we can say U is orthogonal to W in the sense of James if $\|x + \lambda y\|_V \geq \|x\|_V$ for all $x \in U$ and all $y \in W$. Moreover we can define orthogonality with respect to any given semi-inner product on the space V. Faulkner in [Fau77] shows that in this case a subspace U is orthogonal to another subspace W in the sense of James if and only if there exists a semi-inner product on V such that U is orthogonal to W with respect to that semi-inner product.

3.2 Duality

Since the use of RKHS crucially relies on the identification of points with bounded linear functionals on the space we finally want to prove a Riesz representation theorem for s.i.p. spaces. For this we need to further assume that the space is uniformly convex (see Section 5.4).

We also need the following little lemma which is well know for uniformly convex Banach spaces and it is straightforward to check that it agrees with our definition of orthogonality for s.i.p. spaces.

Lemma 3.9

In a continuous s.i.p. space which is uniformly convex and complete for every proper closed subspace there exists a non-zero normal vector.

Proof:

It is known that in a uniformly convex Banach space V for a proper closed subspace W and any $x \notin W$ there exists a unique non-zero closest point, i.e. a vector $y_0 \in W$ such that $\|x - y_0\|_V = \inf\{\|x - y\|_V \,:\, y \in W\}$. Thus setting $z_0 = x - y_0$ is clearly normal to W by Theorem 3.7.

❑

Theorem 3.10 *(Riesz representation theorem)*

Let V be a uniformly convex, complete, continuous s.i.p. space. Then for every $L \in V^*$ there exists a unique vector $y \in V$ such that

$$L(x) = [x, y]_V \qquad \text{for all } x \in V.$$

Furthermore

$$\|y\|_V = \|L\|_{V^*}.$$

Proof:

Part 1: *(Existence)*

If $L \equiv 0$ we can choose $y = 0$ so we can without loss of generality assume $L(x) \neq 0$ for some $x \in V$. But then $\ker(L)$ is a proper closed subspace of V and thus by Lemma 3.9 there exists a non-zero normal vector y_0, i.e.

$$[z, y_0]_V = 0 \tag{17}$$

for all $z \in \ker(L)$.

We can represent every $x \in V$ in the form $x = z + \lambda y_0$ with $z \in \ker(L)$ and $\lambda = \frac{L(x)}{L(y_0)}$ so that

$$L(x) = L(z + \lambda y_0) = L(z) + \lambda L(y_0). \tag{18}$$

Thus we only need to consider choices of y for two easy cases.

- If $x = z \in \ker(L)$ we can choose any $y = \mu y_0$ with $\mu \in \mathbb{F}$ as then by (17)
$$[x,y]_V = \overline{\mu}\,[z,y_0]_V = 0 = L(z).$$
- If on the other hand $x = y_0$ then we see that for $y = \frac{\overline{L(y_0)}}{\|y_0\|_V^2} y_0$ we get precisely
$$[x,y]_V = \frac{L(y_0)}{\|y_0\|_V^2}\,[y_0,y_0]_V = L(y_0).$$

Plugging this into Eq. (18) we see that with $y = \frac{\overline{L(y_0)}}{\|y_0\|_V^2} y_0$ we indeed obtain
$$L(x) = [z,y]_V + \lambda\,[y_0,y]_V = [z+\lambda y_0, y]_V = [x,y]_V$$
as required.

Part 2: *(Uniqueness)*
Suppose there exist two vectors $y, \widetilde{y} \in V$, $y \neq \widetilde{y}$ such that
$$L(x) = [x,y]_V = [x,\widetilde{y}]_V$$
for all $x \in V$. Then choosing $x = y$ we see that
$$\|y\|_V^2 = [y,y]_V = [y,\widetilde{y}]_V \leq \|y\|_V \cdot \|\widetilde{y}\|_V,$$
so $\|y\|_V \leq \|\widetilde{y}\|_V$. Similarly, choosing $x = \widetilde{y}$, we obtain the reverse inequality, thus $\|y\|_V = \|\widetilde{y}\|_V$ and further $\|y\|_V \cdot \|\widetilde{y}\|_V = [y,\widetilde{y}]_V$. But since we have
$$\begin{aligned}\|\widetilde{y}+y\|_V \cdot \|y\|_V \geq |[\widetilde{y}+y,y]_V| &= |[\widetilde{y},y]_V + [y,y]_V| \\ &= \|\widetilde{y}\|_V \cdot \|y\|_V + \|y\|_V^2 = \|y\|_V\left(\|\widetilde{y}\|_V + \|y\|_V\right)\end{aligned}$$
we also have $\|\widetilde{y}+y\|_V = \|\widetilde{y}\|_V + \|y\|_V$. By uniform convexity, which in particular implies strict convexity, this means that in fact $\widetilde{y} = y$.

Part 3: *(Norm equality)*
From our choice of y in part 1 we immediately see that
$$\|y\|_V^2 = \left[\frac{\overline{L(y_0)}}{\|y_0\|_V^2} y_0, \frac{\overline{L(y_0)}}{\|y_0\|_V^2} y_0\right]_V = \frac{|L(y_0)|^2}{\|y_0\|_V^4}\,[y_0,y_0]_V = \frac{|L(y_0)|^2}{\|y_0\|_V^2} \leq \frac{\|L\|_{V^*}^2 \|y_0\|_V^2}{\|y_0\|_V^2}.$$

Conversely by construction we have

$$\|L\|_{V^*} \leq |L\left(\frac{y}{\|y\|_V}\right)| = |\left[\frac{y}{\|y\|_V}, y\right]_V| = |\frac{1}{\|y\|_V}[y,y]_V| = \|y\|_V,$$

which completes the proof.

❑

But even if the space is not uniformly convex and uniformly smooth so that the above Riesz representation theorem does not hold there is a quite nice characterisation of semi-inner products by bounded linear functionals. The following results are taken from [Dra04].

Proposition 3.11

Every semi-inner product $[\cdot,\cdot]_V$ on a normed vector space V is of the form

$$[x,y]_V = L_y(x), \qquad x,y \in V,$$

where L_y is such that $\|L_y\|_{V^*} = \|y\|_V$ and $L_y(y) = \|y\|_V^2$.

It is clear that every L_y as in the proposition induces a semi-inner product. James [Jam64] proved the following characterisation of reflexivity.

Proposition 3.12

A Banach space $\mathcal{B}$ is reflexive if and only if every nonzero continuous linear functional L attains its norm in at least one point $x \in \mathcal{B}$, i.e. $L(x) = \|L\|_{\mathcal{B}^*} \cdot \|x\|_{\mathcal{B}}$.

Putting both together we obtain that a Banach space $\mathcal{B}$ is reflexive if and only if every continuous linear functional L on $\mathcal{B}$ is represented by a semi-inner product $[\cdot,\cdot]_L$ and a point $y_L \in \mathcal{B}$ by $L(x) = [x, y_L]_L$. This was proved by Faulkner in [Fau77].

Moreover the duality of semi-inner products and norm-attaining linear functionals allows for a nice characterisation of sets of norm attainment for linear functionals. In [Sai18] Sain proves that a linear functional $L \in \mathcal{B}^*$ attains its norm at a point $y \in \mathcal{B}$ if and only if there exists a semi-inner product $[\cdot, \cdot]_L$ on $\mathcal{B}$ such that $L(x) = \left[x, \frac{\|L\|}{\|y\|} y\right]_L$ for all $x \in \mathcal{B}$.

3.3 Uniform S.I.P. Spaces

With a Riesz representation theorem at hand we get an identification of V and V^* similarly to Hilbert spaces. By the Cauchy-Schwarz inequality the map $y \mapsto [y, x]_V$ is clearly a continuous linear functional on V, denoted by x^*. By the Riesz representation theorem we have that $x \mapsto x^*$ is a isometric isomorphism from V to V^*. Thus by definition

$$[x, y]_V = y^*(x) = (x, y^*)_V \text{ for all } x, y \in V. \tag{19}$$

Note that a uniformly convex Banach space is reflexive. Further a normed vector space is uniformly Fréchet differentiable if and only if its dual is uniformly convex (c.f. Section 5.4). Thus our assumption implies uniform convexity of the dual and by reflexivity of the space we get Fréchet differentiability of the dual space. So if V is a uniformly convex, uniformly Fréchet differentiable Banach space, then so is V^*. We thus have a unique semi-inner product on V^* which is given by

$$[x^*, y^*]_{V^*} = [y, x]_V \,. \tag{20}$$

In view of Eq. (19) we have

$$[x^*, y^*]_{V^*} = [y, x]_V = x^*(y)$$

and hence linearity in the first argument follows easily by noting

$$[x^* + y^*, z^*]_{V^*} = (x^* + y^*)(z) = x^*(z) + y^*(z) = [x^*, z^*]_{V^*} + [y^*, z^*]_{V^*} \,.$$

All other properties of a semi-inner product follow immediately from $[\cdot, \cdot]_V$ satisfying them. Note that by linearity of $[\cdot, \cdot]_V$

$$[z^*, x^*]_{V^*} + [z^*, y^*]_{V^*} = [x, z]_V + [y, z]_V = [x + y, z]_V = [z^*, (x + y)^*]_{V^*}$$

but in general $(x+y)^* \neq x^* + y^*$ as the duality pairing is linear if and only if $[\cdot,\cdot]_V$ is an inner product. Thus this only implies linearity if we are in the Hilbert space case. This is the crucial difference in these constructions to the usual Hilbert space setting.

We are now in a good position to establish a theory very similar to RKHS as presented above. We have some fundamental structure that is very similar to Hilbert spaces which allows us to apply many standard Hilbert space techniques. To simplify a bit we will call spaces which allow all the above constructions, i.e. spaces which are uniformly convex and uniformly smooth, uniform as is made precise in the following definition.

Definition 3.13 *(Uniform Banach space)*

We say a space V is uniform if it is a uniformly convex and uniformly Fréchet differentiable Banach space.

Thus in particular we call a s.i.p. space V a uniform s.i.p. space if it is uniformly continuous and the induced norm is uniformly convex and complete.

Note that all lp spaces for $p \in (1,\infty)$ with the semi-inner product given above are uniform s.i.p. spaces.

4 Reproducing Kernel Banach Spaces

The theory of reproducing kernel Hilbert spaces presented in Section 2.4 is well established and has been widely used in applications. For various reasons it is desirable to try and generalise the theory to be applicable to a wider range of problems. In particular the possibility to learn in Banach spaces rather than Hilbert spaces would be important. Firstly there may be cases in which it is impossible to embed the data into a Hilbert space, as done in the constructions of Section 2.4 above. This could happen e.g. due to some intrinsic structure of the data. Since Banach spaces are far less restrictive, an embedding into a Banach space might still be possible. Secondly, even if an embedding of the data into a Hilbert space is possible, in applications certain properties of the norm of the function space are often desirable. As an example consider the l_1 norm, which is very widely used as it leads to sparsity of the solution. Thirdly any two Hilbert spaces of the same finite or infinite dimension are isometrically isomorphic. This is far from true for Banach spaces. This means Banach spaces possess a much richer geometric structure. This additional structure may be useful in the development of new learning algorithms. Lastly, as we will see in Section 7, it turns out that if the learning is based on the representer theorem it is in general actually the function space and not the regulariser that determines the solution. This means that changing the regulariser does not change the solution, but we need to learn in a different function space to obtain a different solution.

When starting to think about generalising the RKHS setting to Banach spaces one might first think of using Definition 2.5. More precisely one might try to simply assume the space to be a Banach space such that point evaluations are continuous. But if we then think of $C[0,1]$, the space of continuous functions on $[0,1]$ equipped with the maximum norm, we find that the reproducing kernel would need to be the delta distribution, which is not a function. This shows that we need a way to replace the representation of point evaluations given by the Riesz representation theorem and the inner product for RKHS. It is clear from the presentations in Chapter 3 that semi-inner products provide exactly such a representation. Another, even more generally applicable approach is to use the duality pairing and an isomorphic identification of the dual space with another function space. We will pursue both approaches throughout this chapter, showing how the semi-inner product identification can build on the duality pairing to obtain a very generally applicable theory.

4.1 Reflexive Reproducing Kernel Banach Spaces

We begin by first developing a theory for reflexive reproducing kernel Banach spaces, and then build upon it to construct uniform reproducing kernel Banach spaces. In view of the semi-inner product theory introduced in Section 3 it is clear that for uniform Banach spaces we can hope to obtain much of the same structure as we have seen for RKHS. To obtain a theory that can extend beyond uniform Banach spaces we are starting the discussion by using a general isomorphic identification of the dual space with another function space and then extend this by using the Riesz representation theorem for uniform Banach spaces as this identification. The discussions throughout this section will follow the work by Zhang, Xu and Zhang [ZXZ09] and Zhang and Zhang [ZZ12].

Definition 4.1 *(Reproducing Kernel Banach Space)*

Let X be an arbitrary, non-empty set. A reproducing kernel Banach space (RKBS) is a reflexive Banach space $\mathcal{B}$ of functions $f : X \to \mathbb{F}$ for which $\mathcal{B}^*$ is isometrically isomorphic to a Banach space $\mathcal{B}^\sharp$ of functions on X and such that all point evaluation functionals δ_x and $\delta_x^\sharp$ are continuous on $\mathcal{B}$ and $\mathcal{B}^\sharp$ respectively.

The identification $\mathcal{B}^\sharp$ of the dual $\mathcal{B}^*$ does not need to be unique. All identifications are obviously also isometrically isomorphic to each other so the definition is independent of the particular representation we choose. Hence we can, and will, think of it as having been fixed. We can then make the

identification implicit by treating elements of $\mathcal{B}^*$ as functions on X.
This definition suffices to have a kernel which is in some sense similar to the kernel we saw in the Hilbert space setting. This will be made precise in the following theorem. We will briefly comment on parts of the proof to illustrate the necessity of the assumptions made in Definition 4.1.

Theorem 4.2

Let $\mathcal{B}$ be a RKBS on X. Then there exists a unique function $k : X \times X \to \mathbb{F}$, the reproducing kernel, such that

(i) Reproducing property:
For all $x \in X$, $k(x,\cdot) \in \mathcal{B}$ and $k(\cdot,x) \in \mathcal{B}^*$ and we have

$$f(x) = (f, k(\cdot,x))_{\mathcal{B}} \text{ for all } f \in \mathcal{B}, \tag{21}$$

$$f^*(x) = (k(x,\cdot), f^*)_{\mathcal{B}} \text{ for all } f^* \in \mathcal{B}^*. \tag{22}$$

So in particular

$$k(x,y) = (k(x,\cdot), k(\cdot,y))_{\mathcal{B}}\,. \tag{23}$$

(ii) k spans $\mathcal{B}$ and $\mathcal{B}^*$:

$$\mathcal{B} = \overline{\operatorname{span}}\{k(x,\cdot) \,:\, x \in X\}, \tag{24}$$

$$\mathcal{B}^* = \overline{\operatorname{span}}\{k(\cdot,x) \,:\, x \in X\}.$$

Proof *(Sketch)*:
Clearly in (21) $k(\cdot,y)$ "is" the delta distribution $\delta_y(f) = f(y)$. Through the isomorphism we have a corresponding $k_y \in \mathcal{B}^\sharp$ so that we have a function we can evaluate by setting $k(x,y) = k_y(x)$ similarly to the Hilbert space case.
This means $k(\cdot,x) \in \mathcal{B}^*$ and it further corresponds to a function in $\mathcal{B}^\sharp$. To obtain a symmetry similar to the Hilbert space case and in particular (23) we also need to represent functions in $\mathcal{B}^\sharp$ i.e. (22). This motivates the assumption of point evaluations being continuous on $\mathcal{B}^\sharp$.
Since $\mathcal{B}^\sharp$ and $\mathcal{B}^*$ are isometrically isomorphic, the delta distribution δ_y acting

on $\mathcal{B}^\sharp$ also defines a continuous linear functional on $\mathcal{B}^*$, i.e. an element of $(\mathcal{B}^*)^*$. This shows why we assume $\mathcal{B}$ to be reflexive. Using reflexivity we obtain a unique $k_y \in \mathcal{B}$ which allows to define k similarly to before. It is not hard to check that this way we obtain a unique, well defined $k : X \times X \to \mathbb{F}$ with the desired properties.

❑

Note that this result, unlike the Hilbert space case, is only true in one direction, i.e. while there is only one reproducing kernel for any given RKBS there may be different RKBS having the same reproducing kernel.

We saw in Section 2.4 that one can easily obtain reproducing kernels by embedding the data into a Hilbert space via a feature map Φ. Zhang, Xu and Zhang in [ZXZ09] give a similar construction for RKBS via embedding the data into a reflexive Banach space and its dual space. We will now briefly present this construction.

Theorem 4.3

Let V be a reflexive Banach space with dual space V^* and suppose there exist maps $\Phi : X \to V$ and $\Phi^* : X \to V^*$ such that

$$\overline{\operatorname{span}}\Phi(X) = V, \qquad \overline{\operatorname{span}}\Phi^*(X) = V^*.$$

Then $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ where

$$\begin{aligned}\mathcal{B} &= \{(u, \Phi^*(\cdot))_V \,:\, u \in V\},\\ \|(u, \Phi^*(\cdot))_V\|_{\mathcal{B}} &= \|u\|_V\end{aligned}$$

is a RKBS on X with dual space $(\mathcal{B}^*, \|\cdot\|_{\mathcal{B}^*})$ where

$$\begin{aligned}\mathcal{B}^* &= \{(\Phi(\cdot), u^*)_V \,:\, u^* \in V^*\},\\ \|(\Phi(\cdot), u^*)_V\|_{\mathcal{B}^*} &= \|u^*\|_{V^*}.\end{aligned}$$

The duality pairing is given by

$$((u, \Phi^*(\cdot))_V, (\Phi(\cdot), u^*)_V)_{\mathcal{B}} = (u, u^*)_V \quad u \in V, u^* \in V^*.$$

The reproducing kernel k for $\mathcal{B}$ is

$$k(x,y) = (\Phi(x), \Phi^*(y))_V \quad x,y \in X.$$

While this theory of RKBS is obviously a generalisation of the Hilbert space setting, it also clearly lacks some of its elegance and applicability. This strongly motivates combining these definitions with the theory of semi-inner products presented in Section 3 and leveraging the Riesz representation theorem for a convenient representation of the dual space.

4.1.1 S.I.P. Reproducing Kernel Banach Spaces

As mentioned before, Lumer [Lum61] introduced semi-inner products specifically to extend Hilbert-space like arguments to a wide range of Banach spaces. In view of the theory of RKHS from Section 2.4 and the first results for RKBS just presented it is clear that they also provide a powerful tool to obtain some of the elegance of the theory of RKHS while maintaining a lot of the generality of Banach spaces. Thus we will from now on consider RKBS which are uniformly convex and uniformly Fréchet differentiable. In this case the existence of a Riesz representation theorem provides a convenient identification $\mathcal{B}^\sharp$ of $\mathcal{B}^*$, and the semi-inner product takes the place of the duality pairing. Following the convention at the end of Section 3 we will call such spaces uniform RKBS.

To show that using semi-inner products in conjunction with the above constructions is useful, we first need to establish that we can obtain a kernel function with representing properties with respect to the semi-inner product. The connection between the semi-inner product and the duality paring Eq. (19) then links the two kernels.

Theorem 4.4

Let $\mathcal{B}$ be a uniform RKBS on X and k its reproducing kernel. Then there exists a unique function

$$\kappa : X \times X \to \mathbb{F}$$

the s.i.p. kernel, such that for all $x \in X$, $\kappa(x, \cdot) \in \mathcal{B}$ and

$$f(x) = [f, \kappa(x, \cdot)]_{\mathcal{B}} \text{ for all } f \in \mathcal{B}, x \in X. \tag{25}$$

Moreover

$$k(\cdot, x) = (\kappa(x, \cdot))^* \text{ for all } x \in X \tag{26}$$

and

$$f^*(x) = [k(x, \cdot), f]_{\mathcal{B}} \text{ for all } f \in \mathcal{B}, x \in X. \tag{27}$$

Proof:
By the Riesz representation theorem for every $x \in X$ there exists a unique function $\kappa_x \in \mathcal{B}$ such that

$$f(x) = \delta_x(f) = [f, \kappa_x]_{\mathcal{B}} \text{ for all } f \in \mathcal{B}.$$

As before for $x, y \in X$ we define $\kappa : X \times X \to \mathbb{F}$ by $\kappa(x, y) = \kappa_x(y)$. We immediately have uniqueness, $\kappa(x, \cdot) \in \mathcal{B}$ and (25) holds. It remains to prove the "moreover" part.
To prove (26) for each $x \in X$ we look at the action of $(\kappa(x, \cdot))^*$ on an arbitrary fixed $f \in \mathcal{B}$

$$(f, (\kappa(x, \cdot))^*)_{\mathcal{B}} \overset{(19)}{=} [f, \kappa(x, \cdot)]_{\mathcal{B}} \overset{(25)}{=} f(x) \overset{(21)}{=} (f, k(\cdot, x))_{\mathcal{B}} .$$

Since x and f were arbitrary this implies (26).
For (27) we simply note that

$$f^*(x) \overset{(22)}{=} (k(x, \cdot), f^*)_{\mathcal{B}} \overset{(19)}{=} [k(x, \cdot), f]_{\mathcal{B}} .$$

❑

Clearly the most interesting case is when $\kappa = k$, in which case it is called the *s.i.p. reproducing kernel*. In particular in that case we have

$$k(x, y) = [k(x, \cdot), k(y, \cdot)]_{\mathcal{B}} , \qquad x, y \in X.$$

Again Zhang, Xu and Zhang in [ZXZ09] show that one can construct uniform s.i.p. spaces via a feature map which embeds the data into a uniform Banach space. This provides a nice characterisation of s.i.p. reproducing kernels, similar to the one we have seen before for RKHS and reflexive RKBS.

Theorem 4.5

Let X be an arbitrary, non-empty set, V a uniform Banach space and Φ a map from X to V. Denote by Φ^* the map from X to V^* defined by $\Phi^*(x) = (\Phi(x))^*$. Assume that

$$\overline{\operatorname{span}}\Phi(X) = V \quad \text{and} \quad \overline{\operatorname{span}}\Phi^*(X) = V^*.$$

Then

$$\mathcal{B} = \{[u, \Phi(\cdot)]_V : u \in V\}$$

is a uniform s.i.p. space with semi-inner product

$$[[u, \Phi(\cdot)]_V, [v, \Phi(\cdot)]_V]_{\mathcal{B}} = [u, v]_V .$$

Its dual is given by

$$\mathcal{B}^* = \{[\Phi(\cdot), u]_V : u \in V\}$$

with semi-inner product

$$[[\Phi(\cdot), u]_V, [\Phi(\cdot), v]_V]_{\mathcal{B}^*} = [v, u]_V .$$

The duality pairing is given by

$$([u, \Phi(\cdot)]_V, [\Phi(\cdot), v]_V)_{\mathcal{B}} = [u, v]_V . \tag{28}$$

The reproducing kernel k and s.i.p. kernel κ coincide so that we have the s.i.p. reproducing kernel

$$k(x, y) = [\Phi(x), \Phi(y)]_V . \tag{29}$$

Proof:

Part 1: *($\mathcal{B}$ is a uniform s.i.p. space)*
We first show that the map $u \mapsto [u, \Phi(\cdot)]_V$ is well defined. If $u = 0$ then clearly $[u, \Phi(\cdot)]_V = 0$ for all $x \in X$. Conversely fix $u \in V$ and assume $[u, \Phi(\cdot)]_V = 0$ for all $x \in X$. Then by Eq. (20) also $[\Phi^*(\cdot), u^*]_{V^*} = 0$ for all $x \in X$. By density of $\Phi^*(X)$ and linearity of $[\cdot,\cdot]_V$ this means $u^* = 0$ and hence, as the map $u \mapsto u^*$ is an isometry, $u = 0$. Thus $u \mapsto [u, \Phi(\cdot)]_V$ is a bijection from V to $\mathcal{B}$.
We now show that $[\cdot,\cdot]_{\mathcal{B}}$ is indeed a semi-inner product. This will in particular imply that the map $u \mapsto [u, \Phi(\cdot)]_V$ is isometric as then for $f = [u, \Phi(\cdot)]_V$

$$\|f\|_{\mathcal{B}}^2 = [f, f]_{\mathcal{B}} = [u, u]_V = \|u\|_V^2. \tag{30}$$

To this end let $f = [u, \Phi(\cdot)]_V$, $g = [v, \Phi(\cdot)]_V$ and $h = [w, \Phi(\cdot)]_V \in \mathcal{B}$. All properties of $[\cdot,\cdot]_{\mathcal{B}}$ follow directly from the respective properties of $[\cdot,\cdot]_V$. More precisely, for linearity in the first argument we observe that

$$\begin{aligned}
[\lambda f + g, h]_{\mathcal{B}} &= [\lambda [u, \Phi(\cdot)]_V + [v, \Phi(\cdot)]_V, [w, \Phi(\cdot)]_V]_{\mathcal{B}} \\
&= [\lambda u + v, w]_V \\
&= \lambda [u, w]_V + [v, w]_V \\
&= \lambda [f, h]_{\mathcal{B}} + [g, h]_{\mathcal{B}} .
\end{aligned}$$

Positive definiteness of $[\cdot,\cdot]_{\mathcal{B}}$ is immediate from positive definiteness of $[\cdot,\cdot]_V$ because $u \mapsto [u, \Phi(\cdot)]_V$ is a bijection which maps 0_V to $0_{\mathcal{B}}$.
The Cauchy-Schwarz inequality holds since

$$|[f, g]_{\mathcal{B}}|^2 = |[u, v]_V|^2 \leq [u, u]_V \cdot [v, v]_V = [f, f]_{\mathcal{B}} \cdot [g, g]_V .$$

And finally homogeneity in the second argument is verified by noting that

$$[f, \lambda g]_{\mathcal{B}} = [[u, \Phi(\cdot)]_V, [\lambda v, \Phi(\cdot)]_V]_{\mathcal{B}} = [u, \lambda v]_V = \overline{\lambda} [u, v]_V = \overline{\lambda} [f, g]_{\mathcal{B}} .$$

Thus $[\cdot,\cdot]_{\mathcal{B}}$ indeed defines a semi-inner product and by Eq. (30) completeness, uniform Fréchet differentiability and uniform convexity of the induced norm $\|\cdot\|_{\mathcal{B}}$ follow directly from $\|\cdot\|_V$ having these properties.

Part 2: *($\mathcal{B}^*$ is a uniform s.i.p. space)*
The arguments are very similar as for $\mathcal{B}$. Again if $u = 0$ then clearly $[\Phi(\cdot), u]_V = 0$ for all $x \in X$ and conversely for fixed $u \in V$ if $[\Phi(\cdot), u]_V = 0$ for all $x \in X$ then by density of $\Phi(X)$ and linearity of $[\cdot,\cdot]_V$ we have $u = 0$. Thus $u \mapsto [\Phi(\cdot), u]_V$ is a bijection from V to $\mathcal{B}^*$.

Similarly one establishes all the properties of a semi-inner product for $[\cdot,\cdot]_{\mathcal{B}^*}$ by using the respective property of $[\cdot,\cdot]_V$.

We set $f^* = [\Phi(\cdot), u]_V$, $g^* = [\Phi(\cdot), v]_V$, $h^* = [\Phi(\cdot), u]_V \in \mathcal{B}^*$. To avoid confusion from the non-linearity of the duality mapping we will momentarily denote the dual element of $u \in V$ by $F_u \in V^*$.

When proving linearity we need to be a little more careful here as the order of arguments gets switched from $[\cdot,\cdot]_V$ to $[\cdot,\cdot]_{\mathcal{B}^*}$. We will exploit (20) and the fact that $[f^*, g^*]_{\mathcal{B}^*} = [v, u]_V = [F_u, F_v]_{V^*}$.

We observe that from Eq. (20) we get that $[\Phi(\cdot), u]_V = [F_u, \Phi^*(\cdot)]_{V^*}$ so that

$$\begin{aligned}
[\lambda f^* + g^*, h^*]_{\mathcal{B}^*} &= [\lambda\,[\Phi(\cdot), u]_V + [\Phi(\cdot), v]_V, [\Phi(\cdot), w]_V]_{\mathcal{B}^*} \\
&= [[\lambda F_u, \Phi^*(\cdot)]_{V^*} + [F_v, \Phi^*(\cdot)]_{V^*}, [F_w, \Phi^*(\cdot)]_{V^*}]_{\mathcal{B}^*} \\
&= [[\lambda F_u + F_v, \Phi^*(\cdot)]_{V^*}, [F_w, \Phi^*(\cdot)]_{V^*}]_{\mathcal{B}^*} \\
&= [\lambda F_u + F_v, F_w]_{V^*} \\
&= \lambda\,[F_u, F_w]_{V^*} + [F_v, F_w]_{V^*} \\
&= \lambda\,[w, u]_V + [w, v]_V \\
&= \lambda\,[f^*, h^*]_{\mathcal{B}^*} + [g^*, h^*]_{\mathcal{B}^*}\,.
\end{aligned}$$

Positive definiteness and the Cauchy-Schwarz inequality follow in exactly the same way as in part 1 and also homogeneity is clear since

$$[f^*, \lambda g^*]_{\mathcal{B}^*} = \left[[\Phi(\cdot), u]_V, \left[\Phi(\cdot), \overline{\lambda} v\right]_V\right]_{\mathcal{B}^*} = \left[\overline{\lambda} v, u\right]_V = \overline{\lambda}\,[f^*, g^*]_{\mathcal{B}^*}\,.$$

Thus as before we have that $[\cdot,\cdot]_{\mathcal{B}^*}$ defines a semi-inner product. Completeness, uniform Fréchet differentiability and uniform convexity of the induced norm $\|\cdot\|_{\mathcal{B}^*}$ follow directly from $\|\cdot\|_V$ having these properties.

Part 3: *($\mathcal{B}^*$ is indeed the dual of $\mathcal{B}$ with duality pairing* (28)*)*
By noting that

$$|([u,\Phi(\cdot)]_V,[\Phi(\cdot),v]_V)_{\mathcal{B}}| = |[u,v]_V| \leq \|u\|_V\cdot\|v\|_V = \|[u,\Phi(\cdot)]_V\|_{\mathcal{B}}\cdot\|[\Phi(\cdot),v]_V\|_{\mathcal{B}^*}$$

it is clear that every $[\Phi(\cdot),v]_V \in \mathcal{B}^*$ is a continuous linear functional on $\mathcal{B}$. Since

$$\begin{aligned}([u,\Phi(\cdot)]_V,[\Phi(\cdot),v]_V)_{\mathcal{B}} = [u,v]_V &= [[u,\Phi(\cdot)]_V,[v,\Phi(\cdot)]_V]_{\mathcal{B}}\\ &= ([u,\Phi(\cdot)]_V,([v,\Phi(\cdot)]_V)^*)_{\mathcal{B}}\end{aligned}$$

we have $[\Phi(\cdot),v]_V = ([v,\Phi(\cdot)]_V)^*$ for all $v \in V$.
As $u \mapsto [u,\Phi(\cdot)]_V$, $u \mapsto [\Phi(\cdot),u]_V$ and $u \mapsto u^*$ are bijections this means that $\mathcal{B}^*$ is the dual space of $\mathcal{B}$ with duality pairing (28) as desired.

Part 4: *(The s.i.p. kernel is given by* (29)*)*
Let $f = [u,\Phi(\cdot)]_V$. By the definition of the duality pairing we have

$$f(y) = [u,\Phi(y)]_V = ([u,\Phi(\cdot)]_V,[\Phi(\cdot),\Phi(y)]_V)_{\mathcal{B}} = (f,[\Phi(\cdot),\Phi(y)]_V)_{\mathcal{B}}\,.$$

This is exactly the reproducing property (21) so that

$$k(\cdot,y) = [\Phi(\cdot),\Phi(y)]_V\,.$$

Similarly using the definition of the semi-inner product on $\mathcal{B}$ we see that

$$f(x) = [u,\Phi(x)]_V = [[u,\Phi(\cdot)]_V,[\Phi(x),\Phi(\cdot)]_V]_{\mathcal{B}} = [f,[\Phi(x),\Phi(\cdot)]_V]_{\mathcal{B}}$$

which is precisely the reproducing property (25) so that

$$\kappa(x,\cdot) = [\Phi(x),\Phi(\cdot)]_V\,.$$

But since $(\kappa(x,\cdot))^* = ([\Phi(x),\Phi(\cdot)]_V)^* = [\Phi(\cdot),\Phi(x)]_V$ the s.i.p kernel and the reproducing kernel coincide and

$$k(x,y) = [\Phi(x),\Phi(y)]_V\,.$$

❑

It turns out that we also have a converse result, which provides us with a precise characterisation of s.i.p. reproducing kernels.

Theorem 4.6

A function k on $X \times X$ is a s.i.p. reproducing kernel if and only if it is of the form

$$k(x,y) = [\Phi(x), \Phi(y)]_V$$

for a uniform Banach space V, the feature space, and a mapping, the feature map, $\Phi : X \to V$ such that $\overline{\operatorname{span}}\Phi(X) = V$ and $\overline{\operatorname{span}}\Phi^*(X) = V^*$.

Proof:

We proved in Theorem 4.5 that any function of the form specified in this theorem is a s.i.p. reproducing kernel. Thus we only need to construct a feature space V and feature map Φ for a given s.i.p. reproducing kernel k of a uniform RKBS $\mathcal{B}$ on X.

Let $V = \mathcal{B}$ and set $\Phi(x) = k(x,\cdot)$. Using the reproducing property (25) we see immediately that

$$k_x(y) = k(x,y) = [k(x,\cdot), k(y,\cdot)]_{\mathcal{B}} = [\Phi(x), \Phi(y)]_{\mathcal{B}} .$$

Furthermore by the spanning property of reproducing kernels (24) span $\Phi(X)$ is dense in V.

The converse, that also span $\Phi^*(X)$ is dense in V^*, is proved by contradiction. Assume span $\Phi^*(X)$ is not dense in V^*. Then by the Hahn-Banach theorem there exists a nontrivial linear functional $F \in (V^*)^*$ which is zero on $\overline{\operatorname{span}}\Phi^*(X)$, i.e. $F(\Phi^*(x)) = 0$ for all $x \in X$. By the Riesz representation theorem there exists a unique $f^* \in V^*$ s.t. $F(g^*) = [g^*, f^*]_{V^*}$ for every $g^* \in V^*$ and hence in particular

$$[\Phi^*(x), f^*]_{V^*} = 0 \tag{31}$$

for all $x \in X$. Further as F is nontrivial so is f. Applying the Riesz representation theorem again we obtain a nontrivial $f \in V = \mathcal{B}$ corresponding to

f^* for which, using the reproducing property, we find for every $x \in X$

$$f(x) = [f, k(x,\cdot)]_{\mathcal{B}} = [f, \Phi(x)]_V = [\Phi^*(x), f^*]_{V^*}.$$

As we have seen in (31) above this means $f(x) = 0$ for every $x \in X$. This contradicts f being nontrivial.

❑

4.1.2 Examples

As an example of these constructions consider the following example given by Zhang, Xu and Zhang [ZXZ09]. Let $X = \mathbb{R}$ and $V = L^p(\mathbb{I})$ with $\mathbb{I} = [-\frac{1}{2}, \frac{1}{2}]$. Denote the Fourier transform of a function f by $\hat{f}$ and the inverse Fourier transform by $\check{f}$.
With

$$\Phi(x)(t) = e^{-2\pi ixt},\ \Phi^*(x)(t) = e^{2\pi ixt}, \quad x \in \mathbb{R}, t \in \mathbb{I},$$

we obtain a RKBS

$$\mathcal{B} = \{f \in C(\mathbb{R}) : \operatorname{supp} \hat{f} \subseteq \mathbb{I}, \hat{f} \in L^p(\mathbb{I})\}$$

with dual space

$$\mathcal{B}^* = \{g \in C(\mathbb{R}) : \operatorname{supp} \check{g} \subseteq \mathbb{I}, \check{g} \in L^q(\mathbb{I})\}$$

and kernel

$$k(x,y) = (\Phi(x), \Phi^*(y))_{L^p(\mathbb{I})} = \frac{\sin \pi(x-y)}{\pi(x-y)} = \operatorname{sinc}(x-y).$$

The duality pairing is given by

$$(f,g)_{\mathcal{B}} = \int_{\mathbb{I}} \hat{f}(t)\check{g}(t)\,\mathrm{d}t \quad f \in \mathcal{B}, g \in \mathcal{B}^*.$$

For $p = q = 2$ this construction corresponds to the usual space of band-limited functions. For other values of p we maintain the property of a Fourier transform with bounded support but consider a different L^p norm making $\mathcal{B}$ isometrically isomorphic to $L^p(\mathbb{I})$.

Since unlike Hilbert spaces of the same dimension the $L^p(\mathbb{I})$ spaces are not isomorphic to each other, they exhibit a richer geometric variety which is potentially useful for the development of new learning algorithms.
Note that above example is one dimensional for notational simplicity and similar constructions yield RKBS isomorphic to $L^p_\mu(\mathbb{R}^d)$ where μ is a finite positive Borel measure on $\mathbb{R}^d$ as shown in Zhang and Zhang [ZZ12]. The corresponding RKBS $\mathcal{B}$ consists of functions of the form

$$f_u(x) = \frac{1}{\mu(\mathbb{R}^d)^{\frac{p-2}{p}}} \int\limits_{\mathbb{R}^d} u(t) e^{i\langle x,t\rangle}\,\mathrm{d}\mu(t), \quad x \in \mathbb{R}^d, u \in L^p_\mu(\mathbb{R}^d)$$

with semi-inner product

$$[f_u, f_v]_{\mathcal{B}} = \frac{1}{\|v\|^{p-2}_{L^p_\mu(\mathbb{R}^d)}} \int\limits_{\mathbb{R}^d} u(t)\overline{v(t)}|v(t)|^{p-2}\,\mathrm{d}\mu(t).$$

The reproducing kernel is given by

$$k(x,y) = \frac{1}{\mu(\mathbb{R}^d)^{\frac{p-2}{p}}} \int\limits_{\mathbb{R}^d} e^{i\langle y-x,t\rangle}\,\mathrm{d}\mu(t), \quad x, y \in \mathbb{R}^d.$$

For $d = 1$ and μ the Lebesgue measure on $[-\frac{1}{2}, \frac{1}{2}]$ this reduces to the above example.
The duality mapping in L^p spaces is given by $f^* = \frac{\overline{f}|f|^{p-2}}{\|f\|_p^{p-2}}$ which in the given example means that for an element $f_u \in \mathcal{B}$ the corresponding dual element is given by

$$f_u^* = \frac{\overline{u} \cdot |u|^{p-2}}{\|u\|_p^{p-2}}.$$

Further the duality mapping in a reflexive Banach space is self-inverse so

$$(f_u^*)^* = f_u.$$

4.1.3 Representer Theorem

We finally return to the question of minimisers of the regularisation problem, this time posed in a uniform RKBS, i.e.

$$\min\{\mathcal{E}_z(f) + \lambda\Omega(\|f\|_{\mathcal{B}}) : f \in \mathcal{B}\}$$

for some empirical data $z = (z_1, \ldots, z_m) = ((x_1, y_1), \ldots, (x_m, y_m)) \subset X \times Y$, $m \in \mathbb{N}$, and a uniform RKBS $\mathcal{B}$. Specifically we want to obtain an analogous result to Theorem 2.11, the representer theorem for RKHS. Throughout this thesis we will generally not be concerned about the existence of minimisers. Results on existence and uniqueness of minimisers can be found in e.g. [ZZ12] which is also the paper this section is based on.

Theorem 4.7 *(Representer theorem)*

Let X be a non-empty set and $\mathcal{B}$ a uniform RKBS with s.i.p. reproducing kernel k. Consider the regularisation problem

$$\min\{\mathcal{E}_z(f) + \lambda\Omega(\|f\|_{\mathcal{B}}) : f \in \mathcal{B}\} \tag{32}$$

for $\mathcal{E}_z$ the empirical error for an arbitrary loss function $\mathcal{C}$, and a nondecreasing function $\Omega : [0, \infty) \to \mathbb{R}$. Then there always exists a minimiser $f_0 \in \mathcal{B}$ of (32) such that the dual element of f_0 is of the form

$$f_0^*(x) = \sum_{i=1}^m c_i k(x_i, x)^* = \sum_{i=1}^m c_i k(x, x_i).$$

The proof of this theorem we are going to present is a combination of the proofs for strictly increasing regularisers and nondecreasing regularisers from the paper by Zhang and Zhang [ZZ12]. Their proof for the case of nondecreasing regularisers is shorter and in a way easier than the one presented here, essentially deducing the statement directly from minimal norm interpolation. We chose to combine the two arguments, largely following their argument for strictly increasing regularisers, as it is more instructive. In particular it shows how the Hahn-Banach theorem and duality arguments replace the traditional Hilbert space arguments.

Proof *(Of Theorem 4.7)*:

Let f be a solution of Eq. (32). The set $I_f = \{g \in \mathcal{B} : g(x_i) = f(x_i)\}$ is closed, convex and nonempty as it contains f. Since $\mathcal{B}$ is uniformly convex there exists a unique $f_0 \in I_f$ such that

$$\|f_0 - 0\| = \operatorname{dist}(0, I_f) = \min\{\|g - 0\| : g \in I_f\}. \tag{33}$$

In other words f_0 is the unique solution to the minimal norm interpolation problem.
Now since $f_0 \in I_f$ we have that $f_0(x_i) = f(x_i)$ for all x_i and thus

$$\mathcal{E}_z(f_0) = \mathcal{E}_z(f).$$

Moreover $\|f_0\| \leq \|f\|$ and hence since Ω is nondecreasing

$$\Omega(\|f_0\|_{\mathcal{B}}) \leq \Omega(\|f\|_{\mathcal{B}}).$$

Thus f_0 is also a minimiser of Eq. (32). Assume for contradiction that f_0 is not of the form stated in the theorem. That is $f_0^* \notin \mathcal{K}$ for

$$\mathcal{K} = \operatorname{span}\{k(x_i, \cdot)^* = k(\cdot, x_i) \,:\, i \in \mathbb{N}_m\} \subset \mathcal{B}^*.$$

Since $\mathcal{K}$ is a closed and convex subspace of $\mathcal{B}^*$ by the Hahn-Banach separation theorem there exists a functional $T \in (\mathcal{B}^*)^*$ and a constant $s \in \mathbb{R}$ such that

$$\operatorname{Re} T(f_0^*) < s \leq \operatorname{Re} T(u) \quad \text{ for all } u \in \mathcal{K}.$$

Firstly, since $\mathcal{B}$ is reflexive, there exists $g \in \mathcal{B}$ such that

$$T(v) = v(g) \quad \text{ for all } v \in \mathcal{B}^*.$$

Since u is chosen from an entire subspace we can freely multiply by scalars $\lambda \in \mathbb{F}$ to see that

$$s \leq \operatorname{Re} T(\lambda u) = \operatorname{Re} \lambda T(u).$$

By choosing λ real or purely imaginary and sending it off to either plus or minus infinity we see that this can only hold if $T(u) = 0$ for all $u \in \mathcal{K}$. Thus we must also have $s \leq 0$.
Now by definition of the duality pairing (19) and the reproducing property (25) have

$$0 = T(k(x_i, \cdot)^*) = k(x_i, \cdot)^*(g) \overset{(19)}{=} [g, k(x_i, \cdot)]_{\mathcal{B}} \overset{(25)}{=} g(x_i).$$

But this means that

$$f + tg \in I_f \quad \forall t \in \mathbb{F}.$$

We further have that

$$\operatorname{Re}[g, f_0]_{\mathcal{B}} = \operatorname{Re} f_0^*(g) = \operatorname{Re} T(f_0^*) < s \le 0.$$

But by Theorem 3.5 this means that g provides a direction in which the derivative of the norm at f_0 is negative, more specifically

$$\lim_{t \searrow 0} \frac{\|f_0 + tg\|_B - \|f_0\|_B}{t} = \frac{\operatorname{Re}[g, f_0]_{\mathcal{B}}}{\|f_0\|_B} < 0.$$

Thus for t small enough we have $\|f_0 + tg\|_B < \|f_0\|_B$. But this contradicts Eq. (33), f_0 being the unique minimiser of the minimal norm interpolation problem.

❑

Remark 4.8

The separating hyperplane in some sense replaces the orthogonal decomposition from the Hilbert space case. Since the subspace spanned by the kernel function centred at the data points gets put into the kernel of the linear functional we get that the corresponding function can be thought of as being in the orthogonal complement as in the Hilbert space case. From the orientation of the hyperplane, i.e. the assumed minimiser being in the negative half space we get the required decrease in norm.

These arguments fail for $\operatorname{span}\{k(x_i, \cdot)\} \subset \mathcal{B}$ due to the nonlinearity of the duality mapping. One can pull back the functional defining the hyperplane via the Riesz representation theorem but the lack of symmetry of the semi-inner product, or equivalently the nonlinearity of the second argument and hence of the duality mapping, mean that we cannot guarantee that the obtained function is zero on the data points. This is why we only get a representation of the dual element of the minimiser as a linear combination of the s.i.p. reproducing kernel centred at the data points, i.e. effectively a linear combinations of point evaluations at the data points.

Even though this form of the representer theorem only characterises the dual element of the minimiser, it is a very powerful result when combined with a characterisation equation. We will close this section by presenting, without proof, a case which was presented in [ZZ12], in which we can combine the representer theorem with a characterisation equation to obtain a system of equations which determines the minimiser of Eq. (32). To be able to obtain such a system of equations we restrict ourselves further and consider only cases in which there exists a unique minimiser. More specifically as sufficient conditions for the existence and uniqueness of a minimiser of (32) we assume both $\mathcal{E}_z$ and Ω to be continuous and convex and Ω to be strictly increasing with $\lim\limits_{t\to\infty}\Omega(t) = \infty$. Further details about the characterisation equations and proofs about these conditions can be found in [ZZ12].

Recall from Definition 2.1 that $\mathcal{C}(f(x_i), y_i)$ measures the loss incurred by using $f(x_i)$ to predict the true output y_i.

Theorem 4.9 *(Characterization equations)*

Assume $\mathcal{C}(\cdot,\cdot)$ is a loss function which is differentiable and convex with respect to its first variable for every $x_i \in X$.

Assume further that $\Omega : [0,\infty) \to \mathbb{R}$ is a strictly increasing, differentiable, convex function and satisfies $\lim\limits_{t\to\infty}\Omega(t) = \infty$.

Then $f_0 \neq 0$ is the minimiser of Eq. (32) if and only if

$$\sum_{i=1}^{m} \frac{\partial c}{\partial a}(f_0(x_i), y_i) k(x_i, \cdot)^* + \lambda \frac{\Omega'(\|f_0\|_{\mathcal{B}})}{\|f_0\|_B} f_0^* = 0 \tag{34}$$

where $\frac{\partial c}{\partial a}$ is used to denote the partial derivative with respect to the first variable of c.

The zero function $f_0 = 0$ is the minimiser of Eq. (32) if and only if

$$\|T\|_{\mathcal{B}^*} \le \lambda\Omega'(0)$$

where $T \in \mathcal{B}^*$ is for every $f \in \mathcal{B}$ defined as

$$T(f) = \sum_{i=1}^{m} \frac{\partial c}{\partial a}(0, y_i) f(x_i)$$

where again $\frac{\partial c}{\partial a}$ is used to denote the partial derivative with respect to the first variable of c.

We now combine this result with the representer theorem. To this end note that

$$f_0(x_i) = [f_0, k(x_i,\cdot)]_{\mathcal{B}} = [k(x_i,\cdot)^*, f_0^*]_{\mathcal{B}^*} = \left[k(x_i,\cdot)^*, \sum_{k=1}^m c_k k(x_k,\cdot)^*\right]_{\mathcal{B}^*}.$$

Plugging the statement of the representer theorem and the characterisation (34) into this and noting further that the duality mapping is isometric we obtain

$$\begin{aligned}
&\sum_{i=1}^m \frac{\partial c}{\partial a}(f_0(x_i), y_i) k(x_i,\cdot)^* + \lambda \frac{\Omega'(\|f_0\|_{\mathcal{B}})}{\|f_0\|_{\mathcal{B}}} f_0^* = 0\\
\Leftrightarrow &\sum_{i=1}^m \frac{\partial c}{\partial a}\left(\left[k(x_i,\cdot)^*, \sum_{k=1}^m c_k k(x_k,\cdot)^*\right]_{\mathcal{B}^*}, y_i\right) k(x_i,\cdot)^* + \lambda \frac{\Omega'(\|f_0^*\|)}{\|f_0^*\|} \sum_{j=1}^m c_j k(x_j,\cdot)^* = 0\\
\Leftrightarrow &\sum_{i=1}^m \left(\frac{\partial c}{\partial a}\left(\left[k(x_i,\cdot)^*, \sum_{k=1}^m c_k k(x_k,\cdot)^*\right]_{\mathcal{B}^*}, y_i\right) + \lambda \frac{\Omega'(\|f_0^*\|_{\mathcal{B}^*})}{\|f_0^*\|_{\mathcal{B}^*}} c_i\right) k(x_i,\cdot)^* = 0.
\end{aligned}$$

Assuming linear independence of the $k(x_i,\cdot)^* = k(\cdot,x_i)$ this means that the coefficients of the minimiser satisfy the system of equations

$$\frac{\partial c}{\partial a}\left(\left[k(x_i,\cdot)^*, \sum_{k=1}^m c_k k(x_k,\cdot)^*\right]_{\mathcal{B}^*}, y_i\right) + \lambda \frac{\Omega'\left(\|\sum_{k=1}^m c_k k(x_k,\cdot)^*\|_{\mathcal{B}^*}\right)}{\|\sum_{k=1}^m c_k k(x_k,\cdot)^*\|_{\mathcal{B}^*}} c_i = 0.$$

We thus have obtained the desired system of equations depending on the data points which characterises the solution. In contrast to the Hilbert space case the problem here is often non-convex or nonlinear so one will need to come up with more powerful algorithms to find a solution.

4.2 Non-reflexive Reproducing Kernel Banach Spaces

Reflexive reproducing kernel Banach spaces and in particular uniform RKBS provide a significant extension of the theory of kernel methods to a large variety of Banach spaces. While many common Banach spaces are reflexive, l^1, which is widely used in applications, is not. It is thus desirable to extend the theory further. As stated in Section 3.2 a Banach space is reflexive if

and only if every bounded linear functional is represented by a semi-inner product. We thus can not rely on representing the point evaluations by semi-inner products but have to use the duality pairing.
The constructions we are going to present in this section follow the papers by Song, Zhang and Hickernell [SZH13], and Georgiev, Sánchez-Gonzáles and Pardalos [GSGP14].

Definition 4.10

Let $\mathcal{B}$ and $\mathcal{B}^\sharp$ be Banach spaces of functions on X. The pair $(\mathcal{B}, \mathcal{B}^\sharp)$ is a pair of reproducing kernel Banach spaces (RKBS) with reproducing kernel $k : X \times X \to \mathbb{F}$ if

(i) Point evaluation functionals are continuous on $\mathcal{B}$ and $\mathcal{B}^\sharp$,

(ii) $k(x, \cdot) \in B$ for all $x \in X$ and $k(\cdot, y) \in \mathcal{B}^\sharp$ for all $y \in X$;

(iii) There is a bilinear form $(\cdot, \cdot)_k$ on $\mathcal{B} \times \mathcal{B}^\sharp$ such that

$$\begin{aligned} (f, k(\cdot, y))_k &= f(y) \qquad \forall y \in X, f \in \mathcal{B}, \\ (k(x, \cdot), g)_k &= g(x) \qquad \forall x \in X, g \in \mathcal{B}^\sharp. \end{aligned}$$

We will sketch the constructions presented in the aforementioned papers [SZH13, GSGP14] to obtain a pair of RKBS in the sense of this definition for a given kernel k. Further details and proofs can be found in those papers.
The construction starts from a reproducing kernel and yields a pair of Banach spaces of functions $(\mathcal{B}, \mathcal{B}^\sharp)$ so that the spans of the kernel functions satisfy a certain density property in those spaces and point evaluations are represented by the kernel via a bilinear form on the two spaces.

Let $k : X \times X \to \mathbb{F}$ be a function and define

$$\mathcal{B}_0 = \operatorname{span}\{k(x, \cdot) : x \in X\}, \qquad \mathcal{B}_0^\sharp = \operatorname{span}\{k(\cdot, y) : y \in X\}.$$

Assume that there exists a norm $\|\cdot\|_{\mathcal{B}_0}$ on $\mathcal{B}_0$ such that point evaluations are continuous. The function k and the norm $\|\cdot\|_{\mathcal{B}_0}$ are all that needs to

be known when constructing a pair of RKBS. The rest of the construction follows without any further explicit input.
We define a bilinear form on $\mathcal{B}_0 \times \mathcal{B}_0^\sharp$ by

$$(f,g)_k = \left(\sum_{i=1}^{n} s_i k(x_i,\cdot), \sum_{j=1}^{m} t_j k(\cdot,y_j)\right)_k = \sum_{i=1}^{n}\sum_{j=1}^{m} s_i t_j k(x_i,y_j)$$

for all $f \in \mathcal{B}_0$ and $g \in \mathcal{B}_0^\sharp$. Similarly to previous sections one can easily check that this defines a well-defined bilinear form which satisfies the reproducing property (iii) of Definition 4.10. This allows us to define a norm on the space $\mathcal{B}_0^\sharp$ by

$$\|g\|_{\mathcal{B}_0^\sharp} = \sup_{f \in \mathcal{B}_0, \|f\|_{\mathcal{B}_0} \le 1} |(f,g)_k|.$$

This is a well-defined norm and one can show that point evaluations are continuous on $(\mathcal{B}_0^\sharp, \|\cdot\|_{\mathcal{B}_0^\sharp})$ if and only if point evaluations are continuous on $(\mathcal{B}_0, \|\cdot\|_{\mathcal{B}_0})$. It is also clear that we get a Cauchy-Schwarz type inequality

$$|(f,g)_k| \le \|f\|_{\mathcal{B}_0} \cdot \|g\|_{\mathcal{B}_0^\sharp}, \qquad f \in \mathcal{B}_0, g \in \mathcal{B}_0^\sharp.$$

It remains to complete the pair $(\mathcal{B}_0, \mathcal{B}_0^\sharp)$ to a pair of RKBS $(\mathcal{B}, \mathcal{B}^\sharp)$ in the sense of Definition 4.10. We need to be careful how we obtain the completion to make sure that the required properties, which we built into $\mathcal{B}_0$ and $\mathcal{B}_0^\sharp$, are preserved. Since point evaluations are continuous for any Cauchy sequence $(f_n) \subset \mathcal{B}_0$ the sequence $(f_n(x))$ is Cauchy in $\mathbb{F}$. One can check that the limit $f(x) = \lim_{n\to\infty} f_n(x)$ is well-defined.
We can thus complete $\mathcal{B}_0$ by setting

$$\mathcal{B} = \left\{ f : X \to \mathbb{F} : \exists \text{ Cauchy sequence } (f_n) \subset \mathcal{B}_0 \text{ s.t. } f(x) = \lim_{n\to\infty} f_n(x) \forall x \in X \right\}$$

with the norm $\|f\|_{\mathcal{B}} = \lim_{n\to\infty} \|f_n\|_{\mathcal{B}_0}$.
This completion process yields a well-defined norm $\|\cdot\|_{\mathcal{B}}$ on a Banach space of functions $\mathcal{B}$ such that point evaluations are continuous if and only if the norm $\|\cdot\|_{\mathcal{B}_0}$ satisfies a norm consistency property. More precisely for any Cauchy sequence $(f_n) \subset \mathcal{B}_0$ such that $f_n(x) \xrightarrow[n\to\infty]{} 0$ for every $x \in X$ we have that $\|f_n\|_{\mathcal{B}_0} \xrightarrow[n\to\infty]{} 0$.

We complete $\mathcal{B}_0^*$ by the same procedure and find that the norm $\|\cdot\|_{\mathcal{B}_0^\sharp}$ as defined above automatically satisfies the same norm consistency property so that the resulting space $(\mathcal{B}^\sharp, \|\cdot\|_{\mathcal{B}^\sharp})$ is a Banach space of functions such that point evaluations are continuous. Either by a repeated application of the Hahn-Banach theorem or a repeated limit argument we extend the bilinear form $(\cdot,\cdot)_k$ to $\mathcal{B} \times \mathcal{B}^\sharp$ such that the norm $\|\cdot\|_{\mathcal{B}^\sharp}$ is still such that

$$\|g\|_{\mathcal{B}^\sharp} = \sup_{f \in \mathcal{B}, \|f\|_{\mathcal{B}} \leq 1} |(f,g)_k|$$

and the Cauchy-Schwarz type inequality

$$|(f,g)_k| \leq \|f\|_{\mathcal{B}} \cdot \|g\|_{\mathcal{B}^\sharp} \qquad f \in \mathcal{B}, g \in \mathcal{B}^\sharp$$

holds. Finally one checks that also the reproducing properties

$$\begin{aligned} (f, k(\cdot,y))_k &= f(y) \qquad \forall y \in X, f \in \mathcal{B}, \\ (k(x,\cdot), g)_k &= g(x) \qquad \forall x \in X, g \in \mathcal{B}^\sharp \end{aligned}$$

hold and $(\mathcal{B}, \mathcal{B}^\sharp)$ are a pair of RKBS.
Moreover we note that with these constructions the space $\mathcal{B}^\sharp$ is isometrically and linearly embedded into the dual space of $\mathcal{B}$ by the map

$$\begin{aligned} \mathcal{L} : \mathcal{B}^\sharp &\to \mathcal{B}^*, \\ (\mathcal{L}g)(f) &= (f,g)_k, \qquad f \in \mathcal{B}, g \in \mathcal{B}^\sharp. \end{aligned}$$

One can show that this map is an isomorphism if and only if for any proper closed subspace $V \subsetneq \mathcal{B}$ the orthogonal space

$$V^\perp = \{g \in \mathcal{B}^\sharp : (f,g)_k = 0 \, \forall f \in V\} \subset \mathcal{B}^\sharp$$

is nontrivial. In this case we note that $\mathcal{B}$ satisfies all assumptions of Definition 4.1 except for reflexivity.

The above constructions are summarised in the following proposition.

Proposition 4.11

Let $k : X \times X \to \mathbb{F}$ be a function and let

$$\mathcal{B}_0 = \operatorname{span}\{k(x,\cdot) : x \in X\}, \qquad \mathcal{B}_0^\sharp = \operatorname{span}\{k(\cdot,y) : y \in X\}.$$

Assume that there exists a norm $\|\cdot\|_{\mathcal{B}_0}$ on $\mathcal{B}_0$ for which point evaluations are continuous on $\mathcal{B}_0$ and so that for any Cauchy sequence (f_n) in $(\mathcal{B}_0, \|\cdot\|_{\mathcal{B}_0})$ such that $f_n(x) \to 0$ we have $\|f_n\|_{\mathcal{B}_0} \to 0$.
Then there are Banach spaces completions $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ and $(\mathcal{B}^\sharp, \|\cdot\|_{\mathcal{B}^\sharp})$ of $\mathcal{B}_0$ and $\mathcal{B}_0^\sharp$ respectively, such that $(\mathcal{B}, \mathcal{B}^\sharp)$ is a pair of RKBS with reproducing kernel k. Furthermore

$$|(f,g)_k| \le \|f\|_{\mathcal{B}} \cdot \|g\|_{\mathcal{B}^\sharp} \qquad \forall f \in \mathcal{B}, g \in \mathcal{B}^\sharp$$

and

$$\|g\|_{\mathcal{B}^\sharp} = \sup_{f \in \mathcal{B}, \|f\|_{\mathcal{B}} \le 1} |(f,g)_k| \qquad \forall g \in \mathcal{B}^\sharp.$$

The space $\mathcal{B}^\sharp$ is isometrically isomorphic to $\mathcal{B}^*$ if and only if for any proper closed subspace V of $\mathcal{B}$ the orthogonal space $V^\perp$ is nontrivial. In this case the maps $\phi : X \to \mathcal{B}$ and $\phi^* : X \to \mathcal{B}^*$ given by

$$\phi(x) = k(x,\cdot), \qquad \phi^*(y) = \mathcal{L}(k(\cdot,y))$$

define feature maps such that

$$k(x,y) = (\phi(x), \phi^*(y))_k .$$

As mentioned previously we are extending the theory to non-reflexive Banach spaces in particular to include l^1 spaces. An l^1-type space on an arbitrary set X is defined in [SZH13] as a Banach space of functions on X which are integrable with respect to the counting measure on X. More precisely

$$l^1(X) = \{c = (c_x \in \mathbb{F} : x \in X) : \|c\|_{l^1(X)} = \sum_{x \in X} |c_x| < \infty\}$$

While X may be uncountable in this definition, for every $c \in l^1(X)$ the support of c, $\operatorname{supp}(c) = \{x \in X : c_x \neq 0\}$ must be countable.

We are now going to show how to construct a RKBS $\mathcal{B}$ which is isometrically isomorphic to $l^1(X)$. The above construction of a pair of RKBS is based on providing a function k and a norm $\|\cdot\|_{\mathcal{B}_0}$ with certain properties. Following the paper by Song, Zhang and Hickernell [SZH13] we will present assumptions on k and a way of constructing the norm $\|\cdot\|_{\mathcal{B}_0}$ such that the resulting space $\mathcal{B}$ is of l^1-type.

Let $k\colon X \times X \to \mathbb{F}$ be a bounded function such that $k(x_i, \cdot)$ are linearly independent for all sets of pairwise distinct points $\{x_i \in X : i \in \mathbb{N}_m\}$. Define the norm $\|\cdot\|_{\mathcal{B}_0}$ on $\mathcal{B}_0 = \operatorname{span}\{k(x, \cdot) : x \in X\}$ by

$$\|\sum_{i=1}^{m} c_i k(x_i, \cdot)\|_{\mathcal{B}_0} = \sum_{i=1}^{m} |c_i|.$$

Then $\|\cdot\|_{\mathcal{B}_0}$ satisfies that for any Cauchy sequence (f_n) in $(\mathcal{B}_0, \|\cdot\|_{\mathcal{B}_0})$ such that $f_n(x) \to 0$ we have $\|f_n\|_{\mathcal{B}_0} \to 0$ if and only if, for all pairwise distinct $x_i \in X$

$$\sum_{i=1}^{\infty} c_i k(x_i, x) = 0 \ \forall x \in X \quad \Rightarrow \quad c_i = 0 \ \forall i \in \mathbb{N}$$

We thus obtain a RKBS $\mathcal{B}$ which is isometrically isomorphic to $l^1(X)$ via the map

$$\varphi(c) = \sum_{x \in X} c_x k(x, \cdot) \qquad c \in l^1(X).$$

These constructions are summarised in the following proposition.

Proposition 4.12

Let $k\colon X \times X \to \mathbb{F}$ be a bounded function such that $k(x_i, \cdot)$ are linearly independent for all sets of pairwise distinct points $\{x_i \in X : i \in \mathbb{N}_m\}$. Assume further that, for all pairwise distinct $x_i \in X$

$$\sum_{i=1}^{\infty} c_i k(x_i, x) = 0 \ \forall x \in X \quad \Rightarrow \quad c_i = 0 \ \forall i \in \mathbb{N}.$$

Then for the norm

$$\|\sum_{i=1}^{m} c_i k(x_i, \cdot)\|_{\mathcal{B}_0} = \sum_{i=1}^{m} |c_i|$$

the RKBS $\mathcal{B}$ obtained via the constructions of Proposition 4.11 is isometrically isomorphic to $l^1(X)$.

4.2.1 Examples

In this section we briefly introduce some criteria for kernels which satisfy the assumptions of Proposition 4.12 and give some concrete examples. These examples are again taken from the paper by Song, Zhang and Hickernell [SZH13].

Proposition 4.13

If $k:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{F}$ is of the form

$$k(x,y)=\int\limits_{\mathbb{R}^d} e^{-i\langle x-y,\xi\rangle}\varphi(\xi)\,\mathrm{d}\xi,\qquad x,y\in\mathbb{R}^d.$$

for $\varphi\in L^1(\mathbb{R}^d)$ nonzero almost everywhere on $\mathbb{R}^d$ with respect to the Lebesgue measure then k satisfies the assumptions of Proposition 4.12.

Since this is a Fourier type kernel we get as an immediate corollary that compactly supported functions φ defined on $\mathbb{R}^d$ generate kernels that can lead to l^1-type RKBS.

Corollary 4.14

If $\phi:\mathbb{R}^d\to\mathbb{F}$ is nontrivial, compactly supported and continuous then

$$k(x,y)=\phi(x-y),\qquad x,y\in\mathbb{R}^d.$$

satisfies the assumptions of Proposition 4.12.

Many of the kernels we have seen as generators of RKHS in Section 2.4.2 can in fact also be used to generate l^1-type RKBS. Assuming that the reproducing kernel is defined on a set $X\subseteq\mathbb{R}^d$, examples of functions which can generate l^1-type RKBS by the above results include

- The exponential kernel

$$k(x,y)=\exp(-\|x-y\|_2)=\frac{1}{\pi^d}\int\limits_{\mathbb{R}^d} e^{-i\langle x-y,\xi\rangle}\prod_{i\in\mathbb{N}_d}\frac{1}{1+\xi_i^2}\,\mathrm{d}\xi,\quad x,y\in\mathbb{R}^d;$$

- The Gaussian kernel

$$\begin{aligned} k(x,y) &= \exp\left(-\frac{\|x-y\|_2^2}{\sigma}\right) \\ &= \left(\frac{\sqrt{\sigma}}{2\sqrt{\pi}}\right)^2 \int_{\mathbb{R}^d} e^{-i\langle x-y,\xi\rangle} \exp(-\frac{\sigma}{4}\|\xi\|_2^2)\,\mathrm{d}\xi, \quad x,y \in \mathbb{R}^d; \end{aligned}$$

- B-spline kernels

$$k(x,y) = \prod_{i\in\mathbb{N}_d} B_p(x_i - y_i), \qquad x,y \in \mathbb{R}^d,$$

  where B_p denotes the p-th order B-spline for $p \geq 2$.

- Radial basis functions of compact support

4.2.2 Representer Theorem

Song, Zhang and Hickernell also address the question of representer theorems in [SZH13]. Their approach is different from ours in Chapter 6 in two ways. Firstly they consider the existence of representer theorems a property of the function space rather than the regulariser. Secondly they aim to obtain a representer theorem in the space $\mathcal{B}$ rather than its dual space. This is also in contrast to the results from Section 4.1.3 which showed that for Banach spaces the representer theorem is naturally rooted in the dual space. Nevertheless the results obtained in [SZH13] are interesting and as a corollary contain a representer theorem similar to what we have seen before.
For a set $\{x_1,\ldots,x_m\}$ of finitely many distinct sampling points denote similarly to the RKHS section the Gram matrix by $K_{i,j} = k(x_i,x_j)$. Further denote by $K_x(x)$ the vector $(k(x,x_i))^T$.

Proposition 4.15

Let X be a nonempty set and $(\mathcal{B},\mathcal{B}^\sharp)$ a pair of RKBS on X with reproducing kernel k. Consider the regularisation problem

$$\min\{\mathcal{E}_z(f) + \lambda\Omega(\|f\|_{\mathcal{B}}) : f \in \mathcal{B}\} \tag{35}$$

for $\mathcal{E}_z$ the empirical error for a continuous loss function $\mathcal{C}$ and a continuous, coercive function $\Omega\colon[0,\infty)\to\mathbb{R}$. Then if k is such that for all pairwise distinct $x_1,\ldots,x_{n+1}\in X$

$$\|(K_{i,j})^{-1}K_x(x_{n+1})\|_{l^1(\mathbb{N}_n)}\le 1,$$

there always exists a minimiser $f_0\in\mathcal{B}$ of (35) of the form

$$f_0(x)=\sum_{i=1}^{m}c_ik(x_i,x).$$

We will see in Chapter 6 that our results will also apply to non-reflexive RKBS, giving a unified approach applying to all spaces presented in this chapter and Chapter 2.

5 Duality and Geometry

In this chapter we will introduce a variety of mathematical tools and theory which will be used throughout the proofs of our results in Chapter 6. Many of these results are standard and can be found in various well-known Functional Analysis books. Those results which a reader with a functional analysis background is likely to know are in most cases stated without proofs to fix the notation and to be self-contained for readers from a different background. References where more details and proofs for the results can be found are given. Results with particular relevance to our results will be presented with proof.
There are also some results presented in this chapter which are not very well known. These results will be presented in greater detail and full proofs will be given.

We start the chapter with an introduction into subgradients and directional derivatives in Section 5.1. This theory is standard but is relevant for our results due to its connections to duality mappings which will be made clear in a later section of this chapter. Before introducing duality mappings we present the most important results about annihilators in Section 5.2. Annihilators are crucial throughout our work as they generalise the notion of an orthogonal complement from Hilbert spaces to a general Banach space. Having covered subgradients and annihilators we are in the position to formally introduce duality mappings in Section 5.3. We will point out their connections with subgradients and present a not so well known result, the Beurling-Livingston theorem, linking the duality mapping to an annihilator which is crucial for our results in Chapter 6. After introducing the duality mapping we will discuss some geometrical properties such as smoothness and rotundness in Section 5.4. We will explain how the geometry of a Banach space $\mathcal{B}$ and its dual space $\mathcal{B}^*$ are deeply linked through the duality mapping. This link is Well known but essential to both understanding the geometry of a Banach space and the properties of the duality mapping which are completely determined by the geometry of the space. Finally in Section 5.5 we are going to look into the question whether the distance to a closed subspace of a Banach space is attained for every point. Subspaces for which this is the case are called proximinal. While for reflexive Banach spaces every closed subspace is proximinal, the question whether a subspace is proximinal is very difficult to answer for non-reflexive Banach spaces. As it turns out proximinality is a crucial property for our results for non-reflexive Banach spaces. We thus present some of the known results to characterise proximinal subspaces. Finding general, easily applicable conditions is still an open area of research though.

5.1 Subgradients and Directional Derivatives

The theory presented in this section is standard in Functional Analysis and can be found in various books. The main references we are using for this section are the books by Borwein [BL06, BV10] and Hiriart-Urruty, Lemaréchal [HUL01] and Simons [Sim08]. The books [BL06, HUL01] give a good introduction covering the finite-dimensional case, while the books [BV10, Sim08] also cover the infinte-dimensional case. We will only state the key results we will be using in the discussion of our work for reference and to fix the notation.

Definition 5.1 *(Directional Derivative)*

Let $\mathcal{B}$ be a Banach space and $f : \mathcal{B} \to \mathbb{R}$ a real-valued function on $\mathcal{B}$. The directional derivative of f at $\overline{x} \in \mathcal{B}$ in direction $d \in \mathcal{B}$ is defined as

$$f'(\overline{x}, d) = \lim_{t \searrow 0} \frac{f(\overline{x} + td) - f(\overline{x})}{t}$$

whenever the limit exists.

A priori there is no reason for the directional derivative to exist at a certain point or in a certain direction. For our work we will only be considering the directional derivative of a convex, everywhere continuous function. The next result shows that for such a function the directional derivative exists and is well behaved.

Proposition 5.2

If $f : \mathcal{B} \to (-\infty, \infty]$ is convex then for any $\overline{x} \in \operatorname{core}(\operatorname{dom}(f))$ the directional derivative $f'(\overline{x}, \cdot)$ is everywhere finite and sublinear.

If the directional derivative is linear in d for some $\overline{x}$ then f is Gâteaux differentiable at $\overline{x}$ with derivative $f'(\overline{x}, \cdot)$.

The directional derivative is relevant for our work because it can be used to locally describe the subdifferential which is linked to the duality mappings in such a way that it allows us to construct dual elements with certain desired properties.

Definition 5.3 *(Subdifferential)*

Let $\mathcal{B}$ be a Banach space and $f : \mathcal{B} \to \mathbb{R}$ a real-valued function on $\mathcal{B}$. The subdifferential of f at $\overline{x}$ is the set

$$\partial f(\overline{x}) = \{L \in \mathcal{B}^* \,:\, f(x) - f(\overline{x}) \geq L(x - \overline{x}) \;\forall x \in \mathcal{B}\}.$$

The characterisation of the subdifferential in terms of directional derivatives is summarised by the following Proposition which is a combination of proposition 3.1.6 and theorem 3.1.8 from [BL06].

Proposition 5.4

If $f : \mathcal{B} \to \mathbb{R}$ is convex and $\overline{x} \in \operatorname{dom}(f)$ then for a linear functional $L \in \mathcal{B}^*$ we have $L \in \partial f(\overline{x})$, if and only if $L(\cdot) \leq f'(\overline{x}, \cdot)$.

Moreover for any $\overline{x} \in \operatorname{core}(\operatorname{dom}(f))$ and any $d \in \mathcal{B}$

$$f'(\overline{x}, d) = \max\{L(d) \,:\, L \in \partial f(\overline{x})\}$$

In particular $\partial f(\overline{x})$ is nonempty.

In [BV10] Borwein presents a sandwich theorem (theorem 4.1.18) which essentially allows to squeeze an affine function in between a convex and a concave function. In Chapter 6 we are going to need to construct a subgradient with some control over its norm. It should be possible to show that for the case arising in our proof the affine map obtained from Borwein's sandwich theorem can in fact be chosen as a linear map. This would allow to construct the desired linear functional in the subdifferential of a convex function with a bound on its norm. Unfortunately we have not been able to prove that the affine shift can indeed be chosen to be zero. Thus we will deduce an analogous result from a sandwich theorem presented in the book [Sim08] by Simons, which is a consequence of a stronger version of the Hahn-Banach theorem, the Hahn-Banach-Lagrange theorem.

Theorem 5.5 *(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 that 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 a $u \in \operatorname{dom}(k)$ 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

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

To deduce the required result about subdifferentials with some control over their norm, note that for a convex, everywhere continuous function f Proposition 5.2 shows that the directional derivative is everywhere defined and sublinear so that we can choose $P = f'(\overline{x}, \cdot)$ for some fixed $\overline{x}$ in the sandwich theorem. For simplicity we denote the order relation by $\leq_f$. We let $X = B_V$ be the unit ball in V, and $j(x) = x$ be the canonical embedding of B_V into V. Lastly define k to be identically 0.

With j being the identity map we get

$$j(d) \leq_f \frac{1}{2} j(d_1) + \frac{1}{2} j(d_2) \Leftrightarrow f'(\overline{x}, d - \frac{1}{2} d_1 - \frac{1}{2} d_2) \leq 0.$$

But for any $d_1, d_2 \in B_V$ also $\frac{1}{2}d_1 + \frac{1}{2}d_2 \in B_V$ and $f'(\overline{x}, 0) = 0$ 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 $\overline{x}$ with some control over its behaviour on the unit ball which will allow us to bound its norm.

Corollary 5.6 *(Sandwich theorem for subdifferentials)*

Let V be a nonzero, real vector space, $f : V \to \mathbb{R}$ a convex, everywhere continuous function and $\overline{x} \in V$. Then there exists a linear functional L on V such that $L(\cdot) \le f'(\overline{x}, \cdot)$, i.e. $L \in \partial f(\overline{x})$, and

$$\inf_{d \in B_V} L(d) = \inf_{d \in B_V} f'(\overline{x}, d).$$

5.2 Annihilators

Annihilators provide a natural generalisation of the concept of an orthogonal complement from Hilbert spaces to Banach spaces. Orthogonal complements play a crucial role in many results about learning in Hilbert spaces, in particular representer theorems. It is thus not surprising that annihilators play an important role in the generalisations to Banach spaces. Annihilators are a standard tool in functional analysis so we will just briefly present the properties required in this work. More details can be found e.g. in [Rud91, All11].

Definition 5.7 *(Annihilators, Pre-Annihilators)*

Let $\mathcal{B}$ be a Banach space. The annihilator of a subset $V \subseteq \mathcal{B}$ is defined as the subspace of $\mathcal{B}^*$ of all bounded linear functionals on $\mathcal{B}$ that vanish on V, i.e.

$$V^\perp = \{L \in \mathcal{B}^* \,:\, L(x) = 0 \;\forall x \in V\} \subseteq \mathcal{B}^*.$$

The pre-annihilator of a subset $W \subseteq \mathcal{B}^*$ is defined as the subspace of $\mathcal{B}$ on which every functional of W vanishes, i.e.

$$W_\perp = \{x \in \mathcal{B} \,:\, L(x) = 0 \;\forall L \in W\} \subseteq \mathcal{B}.$$

It is clear that both $V^{\perp}$ and $W_{\perp}$ are closed subspaces. Since $V^{\perp}$ is the intersection of kernels of functionals $\hat{x} \in \hat{\mathcal{B}} \subseteq \mathcal{B}^{**}$ it is also weakly* closed.
Note that the orthogonal complement for continuous semi-inner products defined in Definition 3.8 coincides with the annihilator since

$$W^{\perp} = \{y \in \mathcal{B} : [x,y]_{\mathcal{B}} = 0 \;\forall\, x \in W\} = \{y \in \mathcal{B} : y^{*}(x) = 0 \;\forall\, x \in W\}.$$

There is duality relation between the annihilator, pre-annihilator and the sets generating them which is described in the following result.

Proposition 5.8

Let $\mathcal{B}$ be a Banach space, $V \subseteq \mathcal{B}$ a subset of $\mathcal{B}$ and $W \subseteq \mathcal{B}^{*}$ a subset of $\mathcal{B}^{*}$. Then

(i) $V^{\perp} = (\overline{\operatorname{span}}\{V\})^{\perp}$ and $(V^{\perp})_{\perp}$ is the norm closure of $\operatorname{span}\{V\}$ in $\mathcal{B}$;

(ii) $W_{\perp} = (\operatorname{span} W)_{\perp}$ and $(W_{\perp})^{\perp}$ is the weak* closure of $\operatorname{span}\{W\}$ in $\mathcal{B}^{*}$.

In Chapter 6 we will repeatedly come across annihilators of subspaces of finite codimension. The following result says that the weak* closure from Proposition 5.8 is not required in this case, which is crucial for our results.

Lemma 5.9

Let $\mathcal{B}$ be a Banach space and $W \subset \mathcal{B}^{*}$ a finite subset of $\mathcal{B}^{*}$. Then $(W_{\perp})^{\perp}$ is weak* closed and thus $(W_{\perp})^{\perp} = \operatorname{span}\{W\}$.

The proof follows immediately from the following lemma (Lemma 3.2 in [Bre11]).

Lemma 5.10

Let V be a vector space and $L, L_1, \ldots, L_m$ linear functionals on V such that

$$\bigcap_{i \in \mathbb{N}_m} \ker(L_i) \subseteq \ker(L).$$

Then $L \in \operatorname{span}\{L_i : i \in \mathbb{N}_m\}$.

Proof:
Define a map

$$q : V \to \mathbb{F}^m,$$
$$q(x) = (L_1(x), \dots, L_m(x)).$$

Then $\ker(q) = \bigcap_{i \in \mathbb{N}_m} \ker(L_i) \subseteq \ker(L)$ and thus L factors through q. More precisely $\overline{L}(q(x)) = L(x)$ defines a linear map $\overline{L} : \mathrm{im}(q) \to \mathbb{F}$ on the image of q. Extending $\overline{L}$ to $\mathbb{F}^m$ we obtain a linear map $\overline{L}$ such that the diagram

$$\begin{tikzcd}
V \arrow[rr, "L"] \arrow[dr, "q"'] & & \mathbb{F} \\
& \mathbb{F}^m \arrow[ur, "\overline{L}"'] &
\end{tikzcd}$$

commutes. Then there exist $\lambda_1, \dots, \lambda_m \in \mathbb{F}$ such that $\overline{L}(y_1, \dots, y_m) = \sum\limits_{i=1}^{m} \lambda_i y_i$ and so

$$L(x) = \overline{L}(q(x)) = \sum_{i=1}^{m} \lambda_i L_i(x)$$

for all $x \in V$. Thus $L = \sum\limits_{i=1}^{m} \lambda_i L_i$.

❏

Proof *(Of Lemma 5.9)*:
The fact that $\mathrm{span}\{W\} \subseteq (W_\perp)^\perp$ is obvious since by definition all $L \in W$ are zero on $W_\perp$.
For the converse assume $L \in (W_\perp)^\perp$. Then $\bigcap_{\widetilde{L} \in W} \ker(\widetilde{L}) = W_\perp \subseteq \ker(L)$ and since W is a finite set by Lemma 5.10 $L \in \mathrm{span}\{W\}$.

❏

Lastly there is a useful duality between annihilators and dual spaces of quotient spaces. Similarly dual spaces of closed subspaces can be expressed in terms of the quotient of the dual space by the annihilator.

Proposition 5.11

Let $\mathcal{B}$ be a Banach space and V a closed subspace and $q : \mathcal{B} \to \frac{\mathcal{B}}{V}$ the quotient map.

(i) The map

$$\Phi : \left(\frac{\mathcal{B}}{V}\right)^* \to V^\perp,$$
$$\Phi(L_q) = L_q \circ q$$

is an isometric isomorphism of $\left(\frac{\mathcal{B}}{V}\right)^*$ onto $V^\perp$.

(ii) Extending $L \in V^*$ by the Hahn-Banach theorem to a functional $\overline{L} \in \mathcal{B}^*$ the map

$$\Phi : V^* \to \frac{\mathcal{B}^*}{V^\perp},$$
$$\Phi(L) = \overline{L} + V^\perp$$

is an isometric isomorphism of V^* onto $\frac{\mathcal{B}^*}{V^\perp}$.

5.3 Duality Mappings

When discussing RKBS in Chapter 4 we already saw that the representer theorem in a Banach space is a result about the dual space rather than the space itself. This does not become apparent in the classical version of the representer theorem for Hilbert spaces as the duality mapping is the identity. With the representer theorem being set in the dual space it is clear that the properties of the duality mapping crucially determine the representer theorem one can obtain. With the representer theorem for uniform RKBS in Section 4.1.3 we already saw the shift from a linear to a nonlinear duality mapping which exposed the property of the representer theorem being about the dual space. In our work in Chapter 6 we will step by step further deal with the duality mapping not being univocal, injective and surjective.

In this section we give the fundamental definitions of the duality mapping which are well known. We will then close the section by presenting a less well known but very powerful result, the Beurling-Livingston theorem, which is essential for some of our results in Chapter 6.

Definition 5.12 *(Duality mappings)*

Let $\mu : [0,\infty) \to [0,\infty)$ be a continuous and strictly increasing function such that $\mu(0) = 0$ and $\mu(t) \underset{t\to\infty}{\longrightarrow} \infty$.
A set-valued map $J_\mu : V \to 2^{V^*}$ is called a duality mapping of V into V^* with gauge function μ if $J_\mu(0) = \{0\}$ and for $0 \neq x \in V$

$$J_\mu(x) = \{L \in V^* \,:\, L(x) = \|L\| \cdot \|x\|, \|L\| = \mu(\|x\|)\}\,.$$

The following properties of the duality mapping are well known but essential, see e.g. [Bro69].

Proposition 5.13

For every $x \in V$ the set $J_\mu(x)$ is nonempty, weakly* closed and convex.

A reason to introduce the subdifferential in Section 5.1 was its link with the duality mapping. Some important properties of duality mappings can be deduced from the fact that a mapping is a duality mapping if and only if it is the subgradient of a certain convex function, as shown e.g. in [Asp67].

Proposition 5.14

For a normed linear space V with duality mapping J_μ with gauge function μ define $M : V \to \mathbb{R}$ by

$$M(x) = \int\limits_0^{\|x\|_V} \mu(t)\,\mathrm{d}t. \tag{36}$$

For any $0 \neq x \in V$ we have that $\partial M(x) = J_\mu(x)$. Thus $L \in J_\mu(x)$ if and only if

$$M(y) \geq M(x) + L(y-x) \qquad \forall y \in V.$$

In this work we will be considering the case where μ is the identity and the duality mapping an isometry. This case is commonly referred to as the *normalised duality mapping*. We will omit the subscript and simply write J for the normalised duality mapping.

Finally the following generalised version of the Beurling-Livingston theorem (c.f. [BL62, Bro65b]) is essential for the proof of one of our main result. A proof of this theorem can be found in the work by Browder [Bro65a] which is very general, deducing the result from a result on multi-valued monotone nonlinear mappings. A more direct proof, giving a better idea of the objects occurring in the result, can be found in the work by Blažek [Bla82]. Unfortunately there is an issue in the proof in the paper by Blažek, we thus present a corrected version of it here. The overall intuition of Blažeks proof is correct nonetheless and a summary of it can also be found in a paper by Asplund [Asp67]. For convenience we also include Asplund's summary here.

Theorem 5.15 *(Beurling-Livingston)*

Let V be a real normed linear space with duality mapping J_μ with gauge function μ and W a reflexive subspace of V.

Then for any fixed $x_0 \in V, L_0 \in V^*$ there exists $z \in W$ such that

$$J_\mu(x_0 + z) \cap (W^\perp - L_0) \neq \varnothing.$$

Proof *(Sketch [Asp67])*:
Consider the functional

$$F : V \to \mathbb{R},$$
$$F(x) = M(x - x_0) - L_0(x - x_0).$$

Since F is continuous, convex and coercive, it attains its minimum on the reflexive space W at some point $z \in W$. By the Hahn-Banach theorem F thus has a subgradient at z which is identically zero on W. By Proposition 5.14 this subgradient is a dual element with the stated property.

❑

Proof *(Corrected version of [Bla82])*:
Using the functional M from Proposition 5.14 define a functional $F : V \to \mathbb{R}$ by

$$F(x) = M(x - x_0) - L_0(x - x_0).$$

Since M is continuous, convex with strictly increasing derivative and L_0 is linear, F is clearly continuous, convex and coercive. This means that F attains its minimum on the reflexive subspace W in at least one point, $\overline{z}$ say. Hence, for all $y \in W$

$$\begin{aligned} F(y) - F(\overline{z}) &\geq 0 \\ \Leftrightarrow M(y - x_0) &\geq M(\overline{z} - x_0) + L_0(y - \overline{z}) \\ \Leftrightarrow M(y - x_0) - M(\overline{z} - x_0) + L_0(\overline{z} - x_0) &\geq L_0(y - x_0). \end{aligned} \tag{37}$$

By Proposition 5.14 this means that $L_0\big|_W \in \partial M\big|_W(\overline{z} - x_0) = J_\mu\big|_W(\overline{z} - x_0)$. For simplicity we write $L_0\big|_W = L_W$.

Note that if $x_0 \in W$ and $L_W = 0$ we have that $F(x) = M(x - x_0)$ on W so $\overline{z} = x_0$ and we trivially have $J_\mu(x_0 - x_0) = \{0\} = \{-L_0 + L_0\} \subset W^\perp + L_0$. So we can without loss of generality assume that not both $x_0 \in W$ and $L_W = 0$.

In case $x_0 \in W$ it is clear that M is minimised at x_0. If $L_W \neq 0$ then L_W attains its norm on W in a point z say. Thus it is clear that there exists a minimiser for F of the form $\overline{z} = z + x_0$. More precisely F is minimised where an element of ∂M and ∇L_0 are equal. Since $\partial M(x - x_0) = J_\mu(x - x_0)$ and elements $L_x \in J_\mu(x - x_0)$ are of norm $\|L_x\| = \mu(\|x - x_0\|)$, the fact that ∂M and ∇L_0 are equal implies that the minimiser $\overline{z} = z + x_0$ is such that $\|L_W\|_{W^*} = \mu(\|\overline{z} - x_0\|)$.

If on the other hand $x_0 \notin W$ then we note that $\overline{z}$ being the minimum for F on W implies that $L_z^F(y) \geq 0$ for all $L_z^F \in \partial F(\overline{z})$ and all $y \in W$. But by definition $\partial M(x - x_0) - \partial L_0(x - x_0) \subseteq \partial F(x)$ and thus for every $L_z \in J_\mu(\overline{z} - x_0) = \partial M(\overline{z} - x_0)$ and $L_W = L_0 = \partial L_0(\overline{z} - x_0)$

$$L_z(y) - L_W(y) \geq 0$$

But since L_z is of norm $\mu(\|\overline{z} - x_0\|)$ this means that

$$\mu(\|\overline{z} - x_0\|) \cdot \|y\| \geq L_z(y) \geq L_W(y)$$

for all $y \in W$. Thus $\|L_W\|_{W^*} = \|L_0\big|_W\|_{W^*} \le \mu(\|\overline{z} - x_0\|)$.
Now denote by $\overline{W}$ the space generated by W and x_0 and note that this space is still reflexive. Extend L_W to $L_{\overline{W}}$ on $\overline{W}$ by setting

$$L_{\overline{W}}(x_0) = L_0(\overline{z}) - \mu(\|\overline{z} - x_0\|) \cdot \|\overline{z} - x_0\|.$$

Then

$$\begin{aligned} L_{\overline{W}}(\overline{z} - x_0) &= L_W(\overline{z}) - (L_0(\overline{z}) - \mu(\|\overline{z} - x_0\|) \cdot \|\overline{z} - x_0\|) \\ &= \mu(\|\overline{z} - x_0\|) \cdot \|\overline{z} - x_0\| \end{aligned}$$

so $\|L_{\overline{W}}\|_{\overline{W}^*} \ge \mu(\|\overline{z} - x_0\|)$.
Further $L_{\overline{W}}(y) = L_W(y) \le \mu(\|\overline{z} - x_0\|) \cdot \|y\|$ for all $y \in W$, so $\|L_{\overline{W}}\| > \mu(\|\overline{z} - x_0\|)$ can only happen if the norm is attained for some point $\lambda y + \nu x_0$ for $y \in W$, $\nu \ne 0$. Or equivalently, dividing through by ν, at a point $y + x_0$ for some $y \in W$. But for those points we have

$$\begin{aligned} L_{\overline{W}}(y + x_0) &= L_W(y) + L_0(\overline{z}) - \mu(\|\overline{z} - x_0\|) \cdot \|\overline{z} - x_0\| \\ &\le \mu(\|\overline{z} - x_0\|) \cdot \|y + \overline{z}\| - \mu(\|\overline{z} - x_0\|) \cdot \|\overline{z} - x_0\| \\ &\le \mu(\|\overline{z} - x_0\|) \,|\|y + \overline{z}\| - \|\overline{z} - x_0\|| \\ &\le \mu(\|\overline{z} - x_0\|) \cdot \|y + x_0\| \end{aligned}$$

and thus $\|L_{\overline{W}}\| = \mu(\|\overline{z} - x_0\|)$ and $L_{\overline{W}}(\overline{z} - x_0) = \|L_{\overline{W}}\| \cdot \|\overline{z} - x_0\|$.

Since for $x_0 \in W$ we have $\overline{W} = W$ in either case we have obtained a function $L_{\overline{W}}$ such that

$$L_{\overline{W}}\big|_W = L_0\big|_W \qquad \|L_{\overline{W}}\| = \mu(\|\overline{z} - x_0\|) \qquad L_{\overline{W}}(\overline{z} - x_0) = \|L_{\overline{W}}\| \cdot \|\overline{z} - x_0\|$$

Now extend $L_{\overline{W}}$ by the Hahn-Banach theorem to L_V on V such that

$$\|L_V\| = \|L_{\overline{W}}\| = \mu(\|\overline{z} - x_0\|)$$

and $L_V\big|_{\overline{W}} = L_{\overline{W}}$. Hence $(L_V - L_0)\big|_W = 0$ so $L_V \in W^\perp + L_0$.
It remains to show that $L_V \in J_\mu(\overline{z} - x_0)$ by showing Eq. (37) holds for L_V and every $y \in V$. Notice first that

$$L_V(y - x_0) \le \|L_V\| \cdot \|y - x_0\| = \|L_{\overline{W}}\| \cdot \|y - x_0\| = L_{\overline{W}}\left(\frac{\|y - x_0\|}{\|\overline{z} - x_0\|}(\overline{z} - x_0)\right). \quad (38)$$

But

$$L_{\overline{W}}\left(\frac{\|y-x_0\|}{\|\overline{z}-x_0\|}(\overline{z}-x_0)\right)-L_{\overline{W}}(\overline{z}-x_0)=\left(\frac{\|y-x_0\|}{\|\overline{z}-x_0\|}-1\right)\mu(\|\overline{z}-x_0\|)\|\overline{z}-x_0\|$$
$$=\mu(\|\overline{z}-x_0\|)\left(\|y-x_0\|-\|\overline{z}-x_0\|\right) \quad (39)$$

and further

$$M(y-x_0)-M(\overline{z}-x_0)=\int\limits_{\|\overline{z}-x_0\|}^{\|y-x_0\|}\mu(t)\,\mathrm{d}t\geq\left(\|y-x_0\|-\|\overline{z}-x_0\|\right)\mu(\|\overline{z}-x_0\|), \quad (40)$$

so the left-hand side of Eq. (40) is always at least as big as the left-hand side of Eq. (39). We can thus add the left-hand side of Eq. (39) to the right-hand side of Eq. (37) and the left-hand side of Eq. (40) to the left-hand side of Eq. (37) while preserving the inequality. Equation (37) is in particular true for $\overline{z}$ and in that case also for $L_{\overline{W}}$ as it agrees with L_0 on $\overline{z}$ and x_0, i.e.

$$M(\overline{z}-x_0)-M(\overline{z}-x_0)+L_{\overline{W}}(\overline{z}-x_0)\geq L_{\overline{W}}(\overline{z}-x_0)$$

Thus by adding the left-hand sides of Eq. (39) and Eq. (40) as described we obtain

$$M(y-x_0)-M(\overline{z}-x_0)+L_{\overline{W}}(\overline{z}-x_0)\geq L_{\overline{W}}\left(\frac{\|y-x_0\|}{\|\overline{z}-x_0\|}(\overline{z}-x_0)\right)$$

for all $y\in V$. But since L_V also agrees with $L_{\overline{W}}$ on $\overline{z}$ and x_0 this together with Eq. (38) implies that

$$M(y-x_0)-M(\overline{z}-x_0)+L_V(\overline{z}-x_0)\geq L_V(y-x_0)$$

for all $y\in V$, which is what we wanted to prove. Thus indeed $L_V\in J_\mu(\overline{z}-x_0)$ as claimed. By homogeneity of J_μ clearly $-L_V$ with $-\overline{z}\in W$ is as in the statement of the theorem.

❏

5.4 Geometry

The properties of the duality mapping are deeply linked to the geometry of the space. Consequently also the geometries of the space itself and its dual space are strongly linked. These connections are very well known but since these geometrical properties are determining the properties of the duality mapping, which in turn determines the form of the representer theorem one may be able to obtain, we summarise the most important definitions and results here. It will turn out that in particular a lack of strict convexity will cause difficulties in making precise statements about representer theorems. This is because in a space which is not strictly convex the unit ball contains straight sections. We will see that it is very difficult to make any statements about the behaviour of a regulariser across a straight section of the unit ball. We thus close the section by giving an overview of exposed points, i.e. points which are not contained in the interior of any straight section, and exposed faces. The main references used for the results in this section are the books [HUL01] by Hiriart-Urruty and Lemaréchal and [Meg12] by Megginson. Another good reference is [Köt83].

Definition 5.16

A point x on the sphere $S_r \subset V$ is

(i) rotund if $\|\frac{x+y}{2}\|_V < r$ for all $y \in S_r \setminus \{x\}$;

(ii) smooth if there exists exactly one $L \in V^*$ such that $\|L\| = 1$ and $L(x) = \|x\|_V$.

We are only going to present the most important results characterising rotundity and smoothness and illustrating their link with properties of the dual space via the duality mapping. These results are well known and can be found in many books about Functional Analysis. For a very thorough discussion about rotundity and smoothness the reader is referred to Megginson [Meg12].

Proposition 5.17

Let $L \in V^*$ and $x \in V$ such that $L \in J(x)$.

(i) If L is rotund then x is smooth.

(ii) If L is smooth then x is rotund.

While in general we can only make statements about the structure of a point based on the geometry of its dual element(s), for a reflexive Banach space the statement goes both ways via the identification of the space with its second dual.

Proposition 5.18

A reflexive Banach space is rotund if and only if its dual space is smooth and is smooth if and only if its dual space is rotund.

The above definitions have only been qualitative, defining when a point is smooth or rotund. We can go further and in fact give a measure of "how smooth" or "how rotund" a point is.

Definition 5.19

The modulus of smoothness of the space V is defined as

$$\rho_V : (0,\infty) \to [0,\infty)$$
$$\rho_V(t) = \sup\left\{ \frac{\|x+y\|_V + \|x-y\|_V}{2} - 1 \,:\, \|x\|_V = 1, \|y\|_V = t \right\}$$

The space V is uniformly smooth if

$$\lim_{t \searrow 0} \frac{\rho_V(t)}{t} = 0$$

The modulus of rotundity or modulus of convexity of the space V is defined as

$$\delta_V : [0,2] \to [0,1]$$
$$\delta_V(\varepsilon) = \inf\left\{ 1 - \frac{\|x+y\|_V}{2} \,:\, \|x\|_V = \|y\|_V = 1, \|x-y\|_V \geq \varepsilon \right\}$$

The space V is uniformly convex if $\delta_V(\varepsilon) > 0$ for all $\varepsilon \in (0,2]$.

Using the modulus of smoothness and modulus of rotundity one can also obtain a quantitative version of Proposition 5.18, linking the modulus of smoothness of the space with the modulus of rotundity of the dual space and vice versa. This is a standard result but not required for our results. Details can be found e.g. in [LT79, Meg12].
The above definition of uniform smoothness is in fact equivalent to a condition on the differentiability of the norm, as seen before in the context of uniformly smooth semi-inner product spaces.

Remark 5.20

There are the following equivalent, useful definitions of uniform convexity and uniform smoothness.

- The space V is uniformly smooth if the norm $\|\cdot\|_V$ is uniformly Fréchet differentiable, i.e. if the limit
$$\lim_{t\to 0}\frac{\|x+ty\|_V-\|x\|_V}{t}$$
exists uniformly for $t\in\mathbb{R}$ and all $x,y\in V$ such that $\|x\|_V=\|y\|_V=1$.

- The space V is uniformly convex if for every $\varepsilon>0$ there exists $\delta>0$ such that for all $x,y\in S_V$ with $\|x-y\|_V>\varepsilon$ we have $\frac{\|x+y\|_V}{2}\leq 1-\delta$.

The connections between the geometry of the space V and its dual space V^* presented in this section lead to a number of statements about the properties of the normalised duality mapping depending on the geometric structure of V. These properties are well known and can be found e.g. in Dragomir [Dra04] and the references therein.

Proposition 5.21

We have the following equivalences between properties of the duality mapping J and the geometry of the space V.

(i) J is surjective if and only if V is reflexive.

(ii) J is injective if and only if V is strictly convex.

(iii) J is univocal if and only if V is smooth.

(iv) J is norm-to-weak* continuous exactly at points of smoothness of V.

(v) J is norm-to-norm continuous if and only if V is uniformly smooth.

In view of (i) we say a normed space V is *subreflexive* if the image of J is norm-dense in V^*. Bishop and Phelps prove in [BP61] that every Banach space is subreflexive. They also state that some but not all incomplete normed spaces are subreflexive.

We now turn to a brief discussion of exposed faces which are essential when discussing representer theorems in spaces which are not strictly convex. This is in particular relevant for the case of non-reflexive spaces where we are interested in l^1-type spaces. For more details see e.g. [HUL01, AGP05].

Definition 5.22 *(Exposed Face)*

A non-empty subset F of the ball $B_r \subset \mathcal{B}$ is an exposed face of B_r if there exists an $L \in \mathcal{B}^*$ such that

$$F = \left\{ x \in B_r \,:\, L(x) = \sup_{y \in B_r} L(y) \right\}.$$

A point $x \in \mathcal{B}$ is an exposed point if $\{x\}$ is an exposed face.

It is easy to see that every exposed face is a face, the converse is not true though. It is easy to find an example of an extreme point which is not exposed. But in a sense there are not very many such points, since every extreme point is the limit of exposed points, i.e.

$$\exp(B_r) \subseteq \operatorname{ext}(B_r) \subseteq \overline{\exp(B_r)}$$

Unfortunately exposed faces and points do not share some of the nice properties of faces and extreme points. While the non-empty intersection of faces is always a face, Aizpuru and Garcia-Pacheco show in [AGP08] the existence of a non-exposed point which is the intersection of exposed faces. The same results further show that a point can be exposed in every 2-dimensional subspace containing it, but not an exposed point. This is in contrast to being an extreme point or being smooth or rotund. Those are 2-dimensional properties, meaning that a point is an extreme point or smooth or rotund if and only if it is an extreme point or smooth or rotund in every 2-dimensional subspace containing it.

For a general Banach space $\mathcal{B}$ it is very hard to make any general statements about which points are exposed. It is clear that every rotund point is exposed. Furthermore every smooth exposed point is rotund. Aizpuru and Garcia-Pacheco in [AGP08] try to make a statement about the properties of exposed points which are not smooth and might not be rotund. Since in general there can be points which are not smooth and not exposed they define the following two stronger conditions of non-smoothness.

Definition 5.23

A point x on the sphere $S_r \subset \mathcal{B}$ is

(i) strongly non-smooth if for every $y \in S_r$ such that $\|tx+(1-t)y\|_{\mathcal{B}} = r$ for all $t \in [0,1]$, x is not a smooth point in $\operatorname{span}\{x,y\}$;

(ii) uniformly non-smooth if for every $y \in S_r \setminus \{x\}$, x is not a smooth point in $\operatorname{span}\{x,y\}$.

They prove that in a separable Banach space every strongly non-smooth point is in fact an exposed point and thus the set of exposed points is exactly the union of the set of rotund points and the set of strongly non-smooth points. But in a non-separable Banach space this is false and there may be strongly non-smooth points which are not exposed. They then show that the notion of uniformly non-smooth points is too strong, in the sense that there can be either uniformly non-smooth points which are not exposed, or exposed points which are neither rotund nor uniformly non-smooth. For non-separable Banach spaces we thus do not know how to completely describe the set of exposed points.

In [Day55] Day discusses convexity and smoothness properties of common function spaces. He remarks that there is no example known of a space which is smooth but not strictly convex. But in view of Minkowski spaces, where we construct a norm by starting with a convex set which is symmetric around the origin and construct a norm such that this set is the unit ball for this norm, we see that there is an easy way of constructing counterexamples for most assumptions of the interplay of smoothness and rotundness one may hope to make. This in particular means that as long as one can imagine a symmetric convex set with a point which violates the assumption, one can construct a Minkowski norm which has this set as unit ball. For a detailed discussion of this construction see e.g. [Tho96]. We thus cannot make assumption about the smoothness or rotundness of a point based on the properties of points in its neighbourhood.

5.5 Proximinal Subspaces

We close this chapter by exploring when for a subspace W of a vector space V the distance of a point $x \in V \setminus W$ to the subspace is attained. Subspaces for which the distance is always attained are called proximinal. It is clear that a proximinal subspace necessarily has to be closed.

While for reflexive Banach spaces the distance of a point to the subspace is attained for any closed subspace, the question turns out to be quite difficult in general for non-reflexive Banach spaces. We will show that for finite dimensional subspaces and hyperplanes the question of proximinality can still be answered in a satisfactory way. Some recent results show that there is little hope to be able to extend these results further by giving examples of Banach spaces which in a sense have very few other proximinal subspaces.

We present some further characterisations of proximinal subspaces which hold in general normed vector spaces, these conditions may not always be easy to check though. The question of general and easily applicable conditions for proximinality still remains an open area of research.

Definition 5.24 *(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

$$\|x - y\|_V = \operatorname{dist}(x, W).$$

As already mentioned in the introduction of this section, for reflexive Banach spaces the question of proximinality can be answered positively (see e.g. [Con94]).

Proposition 5.25

If $\mathcal{B}$ is a reflexive Banach space then any closed linear subspace $W \subset \mathcal{B}$ is proximinal.

If the space is not reflexive the question which subspaces are proximinal becomes much more difficult. In the book [Hol75] Holmes presents some conditions which characterise proximinal subspaces. In particular he gives a condition for subspaces of finite codimension which is the case we are going to encounter in Chapter 6. In view of the fact that the norm on the quotient space represents the distance of a point to the subspace it is not surprising that a first characterisation of proximinality is based on properties of the quotient map.

Theorem 5.26 *(Godini's theorem)*

Let V be a real normed vector space and $W \subset V$ a closed subspace of V. Denote by B_V and $B_{\frac{V}{W}}$ the unit balls of V and $\frac{V}{W}$ respectively. Then W is proximinal if and only if

(i) $q(B_V)$, the image of the unit ball of V under the quotient map q, is the unit ball of $\frac{V}{W}$, i.e. $q(B_V) = B_{\frac{V}{W}}$.

(ii) $q(B_V)$, the image of the unit ball of V under the quotient map q, is closed in the quotient space $\frac{V}{W}$.

Proof:

Part 1: *(W proximinal ⇒ (i))*
Assume W is proximinal. It is clear that $q(B_V) \subseteq B_{\frac{V}{W}}$.
To prove the reverse inclusion $B_{\frac{V}{W}} \subseteq q(B_V)$ fix $x \in V$ such that

$$\|x+W\|_{\frac{V}{W}} = 1 = \operatorname{dist}(x,W)$$

Let $y \in W$ be such that the distance of x to W is attained at y, i.e

$$\|x-y\|_V = \inf\{\|x-z\|_V : z \in W\}.$$

In particular $\|x-y\|_V \leq 1$ and $q(x-y) = x+W$ so that $x \in q(B_V)$.

Part 2: *((i) ⇒ W proximinal)*
Let $x \in V \setminus W$ such that $\|x+W\|_{\frac{V}{W}} = \operatorname{dist}(x,W) = 1$. If $q(B_V) = q(B_{\frac{V}{W}})$ then there exists $y \in V$ such that $\|y\|_V = \|q(y)\|_{\frac{V}{W}}$ and $q(y) = x+W$. But then $x-y \in W$ and

$$\|x-(x-y)\|_V = \|y\|_V = \|q(y)\|_{\frac{V}{W}} = \operatorname{dist}(x,W)$$

so that the distance from x to W is attained in $x-y$.

Part 3: *(Condition (i) ⇔ condition (ii))*
It is clear that (i) implies (ii) so we only need to prove the converse.
If $q(B_V)$ is closed and a proper subset of $B_{\frac{V}{W}}$ then there exists a point

$$x+W \in B_{\frac{V}{W}} \setminus q(B_V)$$

which by the Hahn-Banach separation theorem can be strictly separated from $q(B_V)$. By Proposition 5.11 the separating functional on $\frac{V}{W}$ corresponds to a functional $L \in W^\perp$ such that

$$L(x) > \sup\{L(y) : \|y\|_V \leq 1\} = \|L\|_{V^*}.$$

But $|L(x)| \leq \|L\|\|x+W\|_{\frac{V}{W}} = \|L\|\operatorname{dist}(x,W) = \|L\|$ which is a contradiction so $q(B_V) = B_{\frac{V}{W}}$.

❑

Using Godini's theorem one can obtain a further characterisation, in particular of subspaces of finite codimension, which is the case we are going to encounter in Chapter 6. The subspace which will appear in the proofs of our results will be generated by a finite number of linear functionals, the linear functionals defining our optimisation problem. The following result gives a condition on precisely those linear functionals to characterise proximinality, making it very appealing for our work.

Corollary 5.27

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 $L_1, \dots, L_m$ of $W^\perp$ define a map S by

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

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 $\mathbb{R}^m$.

Proof:

Part 1: *(Every finite-dimensional subspace is proximinal)*
We want to use Theorem 5.26 (ii) and show that $q(B_V)$ is closed in $\frac{V}{W}$.
Let $(x_n) \subset B_V$ such that $x_n + W \to x + W$ in $\frac{V}{W}$. Then

$$\|(x_n - x) + W\|_{\frac{V}{W}} = \operatorname{dist}(x_n - x, W) \to 0$$

Letting $y_n \in W$ be vectors which at least almost attain the distance of $x_n - x$ to W we obtain that $x_n - x - y_n = e_n$ for some small vector e_n, in fact $\|e_n\| \to 0$. Since the y_n are almost attaining the distance from $x_n - x$ to W they must be contained in a bounded set so the sequence is bounded in a finite-dimensional subspace so there exists a convergent subsequence, also denoted y_n, with limit point $y \in W$. But then

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} x + y_n + e_n = x + y$$

and $x + y \in B_V$ as the limit of $(x_n) \subset B_V$. Thus $q(x+y) = x + W \in q(B_V)$ and $q(B_V)$ is closed in $\frac{V}{W}$.

Part 2: *(W s.t.* $\operatorname{codim}(W) < \infty$ *is proximinal* $\Leftrightarrow$ *$S(B_V)$ is closed in $\mathbb{R}^m$)* Given a basis $L_1, \dots, L_m$ for $W^\perp$, choose $x_1, \dots, x_m \in V$ such that $L_i(x_j) = \delta_{ij}$. Then $x_1 + W, \dots, x_m + W$ are a basis for $\frac{V}{W}$. By the choice of x_i we have that

$$L_j\left(\sum_{i=1}^m L_i(x)x_i\right) = L_j(x)$$

and thus $x - \sum_{i=1}^m L_i(x)x_i \in (W^\perp)_\perp = W$. This means that

$$x + W = \sum_{i=1}^m L_i(x)(x_i + W)$$

Let

$$T : \mathbb{R}^m \to \frac{V}{W}$$
$$T(c_1, \dots, c_m) = \sum_{i=1}^m c_i(x_i + W)$$

be the natural isomorphism from $\mathbb{R}^m$ to $\frac{V}{W}$. Then $q = T \circ S$ and $S(B_V)$ is closed in $\mathbb{R}^m$ if and only if $q(B_V)$ is closed in $\frac{V}{W}$. Hence the result follows by Theorem 5.26 (ii).

❑

Another case that is relevant for the proofs of our results is proximinality of hyperplanes. Conway in [Con94] proves that the kernel of a linear functional is proximinal exactly for those linear functionals which attain their norm.

Lemma 5.28

Let $\mathcal{B}$ be a Banach space and $L \in \mathcal{B}^*$. Then $\ker(L)$ is proximinal if and only if there exists an $x \in \mathcal{B}$ such that $L(x) = \|L\|_{\mathcal{B}^*} \cdot \|x\|_{\mathcal{B}}$.

Proof:
Throughout the proof for simplicity denote $\ker(L) = W$.

Part 1: *(*$\ker(L)$ *proximinal* $\Rightarrow$ *L attains its norm)*
Suppose W is proximinal. Define $L_q : \frac{\mathcal{B}}{W} \to \mathbb{F}$ by $L_q(x+W) = L(x)$. Then $L_q \in \left(\frac{\mathcal{B}}{W}\right)^*$ and $\|L_q\| = \|L\|$. Since $\dim(\frac{\mathcal{B}}{W}) = 1$ the functional L_q attains its norm, i.e.

$$L_q(x+W) = \|L_q\| \cdot \|x+W\|_{\frac{\mathcal{B}}{W}} \text{ for some } x+W \in \frac{\mathcal{B}}{W}.$$

Because W is proximinal the distance of $x \in \mathcal{B}$ to W is attained, i.e. there exists $y \in W$ such that $\|x+y\|_{\mathcal{B}} = \operatorname{dist}(x,W) = \|x+W\|_{\frac{\mathcal{B}}{W}}$. But then

$$L(x+y) = L(x) = L_q(x+W) = \|L_q\| \cdot \|x+W\|_{\frac{\mathcal{B}}{W}} = \|L\|_{\mathcal{B}^*} \cdot \|x+y\|_{\mathcal{B}}.$$

Part 2: *(L attains its norm* $\Rightarrow$ $\ker(L)$ *is proximinal)*
Assume L attains its norm in $x_0 \in \mathcal{B}$, i.e. $L(x_0) = \|L\|_{\mathcal{B}^*} \cdot \|x_0\|_{\mathcal{B}}$. Since $L \in W^\perp$ by Proposition 5.11 L defines a linear functional on $\frac{\mathcal{B}}{W}$. We have

$$\|x_0+W\|_{\frac{\mathcal{B}}{W}} = \sup_{\overline{L} \in \left(\frac{\mathcal{B}}{W}\right)^*, \|\overline{L}\| \le 1} |\overline{L}(x_0+W)| = \|x_0\|_{\mathcal{B}}.$$

Let $x \in \mathcal{B}$ be such that $x+W \neq 0$ in $\frac{\mathcal{B}}{W}$. Since $\dim(\frac{\mathcal{B}}{W}) = 1$ there exists $\lambda \in \mathbb{F}$ such that $|\lambda| = \frac{\|x+W\|_{\frac{\mathcal{B}}{W}}}{\|x_0+W\|_{\frac{\mathcal{B}}{W}}}$ and $x+W = \lambda(x_0+W)$. Hence $x - \lambda x_0 \in W$. But then

$$\|x-(x-\lambda x_0)\|_{\mathcal{B}} = |\lambda| \cdot \|x_0\|_{\mathcal{B}} = |\lambda| \cdot \|x_0+W\|_{\frac{\mathcal{B}}{W}} = \|x+W\|_{\frac{\mathcal{B}}{W}} = \operatorname{dist}(x,W),$$

so the distance from x to W is attained at $x - \lambda x_0$.

❏

Aside from being needed for our work in Chapter 6 this result says that every space contains proximinal hyperplanes. In fact, since any Banach space $\mathcal{B}$ is subreflexive, i.e. the norm attaining functionals are dense in $\mathcal{B}^*$, the set of proximinal hyperplanes is in a sense dense in the set of closed hyperplanes.

It turns out that the question which subspaces, other than finite-dimensional ones and hyperplanes, are proximinal is much harder to answer in general. While Godini's theorem and its corollary above give a characterisation, their conditions may not be easy to check. Further there exist spaces which have in a sense very few proximinal subsets. Read in [Rea18] and Kadets, Lopez, Martin and Werner in [KLMW18] prove that there exists a Banach space which does not contain any proximinal subspace of finite codimension greater than one. This can be seen as a kind of converse to the above - while the set of proximinal hyperplanes is always dense there may not be any proximinal space of larger finite codimension.
The space which will appear in the proofs of our results, which determines the quality of the representer theorem one can obtain, will in practise always be of finite codimension greater than one, so this example is also a negative example of a space in which we can not hope to obtain a strong representer theorem.

We are going to state a few other characterisations of proximinality which may be useful to determine whether the subspace defined by our optimisation problem in Chapter 6 is proximinal. These condition are taken from the book [Sin70] by Singer where one can find a more detailed discussion and further results.
The next result is interesting for the subspaces arising in our application as for subspaces of finite codimension the quotient space is finite-dimensional and hence reflexive.

Proposition 5.29

Assume V is a real normed vector space and $W \subset V$ a closed subspace of V. If $\frac{V}{W}$ is reflexive then W is proximinal if and only if and for every $x^{**} \in (W^{\perp})^{*} \subset V^{**}$ there exists an $x \in V$ such that

- $\|x\|_V = \|x^{**}\|_{(W^{\perp})^{*}}$,
- $x^{**}(L) = L(x)$ for all $L \in W^{\perp}$.

Moreover when trying to determine whether a subspace is proximinal it is sufficient to look at properties of the unit ball of the subspace within the ambient space.

Proposition 5.30

Let V be a real normed vector space with unit ball B_V and $W \subset V$ a closed subspace of V with unit ball B_W. The subspace W is proximinal if its

(i) unit ball B_W is sequentially compact in the weak topology on V;

(ii) unit ball B_W is proximinal in V.

It turns out that condition (ii) is only a sufficient condition, not a necessary one. Saidi proves in [Sai05] that there exists a proximinal subspace of a Banach space such that its unit ball is not proximinal in the ambient space.

We close the section by briefly addressing the question when every closed subspace of finite codimension of a Banach space is proximinal. This is interesting because in this case our results in Chapter 6 will show that we can always obtain a strong representer theorem. It turns out that these are exactly the reflexive Banach spaces. This result is given in [Sin70] together with more equivalent conditions and other classes of subspaces.

Proposition 5.31

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.

6 Existence of Representer Theorems

In this chapter we present most of the main results of this work. We've presented several versions of the classical representer theorem in the previous chapters. These theorems apply to different optimisation problems in different function spaces. The feature all those statements have in common though is that they all give sufficient conditions for the existence of a solution in a subspace spanned by the data. The aim of this chapter is to answer the question of necessary conditions. This question has been partially answered by Argyriou, Micchelli and Pontil in 2009 [AMP09]. They proof necessary and sufficient conditions for the existence of a solution in the linear span of the data for regularisation and the regularised interpolation problems posed in a Hilbert space. They also give a geometrical interpretation of their result for differentiable regularisers. This provides a good intuition about the result, making it more practically useful.

Throughout the chapter we will follow the same approach Argyriou, Micchelli and Pontil used in their work [AMP09]. While the regularisation problem is more common in applications, the regularised interpolation problem is more convenient to work with. One can show that under mild conditions a representer theorem holds for the regularisation problem if and only if it holds for the regularised interpolation problem with the same regulariser. We can thus restrict our attention to the regularised interpolation problem. We will first present our results for regularised interpolation and subsequently, at the end of this chapter, present the proof for equivalence with the regularisation problem.

Furthermore we will not be concerned about the existence of minimisers in this chapter. There is a rich literature about existence of minimisers for both the regularisation problems and regularised interpolation problem, see e.g. [MP04, ZZ12]. Instead we will assume that for any given data $\{(x_i, y_i) : n \in \mathbb{N}_m\} \subset X \times Y$ the minimum of the regularised interpolation problem (see Eq. (41) below) is attained whenever the interpolation constraints can be satisfied.

We will start the chapter by giving a brief overview of the work by Argyriou, Micchelli and Pontil from 2009 [AMP09] who were, to our knowledge, the first to address the question of necessary conditions for a representer theorem. They proved necessary and sufficient conditions for a representer theorem to exist for Tikhonov regularisation in Hilbert spaces and then gave a geometric interpretation of those conditions under the additional assumption of the regulariser being differentiable.

After the brief overview of their work we will present our work, starting with removing the differentiability assumption on the regulariser for the geometric interpretation in Section 6.2. Following this we will extend the result to the corresponding regularised interpolation problem in uniformly smooth and uniformly convex Banach spaces. The proofs follow the same concepts as in the paper [AMP09] and Section 6.2. The lack of Hilbert space structure requires some extra work, in particular we are using the theory of semi-inner products introduced in Section 3. This extends the applicability of the results to uniform reproducing kernel Banach spaces as introduced in Section 4.1.1. These results have been made available as journal publication in [Sch19b].

Subsequently we will further generalise the results to weaken the assumptions on the function space from being uniformly smooth and uniformly convex to merely reflexive. The proofs again follow the same concepts but use some new machinery, most notably the Beurling-Livingston theorem presented in Section 5. With this extension the statements now apply to all reflexive RKBS in the sense of the definitions presented in Section 4.1. The results presented here have been made available as journal publication in [Sch21].

All forms of representer theorems we have seen above were posed in spaces which are at least reflexive. At this point of our presentation it will become clear that for a representer theorem of the classical form to hold reflexivity is crucial. We will illustrate this with a counterexample in Section 6.5, showing that a representer theorem of the form as presented in the previous sections can not hold in general in a non-reflexive space. This is unfortunate since in applications l^1 regularisation is very common.

We thus propose the concept of *approximate solutions* and thus also an *approximate representer theorem* to extend our results to include non-reflexive

Banach spaces as well. These concepts are motivated by the intuition gained from the previously mentioned counterexample given in Section 6.5. After defining these concepts we show that they indeed allow to extend our results from the earlier sections to non-reflexive Banach spaces and thus apply to all RKBS presented throughout Chapter 4, including non-reflexive RKBS defined in Section 4.2, in particular l^1-type RKBS. These results will appear in conference proceedings in [Sch20].
Throughout the chapter all Hilbert and Banach spaces will be assumed to be real. We will comment on the complex case in Section 7.2.

6.1 Differentiable Regularisers

We start our presentation with the results by Argyriou, Micchelli and Pontil from 2009 [AMP09] which provided the starting point for our work. The classical representer theorem gives a sufficient condition on the regularisation functional Ω for a representer theorem to hold. It is a natural question to ask whether one can easily characterise all regularisers which give rise to a representer theorem, i.e. prove a necessary condition. Argyriou, Micchelli and Pontil did answer this question for Hilbert spaces and provided a geometric intuition of their result for differentiable regularisers Ω. They state that it should be possible to extend this intuition to the non-differentiable case, but leave this for future work. Our work presented in the subsequent sections covers this extension to non-differentiable regularisers and further generalisations of their results.

In this and the next section we will be concerned with the existence of representer theorems for regularised interpolation problems in real Hilbert spaces. Throughout both sections $\mathcal{H}$ will always denote a Hilbert space. We are interested in problems of the form

$$\min\left\{\Omega(f) \,:\, f \in \mathcal{H}, \langle f, x_i\rangle_{\mathcal{H}} = y_i, \forall i \in \mathbb{N}_m\right\}. \tag{41}$$

Note that while in applications we will often be interested in interpolation constraints of the form $f(x_i) = y_i$ for some nonlinear functions f, this case is also included in the above setting. Since $\mathcal{H} = \mathcal{H}^*$ we can represent function evaluation at the data points by inner products and via RKHS as presented in Section 2.4 we can introduce nonlinearities.

Our goal is to classify all regularisers which give rise to a linear representer theorem. We will call such regularisers admissible, as made precise in the following definition.

Definition 6.1 *(Admissible Regularizer)*

We say that a function $\Omega : \mathcal{H} \to \mathbb{R}$ is admissible if for any $m \in \mathbb{N}$ and any given data $\{x_1, \dots, x_m\} \subset \mathcal{H}$ and $\{y_1, \dots, y_m\} \subset Y$ such that the interpolation constraints can be satisfied the regularised interpolation problem Eq. (41) admits a solution f_0 such that there exist coefficients $\{c_1, \dots, c_m\} \subset \mathbb{R}$ such that

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

Now we are in the position to give the first result from [AMP09] which states that Ω being admissible is equivalent to it being non-decreasing along orthogonal directions.

Lemma 6.2

A function $\Omega : \mathcal{H} \to \mathbb{R}$ is admissible if and only if for every $f, f_\perp \in \mathcal{H}$ such that $\langle f, f_\perp \rangle_{\mathcal{H}} = 0$ we have

$$\Omega(f + f_\perp) \geq \Omega(f). \tag{42}$$

Proof:

Part 1: *(Ω admissible $\Rightarrow$ non-decreasing along orthogonal directions)*
Fix any $f \in \mathcal{H}$ and consider the regularised interpolation problem

$$\min\{\Omega(g) \, : \, g \in \mathcal{H}, \langle g, f \rangle_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}\}.$$

As Ω is assumed to be admissible there exists a solution in $\operatorname{span}\{f\}$ which clearly is f itself. But if $f_\perp$ is such that $\langle f, f_\perp\rangle_{\mathcal{H}} = 0$ then $\langle f + f_\perp, f\rangle_{\mathcal{H}} = \langle f, f\rangle_{\mathcal{H}}$ so $f + f_\perp$ also satisfies the constraints and hence necessarily $\Omega(f + f_\perp) \geq \Omega(f)$ as claimed.

Part 2: *(Non-decreasing along orthogonal directions ⇒ Ω admissible)*
Conversely fix any data $(x_i, y_i) \in \mathcal{H} \times Y$ for $i \in \mathbb{N}_m$ such that the constraints in Eq. (41) can be satisfied. Let f be a solution to the regularised interpolation problem and decompose it as $f = f_0 + f_\perp$ where $f_0 \in \operatorname{span}\{x_i : i \in \mathbb{N}_m\}$ and $f_\perp \in \operatorname{span}\{x_i : i \in \mathbb{N}_m\}^\perp$. Then by assumption

$$\Omega(f) = \Omega(f_0 + f_\perp) \geq \Omega(f_0)$$

and clearly $y_i = \langle f_0 + f_\perp, x_i\rangle_{\mathcal{H}} = \langle f_0, x_i\rangle_{\mathcal{H}}$ so f_0 is a solution to the problem.

❑

Using this result it is shown in [AMP09] that for differentiable functions admissibility is actually equivalent to being a radially symmetric, nondecreasing function.

Theorem 6.3

Assume $\dim(\mathcal{H}) \geq 2$. A differentiable function $\Omega : \mathcal{H} \to \mathbb{R}$ is admissible if and only if it is of the form

$$\Omega(f) = h\left(\langle f, f\rangle_{\mathcal{H}}\right) \tag{43}$$

for some nondecreasing function $h : [0, \infty) \to \mathbb{R}$.

Proof *(Sketch)*:
It is immediately clear that every Ω of the form (43) is admissible, so we only need to prove the converse, that every differentiable, admissible Ω is of this form. The proof uses that the property of being nondecreasing along orthogonal directions means that

$$\langle \nabla\Omega(f), f_\perp\rangle_{\mathcal{H}} = \lim_{t\to 0} \frac{\Omega(f + tf_\perp) - \Omega(f)}{t} = 0 \tag{44}$$

since the numerator is positive for every $t \in \mathbb{R}$.
For $\mathcal{H} = \mathbb{R}^d$ we write for fixed f_0 of unit norm $f = \|f\| U f_0$ where $U \in SO(n)$ is a rotation. Writing $U = e^D$ for some skew-symmetric matrix D we can consider the path

$$z(\lambda) = \|f\| e^{\lambda D} f_0$$

and using Eq. (44) we find that Ω is constant along this path.
For general $\mathcal{H}$ we use the path

$$z(\lambda) = \frac{(1-\lambda) f_0 + \lambda f}{\|(1-\lambda) f_0 + \lambda f\|_{\mathcal{H}}} \|f\|_{\mathcal{H}}.$$

This makes clear that we are essentially arguing that being nondecreasing in orthogonal directions means that tangential derivatives are zero.

❑

This is a satisfying result, classifying differentiable, admissible regularisers for regularised interpolation problems (Eq. (41)) entirely. But in view of Lemma 6.2 it is not difficult to see that there are non-differentiable, admissible regularisers, e.g. $\Omega(f) = \lceil \|f\|_{\mathcal{H}} \rceil$. The question of removing the differentiability assumption is thus a very natural one to ask. It is mentioned in [AMP09] but left for future work.

6.2 Non-differentiable Regularisers

In this section we will show how to remove the assumption of differentiability of the regulariser in Theorem 6.3. The use of the difference quotient in the proof of this result indicates that one needs a new idea to do this. Our approach is split into two parts. First we prove that in fact the bound along tangents from Lemma 6.2 can be extended to apply to a significantly larger region of the space. Subsequently we show with a mollification argument that this wider bound is sufficient to give a clear description of the regulariser Ω, showing that in fact it has to be almost radially symmetric in a sense which will be made precise in the statement below.

To extend the tangential bound from Lemma 6.2 to hold for a larger region of the space we first notice that the proof of Theorem 6.3 essentially is based on showing that the tangential derivative of Ω is zero. While at that point it is only a language nuance whether we are speaking of Ω as being nondecreasing along orthogonal directions or along tangential directions, it is exactly this which gives the intuition for extending the bound to a significantly larger region. In the later sections of this chapter, when we are considering more general Banach spaces, the two terms will also cease to be equivalent and the results will generalise to tangential directions.

The way to extend the bound obtained in Lemma 6.2 is to chain the tangential bound repeatedly. This way we can in fact reach every point in the space which lies outside the ball where we started.

Lemma 6.4

If for all $f, f_T \in \mathcal{H}$ such that $\langle f_T, f\rangle_{\mathcal{H}} = 0$ we have $\Omega(f) \leq \Omega(f + f_T)$ then for any fixed $\hat{f}$ we have that

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

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

Proof:

Part 1: *(Bound Ω on the half space given by the tangent plane through $\hat{f}$)*
We start by showing that Ω is radially non-decreasing. Since it is nondecreasing along tangential directions this immediately gives the claimed bound for the entire half space given by the tangent plane 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. 4. Fix some $\hat{f} \in \mathcal{B}$ and $1 < \lambda \in \mathbb{R}$ and set $f = \lambda \cdot \hat{f}$. We want to show that $\Omega(f) \geq \Omega(\hat{f})$. Let $f_T \in \mathcal{H}$ be such that $\langle f_T, \hat{f}\rangle_{\mathcal{H}} = 0$. Now for $t \in \mathbb{R}$ let

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

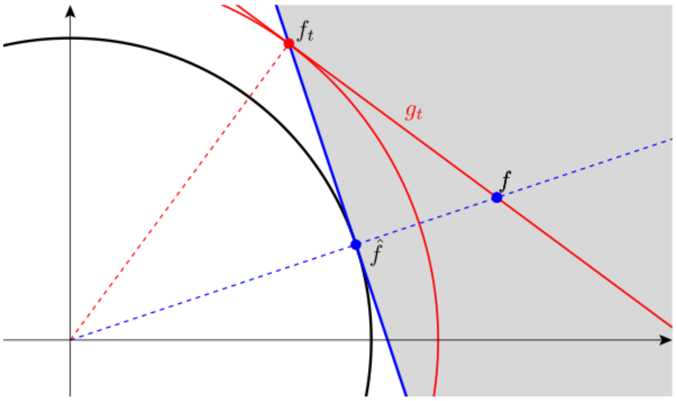


Figure 4: We can extend the tangential bound to the ray $\lambda \cdot \hat{f}$ by finding the point f_t along the tangent from where the tangent at f_t hits the desired point f on the ray. Via the tangents to points along the ray the bound then extends to the shaded half space.

so that $f_t + g_t = f$. We want to show that for some t the segment g_t is tangent at f_t to apply Lemma 6.2. We have

$$\langle g_t, f_t \rangle_{\mathcal{H}} = (\lambda - 1) \left\langle \hat{f}, \hat{f} \right\rangle_{\mathcal{H}} + (\lambda - 1)t \left\langle \hat{f}, f_T \right\rangle_{\mathcal{H}} - t \left\langle f_T, \hat{f} \right\rangle_{\mathcal{H}} - t^2 \left\langle f_T, f_T \right\rangle_{\mathcal{H}}$$

so that, since $\left\langle \hat{f}, f_T \right\rangle_{\mathcal{H}} = 0$ we get

$$0 = \langle g_t, f_t \rangle_{\mathcal{H}} = (\lambda - 1) \|\hat{f}\|^2 - t^2 \|f_T\|^2$$
$$\Leftrightarrow t = \sqrt{\lambda - 1} \frac{\|\hat{f}\|}{\|f_T\|}$$

This means for $t = \sqrt{\lambda - 1} \frac{\|\hat{f}\|}{\|f_T\|}$ the segment g_t is tangent at f_t and thus by Lemma 6.2

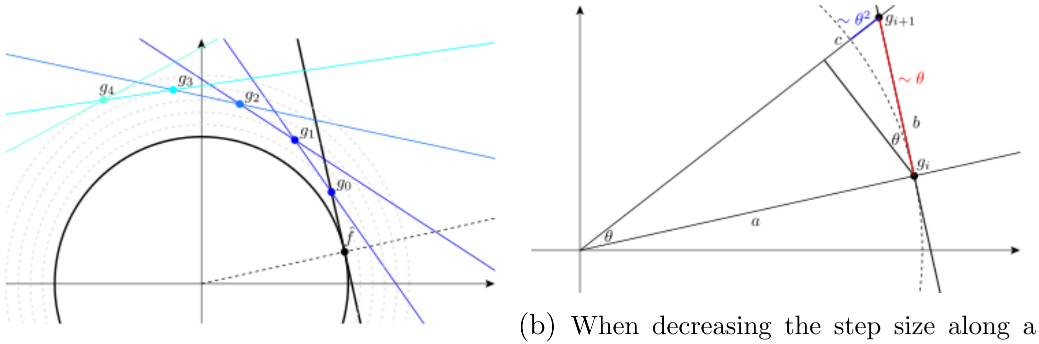
$$\Omega(\hat{f}) \leq \Omega(f_t) \leq \Omega(f_t + g_t) = \Omega(f)$$

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 plane through $\hat{f}$, i.e. the shaded region in Fig. 4.

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

a tangent into the half space just considered, we follow a tangent in the opposite direction around the circle, as shown in Fig. 5a. 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 with each step we move considerably further around the circle than away from it, as shown in Fig. 5b.



(a) By repeatedly taking steps along tangents we can move all the way around the circle.

(b) 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.

Figure 5: Extending the bound around the circle

This is clear from Fig. 5b by noting that for small angles θ

$$\frac{b}{a} = \tan(\theta) \approx \theta$$

and

$$\frac{c}{b} = \sin(\theta) \approx \theta$$

Since $a = \|\hat{f}\|$ is constant this means that the step along the tangent is of order θ for a small step size while the step away from the circle in this case is of order θ^2. So the step size away from the circle decreases much faster than around it as claimed. This proves that indeed by making θ small enough we can reach any point arbitrarily close to the ball of radius $\|\hat{f}\|$.

Combining both arguments 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.

❑

Remark 6.5

Note that Lemma 6.2 in particular implies that $\Omega(0) \leq \Omega(f)$ so Ω has a global minimum at 0 and we can without loss of generality assume $\Omega(0) = 0$.

Using Lemma 6.4 we can show that in fact an admissible regulariser has to be almost radially symmetric in the sense made precise in the following theorem.

Theorem 6.6

A function $\Omega : \mathcal{H} \to \mathbb{R}$ is admissible if and only if it is of the form

$$\Omega(f) = h(\langle f, f \rangle_{\mathcal{H}})$$

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

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.

Proof:

Part 1: *(Ω continuous in radial direction implies Ω radially symmetric)*
We begin by showing 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 that Ω is admissible and continuous in radial direction 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 of norm r such that, without loss of generality, $\Omega(f) > \Omega(g)$.
But then by Lemma 6.4 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 $f_0 \in \mathcal{H}$ of unit norm, define the mollified regulariser by

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

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

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

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

$$\begin{aligned}\widetilde{\Omega}(f + f_T) = \int\limits_{-1}^{0} \rho(t)\Omega\left((\|f + f_T\| - t)\frac{f + f_T}{\|f + f_T\|}\right) \mathrm{d}t \\ \geq \int\limits_{-1}^{0} \rho(t)\Omega\left((\|f\| - t)\,\frac{f}{\|f\|}\right) \mathrm{d}t = \widetilde{\Omega}(f). \quad (45)\end{aligned}$$

Note that by Lemma 6.4 we have that

$$\Omega\left((\|f+f_T\|-t)\frac{f+f_T}{\|f+f_T\|}\right) \geq \Omega\left((\|f\|-t)\frac{f}{\|f\|}\right)$$

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 $\langle f_T, f\rangle_{\mathcal{H}} = 0$. This proves that the integral estimate Eq. (45) 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 parts 1 and 2 the radial mollification $\widetilde{\Omega}$ of a given regulariser Ω is of the form

$$\widetilde{\Omega}(f) = h\left(\langle f, f\rangle_{\mathcal{H}}\right)$$

for some continuous, non-decreasing h. But if we consider Ω along any two distinct, fixed directions given by $f_1, f_2 \in \mathcal{H}$, $f_1 \neq f_2$, $\|f_1\| = \|f_2\| = 1$ as $\Omega(t \cdot f_i) = h_{f_i}\left(\langle t\cdot f_i, t\cdot f_i\rangle_{\mathcal{H}}\right)$ then the mollifications of 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 the 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.

The converse, that any function Ω which is almost radially symmetric in the sense of the theorem is tangentially nondecreasing, is obvious.

❑

This result provides a complete answer to the question of which regularisers are admissible for regularised interpolation on Hilbert spaces. As discussed in the introduction of Chapter 4 there are various reasons to consider learning in Banach spaces rather than just Hilbert spaces. This is going to be the aim of the following sections.

6.3 Uniform Banach Spaces

In this section we are going to show how the ideas of Lemma 6.2, Lemma 6.4 and Theorem 6.6 will generalise to uniform Banach spaces using the theory of semi-inner products introduced in Section 3. Recall that we defined a uniform Banach space to be a Banach space which is uniformly convex and uniformly smooth. This means the space can be seen as *almost Hilbert* with the semi-inner product having many of the properties of an inner product, including a Riesz representation theorem. This in particular extends the results to apply to uniform s.i.p. RKBS as introduced in Section 4.1.1. As an example of uniform Banach spaces the reader can think of l^p spaces for any $p \in (1, \infty)$.
More precisely in this section we are considering the following generalisation of the regularised interpolation problem Eq. (41)

$$\min \{\Omega(f) \,:\, f \in \mathcal{B}, [f, x_i]_{\mathcal{B}} = y_i \;\forall i \in \mathbb{N}_m\} \tag{46}$$

where the domain $\mathcal{B}$ of the interpolation problem is a real uniform Banach space. Throughout the section $\mathcal{B}$ will always denote a uniform Banach space. In the previous Hilbert space 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. But our work, similarly as the work by Micchelli and Pontil [MP04], 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. We have seen this before in Theorem 4.7, the representer theorem for uniform RKBS. The following definition is the analogue of the previous definition of admissibility.

Definition 6.7 *(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 $\{x_1, \dots, x_m\} \subset \mathcal{B}$ and $\{y_1, \dots, y_m\} \subset Y$ such that the interpolation constraints can be satisfied the regularised interpolation problem Eq. (46) admits a solution f_0 such that there exist coefficients $\{c_1, \dots, c_m\} \subset \mathbb{R}$ such that the dual element of f_0 is of the form

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

With this definition at hand it is now again 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. It has been shown by a Hahn-Banach argument similar as e.g. in Zhang, Zhang [ZZ12], which we presented above in the proof of Theorem 4.7, that the same is true for uniform Banach spaces. It turns out that also our results for Hilbert spaces generalise exactly to uniform Banach spaces.

Theorem 6.8

A function $\Omega : \mathcal{B} \to \mathbb{R}$ is admissible if and only if it is of the form

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

for some non-decreasing $h : [0, \infty) \to \mathbb{R}$ 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 $\Omega(f)$ is only constrained by the monotonicity property, i.e. it has to lie 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.

The proof of this result follows a similar line of argumentation as seen in Sections 6.1 and 6.2. The first result from [AMP09] presented as Lemma 6.2 above, which says that admissible regularisers are nondecreasing along orthogonal directions, generalises to an analogous statement for semi-inner products. With the lack of symmetry of the semi-inner product the order of the arguments is crucial and we will see that the result can be summarised as saying that an admissible regulariser is nondecreasing along tangents.
But before we state and prove this result we are going to show that we can extend this tangential bound in the same way as before and a function that is non-decreasing in tangential directions is in fact non-decreasing in norm.

Lemma 6.9

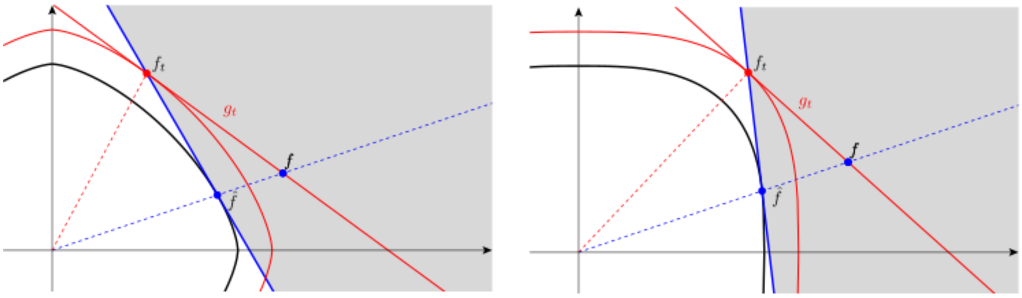
If for all $f, f_T \in \mathcal{B}$ such that $[f_T, f]_{\mathcal{B}} = 0$ we have $\Omega(f) \le \Omega(f + f_T)$ then for any fixed $\hat{f}$ we have that

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

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

Proof:

Part 1: *(Bound Ω on the half space given by the tangent plane through $\hat{f}$)*
The proof strategy is the same as in Lemma 6.4. We first show that Ω is radially nondecreasing by moving out along a tangent and back along another tangent to hit any point along the ray $\lambda \cdot \hat{f}$ for $\lambda > 1$. Via the tangents at those points this again immediately gives the bound for the entire half space spanned by the tangent plane through $\hat{f}$. That this intuition from Hilbert spaces remains true is illustrated in Figs. 6a and 6b
As before we fix some $\hat{f} \in \mathcal{B}$ and $1 < \lambda \in \mathbb{R}$ and set $f = \lambda \cdot \hat{f}$. We want to show that $\Omega(f) \ge \Omega(\hat{f})$. Let $f_T \in \mathcal{B}$ be such that $[f_T, \hat{f}]_{\mathcal{B}} = 0$ or equivalently



(a) Extending the tangential bound to the shaded half space in $l^{\frac{3}{2}}$

(b) Extending the tangential bound to the shaded half space in l^3

Figure 6: One can move "out and back" along tangents to extend the bound to the half space given by the tangent plane through $\hat{f}$ in l^p-spaces

$\|\hat{f} + t \cdot f_T\| > \|\hat{f}\|$ for all $t \neq 0$. Now for $t \in \mathbb{R}$ 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

$$\|f_t + s \cdot g_t\| > \|f_t\| \text{ for all } s \in (0,1) \tag{47}$$

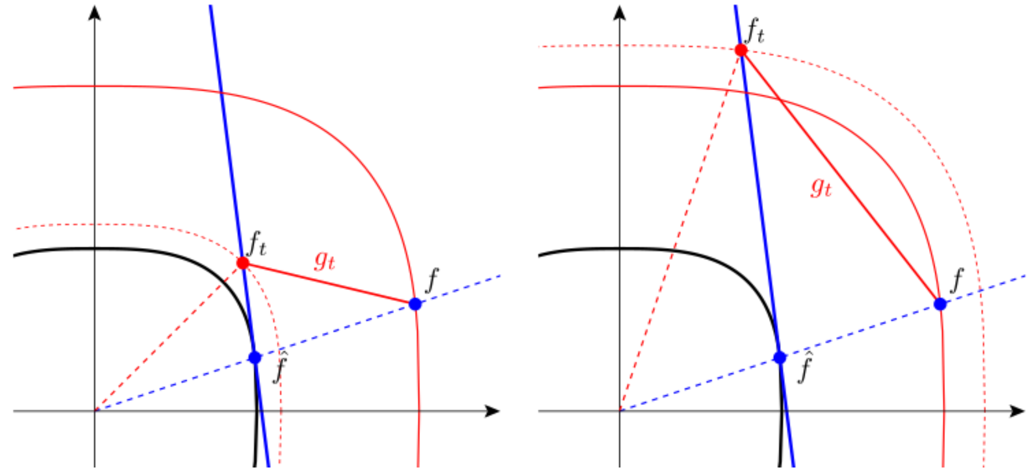
as illustrated in Fig. 7a. On the other hand, as illustrated in Fig. 7b, for t large enough so that $\|f_t\| > \|f\|$ we thus must have

$$\|f_t + s \cdot g_t\| < \|f_t\| \text{ for } s \text{ small enough.} \tag{48}$$

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 duality mapping is norm-to-norm continuous by Proposition 5.21 $\frac{f_t^*(g_t)}{\|f_t\|}$ is clearly continuous in t. By the above discussion the expression is



(a) For small t $\|f_t + s \cdot g_t\|$ must be increasing. (b) For large t $\|f_t + s \cdot g_t\|$ must be decreasing.

Figure 7: The norm derivative changes sign along the tangent $t \cdot f_T$ so there has to be a point where it is zero, i.e. a tangent.

positive for small t and negative for large t so by the intermediate value theorem there exists t_0 such that

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

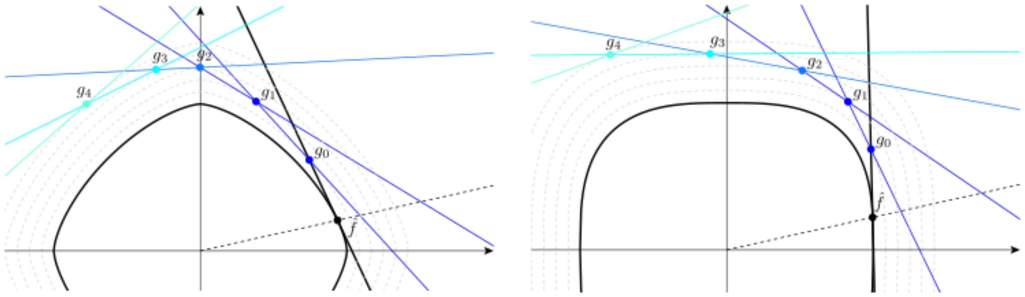
so that indeed $[g_{t_0}, f_{t_0}]_{\mathcal{B}} = 0$ and thus g_{t_0} is tangential to f_{t_0}. But this means that $\Omega(f) \geq \Omega(f_{t_0}) \geq \Omega(\hat{f})$ as claimed.

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 Figs. 6a and 6b.

Part 2: *(Extend the bound around the circle)*

Secondly we show that we can again extend the bound further to apply all the way around the circle, namely $\Omega(f) \geq \Omega(\hat{f})$ for all f such that $\|f\| > \|\hat{f}\|$, in the same way as before in Lemma 6.4.

More precisely we consider $f_t = \hat{f} + t \cdot f_T$ 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. We fix another point along that tangent and repeat the process, moving around the circle. This is illustrated for the uniform Banach space case in Figs. 8a and 8b. We claim that just as in the Hilbert space case we can this way move around the circle while staying arbitrarily close to it by making the step size along each tangent small enough. 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. 9.



(a) Extend the bound around the circle in $l^{\frac{3}{2}}$ (b) Extend the bound around the circle in l^3

Figure 8: By repeatedly taking steps along tangents we can move all the way around the circle.

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 $\mathcal{B}$ as defined in Definition 5.19. 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 $\|f - f_T\| > \|f\|$. 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. 9.

Combining both arguments this proves that we can reach any point with

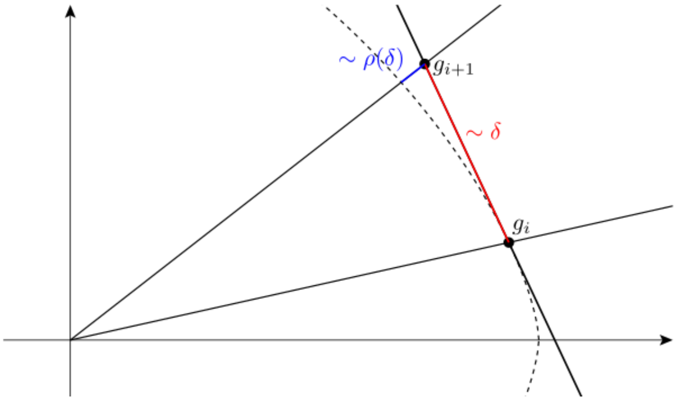


Figure 9: For a step along the tangent of order δ the step away from the circle is of order $\rho(\delta)$.

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. This is the analogous result of Argyriou, Micchelli and Pontils result from [AMP09] that any admissible regulariser on a Hilbert space is non-decreasing in orthogonal direction (see Lemma 6.2 above). With orthogonality not being symmetric for semi-inner products and in view of the intuition gained from the equivalence with James orthogonality we see that in fact the result can be summarised as admissible regularisers being nondecreasing along tangents.

Lemma 6.10

A function $\Omega : \mathcal{B} \to \mathbb{R}$ is admissible if and only if for every $f, f_T \in \mathcal{B}$ such that $[f_T, f]_{\mathcal{B}} = 0$ we have

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

Proof:

The fact that admissibility implies being tangentially nondecreasing follows in exactly the same way as for Lemma 6.2. We are going to repeat the short argument here for convenience. It is the reverse direction, the existence of a linear functional of the claimed form for any given data, which is more involved.

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 $\operatorname{span}\{f^*\}$ which by injectivity and homogeneity of the duality mapping 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.

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 $f_0^* \in \operatorname{span}\{x_i^*\}$ we are done so assume it is not. We let

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

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 we have $Z^* \cap X^* = \{0\}$ and further, since by assumption $f_0^* \notin X^*$, also $\operatorname{span}\{f_0^*\} \cap X^* = \{0\}$.

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 we have $f_0 \notin Z$ and thus $\operatorname{codim}(\operatorname{span}\{f_0, Z\}) = m - 1$. But since $X^* = \operatorname{span}\{x_i^*\}$ and the duality mapping is a homeomorphism, X is homeomorphic to a linear space of dimension m. This means that that $X \cap \operatorname{span}\{f_0, Z\}$ is homeomorphic to a one-dimensional space and hence in particular contains a nonzero element.
Now fix such $0 \neq f \in X \cap \operatorname{span}\{f_0, Z\}$. As we noted earlier f being nonzero means that $f \notin \operatorname{span}\{f_0\}$ and $f \notin Z$. Thus $f = \lambda f_0 + \mu g$ for $\lambda, \mu \neq 0, g \in Z$. By homogeneity of the duality mapping $\lambda \cdot X = X$ and so

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

and thus

$$\frac{1}{\lambda} f = f_0 + \frac{\mu}{\lambda} g = f_0 + \widetilde{g} \in X \cap \operatorname{span}\{f_0, Z\} \tag{49}$$

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 dual elements of the data points and $f_T \in Z$. By definition of Z that means 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 $[x^*, f_T^*]_{\mathcal{B}^*} = 0$ for all $x^* \in X^*$ and hence we see that for $\overline{f_0} = f_0 + f_T \in X$ we get that

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

By the equivalence of s.i.p. orthogonality 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 6.9 we know that a function which is non-decreasing along tangential directions is non-decreasing in norm, so $\|\overline{f_0}\| < \|f_0\|$ implies that $\Omega(\overline{f_0}) \leq \Omega(f_0)$. We thus have found a solution with dual element in the span of the dual elements of the data points as claimed.

❑

One can now put those two results together in the same way as in the proof of Theorem 6.6 to proof Theorem 6.8, that all admissible regularisers on a uniform Banach space have to be almost radially symmetric. Since the proofs are identical it will be omitted here.
By drawing some pictures of l_n^1 one can get the intuition that the extension of the tangential bound as in Lemma 6.9 should also be possible in non-uniform spaces. This raises the question of how the result of admissible regularisers being non-decreasing along tangents extends to non-uniform spaces. Since, as stated at the end of Section 3.2, every linear functional is represented by a semi-inner product if and only if the space is reflexive, a natural assumption to replace uniformity of the space is reflexivity.

6.4 Reflexive Banach Spaces

In this section we use the intuition gained from the semi-inner product theory in Section 6.3 to generalise our results further to weaken the assumptions of uniform convexity and uniform smoothness to reflexivity. Thus all reflexive RKBS as presented in Section 4.1 are examples of spaces covered by our results. A particular example of a reflexive Banach space which is not uniform is l_n^1. One could still work with semi-inner products but we found it more convenient to work with the duality mapping directly. The necessary theory was presented in Chapter 5. The most notable difference to the previous sections is that in a Banach space which is not strictly convex or smooth the duality mapping is not injective or univocal respectively. This requires some further adjustments to our previous definitions.

Throughout this section we let $\mathcal{B}$ be a reflexive Banach space with duality mapping

$$J(x) = \{L \in \mathcal{B}^* : L(x) = \|L\| \cdot \|x\|, \|L\| = \|x\|\}.$$

With the lack of a one to one identification of points and dual elements we now consider evaluations of linear functionals instead of data points. Of course these could still be point evaluations but may also be other linear maps such as e.g. local averages of the form

$$L(f) = \int_{\mathcal{B}} f(x)\,\mathrm{d}P(x).$$

where P is a probability measure on $\mathcal{B}$.

We will consider the following further generalisation of the regularised interpolation problems Eqs. (41) and (46).

$$\min\left\{\Omega(f) \,:\, f \in \mathcal{B},\, L_i(f) = y_i \;\forall i \in \mathbb{N}_m\right\}. \tag{50}$$

With the weaker properties of the duality mapping we also need to define a new notion of admissibility of Ω for this problem, corresponding to what we have seen in previous sections.

Definition 6.11 *(Admissible Regularizer)*

We say that 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. (50) admits a solution f_0 such that there exist coefficients $\{c_1, \ldots, c_m\} \subset \mathbb{R}$ such that

$$\hat{L} = \sum_{i=1}^{m} c_i L_i \in J(f_0).$$

It turns out that in this setting it is in general not possible anymore to give as concise a description of the regulariser as we did previously in Theorems 6.3, 6.6 and 6.8. We are going to see that in particular the lack of strict convexity means that the exact form of the regulariser is harder to describe. In an arbitrary reflexive Banach space there can be a lot of variation in the geometry of the space in terms of rotundness, smoothness and exposed points which makes it very difficult to make generally applicable statements.

We will thus start by discussing to what extent the tangential bound can be generalised, as this result can still be stated in a concise, general form. We will subsequently look into the geometric interpretation. Here we do not see how to make a statement for any reflexive Banach space due to the geometric variety. We will thus impose another assumption on the geometry of

the space. As it is a lack of strict convexity which causes most difficulties in making a general statement we are, for now, only going to discuss the case of strictly convex spaces. As spaces which are not strictly convex appear in applications, this excludes some cases of interest, in particular l^1. But since only the finite dimensional l^1_n is reflexive we postpone the discussion of a class of spaces which contains in particular all l^1-type spaces to a later point, when we are dealing with non-reflexive Banach spaces. The results presented there will equally apply to reflexive Banach spaces such as l^1_n.
Our conjecture is that similar results can be proved for any reflexive Banach space. Once a space has been fixed the problem of geometric variety is eliminated and it should be possible to run arguments similar to what we are presenting below.

Let us begin by discussing the tangential bound which will make clear why strict convexity is the crucial property for a closed form result like in the previous sections. For simplicity we are going to just write "the ball" to refer to the norm ball of any radius r, i.e. $B_r = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq r\}$.

Lemma 6.12

A function $\Omega : \mathcal{B} \to \mathbb{R}$ is admissible if and only if for every exposed face of the 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

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

We are going to refer to the points that this statement applies to as *admissible points.*

Remark 6.13

Note that this in particular means that every exposed point is admissible and the bound applies to every functional exposing it. Further if the point is rotund the bound applies to every functional attaining its norm at the point.

Proof *(Of Lemma 6.12)*:

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

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

Since Ω is assumed to be admissible there exists a solution f_0 such that $c \cdot L \in J(f_0)$.
Now if there does not exist $g \in \mathcal{B}$ such that $g \neq f$ and $L \in J(g)$ then this can only be f itself as in previous sections. So then as before 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 $g \in \mathcal{B}$ such that $L \in J(g)$ we have no way of making a statement about how $\Omega(f)$ and $\Omega(g)$ compare. All we can say is that in this face containing f and g there is at least one point, where the minimum of Ω is attained. 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 claimed tangential bound.

Part 2: *(Nondecreasing along tangential directions $\Rightarrow$ Ω admissible)*
Conversely fix any data $(L_i, y_i) \in \mathcal{B}^* \times Y$ for $i \in \mathbb{N}_m$ such that the constraints can be satisfied. Let f_0 be a solution to the regularised interpolation problem. If $\operatorname{span}\{L_i\} \cap J(f_0) \neq \varnothing$ we are done, so assume not. We let

$$Z = \{f_T \in \mathcal{B} \, : \, L_i(f_T) = 0 \, \forall i \in \mathbb{N}_m\} = \bigcap_{i \in \mathbb{N}_m} \ker L_i$$

We want to show that there exists $f_T \in Z$ such that $\operatorname{span}\{L_i\} \cap J(f_0 + f_T) \neq \varnothing$. To see that this is true choose $V = \mathcal{B}$ and $W = Z$ in Theorem 5.15. Since Z

is a closed subspace of a reflexive space it is itself reflexive. Further choose $x_0 = f_0$ and $L_0 = 0$. Then the theorem says that there exists $f_T \in Z$ such that

$$J(f_0 + f_T) \cap (Z^\perp + 0) \neq \varnothing.$$

But $Z = \{L_i\}_\perp$ and so by Lemma 5.9 $Z^\perp = \operatorname{span}\{L_i\}$. Thus there exists $\hat{f} = f_0 + f_T$ which satisfies the interpolation constraints and such that

$$J(\hat{f}) \cap \operatorname{span}\{L_i\} \neq \varnothing.$$

Further for $\hat{L} \in J(\hat{f}) \cap Z^\perp$ we have $-f_T \in \ker(\hat{L})$. If $f_0 + f_T$ is exposed by $\hat{L}$ then the tangential bound applies and

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

so $\hat{f}$ is a solution of the regularised interpolation problem.
If on the other hand $f_0 + f_T$ is not exposed by $\hat{L}$, then it is contained in a face exposed by $\hat{L}$. But then for any $\overline{f_T} \in \mathcal{B}$ such that $\hat{f} + \overline{f_T}$ is still contained in this face we have that $\hat{L} \in J(f_0 + f_T + \overline{f_T})$ and $\overline{f_T} \in \ker(\hat{L})$ 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.

❑

From part 2 we see that points within a face F of the ball B_r are equivalent and we can in a sense move freely within the face. We will return to this point in Sections 6.6.1 and 7, where the implications of it will become more clear.

Having this theoretical result that an admissible regulariser is still in a sense tangentially nondecreasing, we are now going to turn to examining to what extent we can still give a similar geometrical interpretation as in the previous sections. We are going to see to what extent the circular bound as in Lemmas 6.4 and 6.9 remains true.

6.4.1 Strictly Convex Spaces

As mentioned previously for now we are only going to discuss the easiest case, namely when the space is strictly convex so every point in the ball is rotund and thus exposed. This means every point is admissible and we are in a situation similar to before. Thus in this section $\mathcal{B}$ is always strictly convex and reflexive. An example of a space which is reflexive, strictly convex but not uniformly convex is l_n^1 with the norm $\|\cdot\|_1 + \|\cdot\|_2$. The discussion of some spaces which fail to be strictly convex is postponed to when we are considering non-reflexive Banach spaces.

Lemma 6.14

If for every $f \in \mathcal{B}$ and all $f_T \in \bigcup_{L \in J(f)} \ker(L)$ we have $\Omega(f) \leq \Omega(f + f_T)$ then for any fixed $\hat{f} \in \mathcal{B}$ we have that

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

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

Proof:

Since the space is assumed to be strictly convex every point is exposed. The space may not be smooth in which case the duality mapping J is not univocal but for a non-smooth, rotund point f every $L \in J(f)$ exposes it. Thus Lemma 6.12 applies to all points $f \in \mathcal{B}$ and all functionals $L \in J(f)$. We thus do not need to worry about whether or not a point is an exposed point and whether it is exposed by a given functional attaining its norm at the point.

This means that we can follow the same general idea of argumentation as we did in the previous sections.

Part 1: *(Bound Ω on the half spaces given by the tangent planes through $\hat{f}$)*
We again start by showing that Ω is radially nondecreasing. With J possibly not being univocal this then gives a bound for all half spaces given by a tangent plane through $\hat{f}$, given by some $L \in J(\hat{f})$.
We fix some $\hat{f} \in \mathcal{B}$ and $\lambda > 1$ and set $f = \lambda \cdot \hat{f}$. To show that $\Omega(f) \geq \Omega(\hat{f})$ fix any $L_1 \in J(\hat{f})$ and $f_T \in \ker\{L_1\}$ and set

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

so that $f_t + g_t = f$. By the choice of f_T we have that $\Omega(\hat{f}) \leq \Omega(f_t)$ and as before we now need to show that there exists t_0 such that there exists $L_{t_0} \in J(f_{t_0})$ such that $g_{t_0} \in \ker\{L_{t_0}\}$. This would mean that $\Omega(f_{t_0}) \leq \Omega(f_{t_0} + g_{t_0}) = \Omega(f)$ as claimed.

To show that such a t_0 indeed exists we will consider choices $L_t \in J(f_t)$ for every t. Note first that, by definition of g_t

$$\begin{aligned} L_t(g_t) = 0 &\Leftrightarrow (\lambda - 1)L_t(\hat{f}) = tL_t(f_T) \\ &\Leftrightarrow \lambda L_t(\hat{f}) = L_t(\hat{f}) + tL_t(f_T) = L_t(f_t) \\ &\Leftrightarrow \lambda L_t(\hat{f}) = \|f_t\|^2, \end{aligned} \tag{51}$$

which gives us an equivalent condition to find a suitable t_0.
We now define the set-valued function $F : [0, \infty) \to \mathcal{P}(\mathbb{R})$

$$F(t) = \{\lambda L_t(\hat{f}) \subset \mathbb{R} : L_t \in J(f_t)\}$$

By Proposition 5.13 $J(f)$ is non-empty, weakly* closed and convex for every $f \in \mathcal{B}$ so the value of $F(t)$ is either a single value or an interval in $\mathbb{R}$.
It is known that if $\mathcal{B}$ is smooth then J is univocal and norm to weak* continuous so that F is clearly continuous. We show that if $\mathcal{B}$ is not smooth the function F is still almost continuous in the sense that in any jump the

function is interval valued and the interval connects both ends of the jump. To show this fix an arbitrary $t \in [0,\infty)$ and let $s \to t$. Then $f_s \to f_t$ in norm and hence for any choice of $L_s \in J(f_s)$ we have that $\|L_s\| = \|f_s\| \leq M$ for some constant M. Thus passing to a subsequence if necessary $L_s \overset{*}{\rightharpoonup} \widetilde{L}$, in particular $L_s(\hat{f}) \underset{s\to t}{\longrightarrow} \widetilde{L}(\hat{f})$.
We want to show that this $\widetilde{L}$ is indeed contained in $J(f_t)$. By standard results (c.f. Brezis [Bre11] Proposition 3.13 (iv)) we know that $L_s(f_s) \to \widetilde{L}(f_t)$ but also $L_s(f_s) = \|f_s\|^2 \to \|f_t\|^2$ so that

$$\widetilde{L}(f_t) = \|f_t\|^2. \tag{52}$$

Further $\|\widetilde{L}\| \leq \liminf \|L_s\| = \|f_s\|$ (c.f. Brezis [Bre11] Proposition 3.13 (iii)) and thus

$$\|\widetilde{L}\| \cdot \|f_t\| \leq \liminf \|L_s\| \cdot \lim \|f_s\| \leq \lim L_s(f_s) = \widetilde{L}(f_t) \leq \|\widetilde{L}\| \cdot \|f_t\|,$$

which means that

$$\widetilde{L}(f_t) = \|\widetilde{L}\| \cdot \|f_t\|. \tag{53}$$

Putting Eq. (52) and Eq. (53) together gives

$$\|f_t\|^2 = \widetilde{L}(f_t) = \|\widetilde{L}\| \cdot \|f_t\|,$$

which shows that indeed $\|\widetilde{L}\| = \|f_t\|$ and hence $\widetilde{L} \in J(f_t)$.
But this means that for $s \to t$ and any choice of $F(s)$ where F is not single valued there exists an $x \in F(t)$ such that $F(s) \to x$. This proves the claim that F is "effectively" continuous, in the sense that whenever the function would have a jump it is set valued and its interval value closes the gap between either side of the jump. This means that an intermediate value theorem holds for the function F.
Going back to Eq. (51) we see that it is satisfied if and only if $\|f_{t_0}\|^2 \in F(t_0)$. For $t = 0$, i.e. $f_0 = \hat{f}$, we have

$$F(0) = \lambda L_0(\hat{f}) = \lambda \|\hat{f}\|^2 > \|\hat{f}\|^2 = \|f_0\|^2.$$

On the other hand

$$F(t) = \lambda L_t(\hat{f}) \leq \lambda \|L_t\| \cdot \|\hat{f}\| = \lambda \|f_t\| \cdot \|\hat{f}\|.$$

But since $\|f_t\| \xrightarrow[t\to\infty]{} \infty$ we have $\lambda\|\hat{f}\| < \|f_t\|$ for t large enough and thus

$$F(t) = \lambda L_t(\hat{f}) \leq \lambda\|f_t\| \cdot \|\hat{f}\| < \|f_t\|^2$$

for large t. Since $\|f_t\|^2$ is continuous in t and the intermediate value theorem holds for F this means that there exists a t_0 such that $\|f_{t_0}\|^2 \in F(t_0)$ which means that there exists $L_{t_0} \in J(f_{t_0})$ such that Eq. (51) is satisfied. For this t_0 indeed

$$\Omega(\hat{f}) \leq \Omega(f_{t_0}) \leq \Omega(f_{t_0} + g_{t_0}) = \Omega(f).$$

Part 2: *(Extend the bound around the circle)*
Extending the bound around the circle as was done in the previous cases is in fact trivial. For points of smoothness of the norm this has already been shown to be possible in Lemma 6.9. In points of non-smoothness we have more than one tangent to the ball. But as the tangential bound on Ω holds for every tangent it is obviously always possible to choose a tangent which stays arbitrary close to the circle.

❑

Seeing that this result is effectively the same as Lemma 6.9 it is not surprising that we can obtain the same closed form characterisation of admissible regularisers as before.

Theorem 6.15

A function $\Omega : \mathcal{B} \to \mathbb{R}$ is admissible if and only if it is of the form

$$\Omega(f) = h(\|f\|_{\mathcal{B}})$$

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 $\Omega(f)$ is only constrained by the monotonicity property, i.e. it has to lie between $\lim\limits_{t\nearrow r} h(t)$ and $\lim\limits_{t\searrow r} h(t)$.

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.

The previous proof for Theorem 6.8 is in fact still entirely valid. Note in particular that from the fact that for any $f_T \in \bigcup_{L \in J(f)} \ker(L)$ we have

$$L(f + f_T) = L(f) = \|L\| \cdot \|f\| \leq \|L\| \cdot \|f + f_T\|$$

and so $\|f\| \leq \|f + f_T\|$. By strict convexity the inequality is in fact strict so that the bound for the mollification in the proof of Theorem 6.6 remains valid.
It is also clear that part 1 of the proof of Lemma 6.12 holds for $f = 0$ so 0 is an admissible point. Thus Ω is still without loss of generality minimised at 0 with $\Omega(0) = 0$.
All other parts of the proof are also clearly still valid and we omit the proof.

Another important case is the one of uniformly non-rotund spaces which we are going to define later in Section 6.6.1. This case includes in particular the finite dimensional and hence reflexive space l_n^1. Because uniformly non-rotund spaces also include many important non-reflexive spaces we postpone their discussion until Section 6.6. While we are going to be using a more general notion of solution there it is clear from the proofs of Lemmas 6.4, 6.9 and 6.14 that these results do not depend on the exact notion of solution. The discussions in Section 6.6.1 thus also exactly apply to reflexive, uniformly non-rotund spaces such as l_n^1.

6.5 Reflexivity Is Necessary

So far we have seen that the form of the representer theorem is crucially determined by the properties of the duality mapping. We have dealt with the duality mapping being nonlinear when extending from Hilbert spaces to uniform Banach spaces. We have further dealt with the duality mapping

possibly not being injective or univocal when considering reflexive Banach spaces. As stated in Proposition 5.21 the duality mapping is surjective if and only if the space is reflexive. This immediately suggests that one can not do better than reflexivity in the assumptions on the space without loosening other assumptions. In a non-reflexive Banach space we can find L_i which are not the image of any element in $\mathcal{B}$ under the duality mapping so that there is no hope of finding a solution in the sense of Definition 6.11.
As an example consider $\mathcal{B} = l_1$ with $\mathcal{B}^* = l_\infty$. Let $L_1 = (x_i)$ where $x_i = \frac{i}{i+1}$ for i odd and $x_i = 0$ for i even and $L_2 = (y_i)$ where $y_i = \frac{i}{i+1}$ for i even and $y_i = 0$ for i odd, i.e.

$$L_1 = (\frac{1}{2}, 0, \frac{3}{4}, 0, \ldots) \text{ and } L_2 = (0, \frac{2}{3}, 0, \frac{4}{5}, \ldots).$$

Then $\|L_1\|_\infty = \|L_2\|_\infty = 1$ but there cannot be an l_1-sequence of norm 1, x say, such that $L_1(x) = 1$ or $L_2(x) = 1$. So $L_1, L_2 \notin J(\mathcal{B})$. It is also clear by construction that

$$\operatorname{span}\{L_1, L_2\} \cap J(\mathcal{B}) = \{0\}.$$

This means there is no hope of finding a solution in the sense of Definition 6.11 with a dual element in the linear span of the defining linear functionals, i.e. there cannot be an $f_0 \in \mathcal{B}$ such that

$$J(f_0) \cap \operatorname{span}\{L_1, L_2\} \neq \emptyset.$$

Moreover noticing that the proof of Lemma 6.12 crucially relies on the Beurling-Livingston theorem (Theorem 5.15), which only requires that the subspace $Z = \{f_T \in \mathcal{B} : L_i(f_T) = 0 \, \forall i \in \mathbb{N}_m\}$ and not that $\mathcal{B}$ is reflexive, one might hope to be able to generalise the result a little further. But the main reason for thinking about removing, or at least weakening, the reflexivity assumption on $\mathcal{B}$ is l^1 which is not reflexive and very commonly used in applications. But it is clear that the Beurling-Livingston theorem cannot apply in this case. This is because Z is the intersection of kernels of finitely many linear functionals and thus an infinite-dimensional subspace of l^1. But l^1 does not contain any infinite-dimensional, reflexive subspace. To see this note that l^1 and all of its subspaces have the Schur property, that is norm and weak sequential convergence are equivalent. But if a subspace were reflexive then weak and weak* convergence would be equivalent. Thus the Banach-Alaoglu theorem would say that the unit ball is weak* so weak so norm compact which is a contradiction if the subspace is infinite-dimensional. Since the proof of the Beurling-Livingston theorem as given in Section 5.3 relies on reflexivity of Z for the existence of a minimiser we see that there is no hope of it applying in the l^1 case.

These two arguments make clear that if we want to extend the results to include l^1 we need to weaken one of our other assumptions to have any hope of success. This is going to be the aim of the next section.

6.6 Non-reflexive Banach Spaces

As shown in Section 6.5 reflexivity is necessary to have any hope of proving the existence of a solution in the spirit of previous results. The only hope to obtain a result for non-reflexive Banach spaces is to weaken another assumption. We propose to consider a notion of an *approximate solution* and hence an *approximate representer theorem*. This will be justified and made more precise below. We are going to show that for this weaker concept of solutions we can indeed obtain the immediate generalisations of the previous sections. Thus the results finally apply to all types of spaces which appeared throughout this thesis, including non-reflexive RKBS as introduced in Section 4.2. In particular l^1 is an example of a non-reflexive Banach space, and the main motivation for this section because of its use in applications.

In this section we let $\mathcal{B}$ be a Banach space which might not be reflexive with duality mapping

$$J(x) = \{L \in \mathcal{B}^* \,:\, L(x) = \|L\| \cdot \|x\|, \|L\| = \|x\|\}\,.$$

We then consider essentially the same regularised interpolation problem as before

$$\inf\{\Omega(f) \,:\, f \in \mathcal{B},\, L_i(f) = y_i \;\forall i \in \mathbb{N}_m\}\,. \tag{54}$$

The only difference compared to before is that we now consider the infimum since the question of attainment is more delicate as it was in a reflexive Banach space. More precisely there are two crucial differences compared to what we have seen before.

- Firstly it is known that even for $\Omega(f) = \|f\|_{\mathcal{B}}$ there need not be a solution to the regularised interpolation problem. Thus we a priori cannot expect a solution to always exist anymore, which is why we changed from considering the minimum to the infimum in Eq. (54). But it is easy to see that if some function $\overline{f} \in \mathcal{B}$ satisfies the interpolation constraints then minimising
$$\inf\{\|f\|_{\mathcal{B}} : L_i(f) = y_i \ \forall i \in \mathbb{N}_m\}$$
is equivalent to minimising
$$\inf\{\|\overline{f} + f_T\|_{\mathcal{B}} : f_T \in \bigcap_{i \in \mathbb{N}_m} \ker(L_i)\}.$$
Or in other words the infimum of the minimal norm interpolation is attained for a function f_0 if and only if the distance of 0 to the affine space $\overline{f} + \bigcap_{i \in \mathbb{N}_m} \ker(L_i) = \overline{f} + Z$ is attained at $f_0 \in \overline{f} + Z$. Now different values for the y_i correspond to different shifts of Z so that the distance is attained at different points. For a solution to always exist we thus need Z to be proximinal. We thus will not assume that a solution of (54) always exists as we did throughout the previous sections, but we will rather make the weaker assumption that a solution to Eq. (54) always exists if Z is proximinal, i.e. if the distance of any point in $\mathcal{B}$ to Z is attained (Definition 5.24).

- The second crucial difference to reflexive Banach spaces is that now some or all of the L_i might not be in the image of the duality mapping. And even if they all are in the image of the duality mapping, since the duality mapping is not linear it is not hard to construct an example where the linear span is not entirely contained in the image of the duality mapping. As discussed in Section 6.5 this requires another adjustment of our assumptions. With each generalisation step taken in the previous sections we have slightly generalised the notion of admissibility in the natural way to reflect the properties of the class of spaces considered. We thus follow the same approach here and generalise the notion of admissibility to reflect the properties of a non-reflexive Banach space. More precisely we argued in Section 6.5 that the lack of reflexivity was an issue as it means that the duality mapping ceases to be surjective. But as stated in Section 5.4 every Banach space is subreflexive, i.e. the image of the duality mapping is norm dense in the dual space. This suggests that, while we cannot hope for a solution with dual element in the linear span of the functionals, we might expect to be able to get arbitrarily close to the linear span.

These two points lead to the aforementioned notion of approximate solution and approximate representer theorem. In the case of the intersection of the kernels of the L_i being proximinal we obtain the previous exact representer theorem, otherwise we can only hope for approximate solutions.

Definition 6.16 *(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 $\{L_1, \dots, L_m\} \subset \mathcal{B}^*$ and $\{y_1, \dots, y_m\} \subset Y$ such that the interpolation constraints can be satisfied the regularised interpolation problem Eq. (50) either

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

$$\hat{L} = \sum_{i=1}^{m} c_i L_i \in J(f_0)$$

if $\bigcap_{i \in \mathbb{N}_m} \ker(L_i)$ is proximinal;

(ii) Or otherwise admits for every $\varepsilon > 0$ an approximate solution f_0^ε such that

$$\Omega(f_0^\varepsilon) \leq \inf\{\Omega(f) \,:\, f \in \mathcal{B},\, L_i(f) = y_i \; \forall i \in \mathbb{N}_m\} + \varepsilon$$

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

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

Remark 6.17

(i) In the case of reflexive Banach spaces $\bigcap_{i \in \mathbb{N}_m} \ker(L_i)$ was reflexive. It will become clear in the proof of Lemma 6.18 that in fact $\bigcap_{i \in \mathbb{N}_m} \ker(L_i)$ being proximinal is the minimal requirement for the existence of a solution in the sense of Definition 6.16 (i). General and easily applicable criteria for a subspace to be proximinal are still an open area of research. We have presented some relevant results which characterise proximinal subspaces in Section 5.5.

(ii) To see that Definition 6.16 (ii) is the best we can hope for in general consider the case $\mathcal{B} = l^1$, $\mathcal{B}^* = l^\infty$. Let

$$L = \left(\frac{n}{n+1}\right)_{n \in \mathbb{N}} = \left(\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots\right)$$

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 6.16 (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

$$\|\hat{L} - L\|_{l^\infty} \geq \hat{L}_i - L_i > (1 + \delta) - 1 = \delta > 0$$

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 $\varepsilon > 0$. That this is indeed possible will be proved below.

As already in the reflexive case the approach to be taken now is to prove that being in a sense nondecreasing along tangents is still a necessary and sufficient condition for admissibility. Subsequently we can look at the geometric interpretation of this result for a certain class of spaces to reduce the geometric variety. This class is chosen to include in particular l^1 for its use in applications. The result we obtain is the same as for reflexive Banach spaces.

Lemma 6.18

A function $\Omega : \mathcal{B} \to \mathbb{R}$ is admissible if and only if for every exposed face of the 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

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

We are going to refer to the points that this statement 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

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

As Ω is assumed to be admissible and $\ker(L)$ is proximinal by Lemma 5.28 there exists a solution f_0 such that $c \cdot L \in J(f_0)$. This means that we are in case (i) of Definition 6.16. We can thus argue exactly as in the case of a reflexive space, the short proof is repeated for convenience.

If there does not exist a $g \in \mathcal{B}$ such that $g \neq f$ and $L \in J(g)$ then this solution can only be f itself as in previous sections. So then as before 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 a $g \in \mathcal{B}$ such that $L \in J(g)$ we have no way of making a statement about how $\Omega(f)$ and $\Omega(g)$ compare. All we can say is that in this face there is at least one point where the minimum of Ω is attained. 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 ⇒ Ω admissible)*
Conversely fix any data $(L_i, y_i) \in \mathcal{B}^* \times Y$ for $i \in \mathbb{N}_m$ such that the constraints can be satisfied and let

$$Z = \{f_T \in \mathcal{B} \,:\, L_i(f_T) = 0 \,\forall i \in \mathbb{N}_m\} = \bigcap_{i \in \mathbb{N}_m} \ker L_i.$$

We now have two possible cases. Either Z is proximinal and we are looking for a solution in the sense of Definition 6.16 (i), or Z is not proximinal and we are in the situation of Definition 6.16 (ii).

Case 1: Assume first that Z is proximinal, so we are looking for a solution in the sense of Definition 6.16 (i). By assumption there exists a solution to the regularised interpolation problem. Let f_0 be such a solution. If $\operatorname{span}\{L_i\} \cap J(f_0) \neq \varnothing$ then f_0 is a solution in the sense of Definition 6.16 (i) and we are done, so assume not. This means we want to show that there exists $f_T \in Z$ such that $\operatorname{span}\{L_i\} \cap J(f_0 + f_T) \neq \varnothing$.

In part 2 of the proof of Lemma 6.12 we applied the Beurling-Livingston theorem (Theorem 5.15) with $V = \mathcal{B}$, $W = Z$, $x_0 = f_0$ and $L_0 = 0$. Examining the proof of Theorem 5.15 we see that in that case the proof starts by minimising the functional

$$F(f) = M(f - f_0) = \int_0^{\|f - f_0\|} t \,\mathrm{d}t = \frac{\|f - f_0\|^2}{2}.$$

The existence of a minimiser is guaranteed by noting that F is continuous, convex and coercive so it attains its minimum on the, in that case reflexive, space Z. In the current case Z may not be reflexive but we notice that the functional to be minimised is $\frac{\|f - f_0\|^2}{2}$ which, as $f_0 \notin Z$, clearly attains its minimum if and only if the metric projection of f_0 onto Z exists. In other words the minimiser of F in Z always exists if and only if Z is proximinal, which it is by assumption. The rest of the proof of the Beurling-Livingston theorem (Theorem 5.15) does not use any further properties of Z, it solely relies on the point in question being the minimiser of F. The rest of the proof thus remains valid and we thus obtain an $f_T \in Z$ such that

$$J(f_0 + f_T) \cap (Z^\perp + 0) \neq \varnothing.$$

We can proceed as before in part 2 of the proof of Lemma 6.12. Again we repeat the remaining short argument for convenience.
Since $Z = \{L_i\}_\perp$ by Lemma 5.9 we have that $Z^\perp = \operatorname{span}\{L_i\}$ so there exists $\hat{f} = f_0 + f_T$ which satisfies the interpolation constraints and is such that $J(\hat{f}) \cap \operatorname{span}\{L_i\} \neq \varnothing$. It remains to show that $\hat{f}$ indeed minimises Ω.

For $\hat{L} \in J(\hat{f}) \cap Z^\perp$ we have $-f_T \in \ker(\hat{L})$. If $f_0 + f_T$ is exposed by $\hat{L}$ then the tangential bound applies and

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

so $\hat{f}$ is a solution of the regularised interpolation problem.
If on the other hand $f_0 + f_T$ is not exposed by $\hat{L}$, then it is contained in a face exposed by $\hat{L}$. But then for any $\overline{f_T} \in \mathcal{B}$ such that $\hat{f} + \overline{f_T}$ is still contained in this face we have that $\hat{L} \in J(f_0 + f_T + \overline{f_T})$ and $\overline{f_T} \in \ker(\hat{L})$ 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 on the other hand Z is not proximinal we are in the case of Definition 6.16 (ii). The existence of a solution to the regularised interpolation problem is not guaranteed but there exists f_0^ε which almost attains the infimum, i.e.

$$\Omega(f_0^\varepsilon) \leq \inf\left\{\Omega(f) \,:\, f \in \mathcal{B},\, L_i(f) = y_i \,\forall i \in \mathbb{N}_m\right\} + \varepsilon.$$

If $\operatorname{dist}(J(f_0^\varepsilon), \operatorname{span}\{L_i\}) \leq \varepsilon$ then f_0^ε is a solution in the sense of Definition 6.16 (ii) and we are done, so assume not. This means we want to show that for every $\varepsilon > 0$ there exists $f_T^\varepsilon \in Z$ such that $\operatorname{dist}(J(f_0^\varepsilon + f_T^\varepsilon), \operatorname{span}\{L_i\}) \leq \varepsilon$.

We again look at the proof of the Beurling-Livingston theorem (Theorem 5.15). With Z not proximinal we do not get a minimiser of the functional

$$F(f) = M(f - f_0^\varepsilon) = \int_0^{\|f - f_0^\varepsilon\|} t \,\mathrm{d}t = \frac{\|f - f_0^\varepsilon\|^2}{2}$$

anymore. But considering its restriction to Z, Ekeland's variational principle [Eke74] states that for every $\varepsilon > 0$ there exists an $f_T^\varepsilon \in Z$ which almost minimises $F\big|_Z$, i.e.

$$F(f_T^\varepsilon) \leq \inf_{f \in Z} F(f) + \varepsilon$$

and further

$$F(f_T^\varepsilon) - F(g) < \varepsilon \cdot \|f_T^\varepsilon - g\| \quad \forall f_T^\varepsilon \neq g \in Z \tag{55}$$

Choosing $g = f_T^\varepsilon + th$ for $h \in Z$ in Eq. (55) we obtain

$$\begin{aligned} F(f_T^\varepsilon) - F(f_T^\varepsilon + th) &< \varepsilon \cdot \|f_T^\varepsilon - (f_T^\varepsilon + th)\| \\ \Leftrightarrow -\varepsilon \cdot \|h\| &< \frac{F(f_T^\varepsilon + th) - F(f_T^\varepsilon)}{|t|}. \end{aligned}$$

This means that

$$F'(f_T^\varepsilon, h) = \lim_{t \searrow 0} \frac{F(f_T^\varepsilon + th) - F(f_T^\varepsilon)}{t} > -\varepsilon \cdot \|h\|. \tag{56}$$

We now apply the sandwich theorem (Corollary 5.6) to show that there exists a linear functional in the subdifferential of F with small norm. The theorem says that there exists an $L \in Z^*$ such that $L(\cdot) \leq F'(f_T^\varepsilon, \cdot)$, which means that $L \in \partial F(f_T^\varepsilon)$ by Proposition 5.4. Furthermore for B_Z the unit ball in Z

$$\inf_{h \in B_Z} L(h) = \inf_{h \in B_Z} F'(f_T^\varepsilon, h) \overset{(56)}{>} -\varepsilon \cdot \|h\|$$

which implies that $\|L\|_{Z^*} < \varepsilon$.

We again proceed similarly to the proof of the Beurling-Livingston theorem. As before we have a linear functional on Z which needs to be extended to $\mathcal{B}$ in a suitable way. The intuition now is that while previously the functional was 0 on Z, and hence in the span of the L_i, it is now of small norm on Z and thus close to the span of the L_i as claimed.
To extend L to $\mathcal{B}$ we follow the ideas of part of the proof of the Beurling-Livingston theorem. We 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 $\varepsilon \leq \|f_T^\varepsilon - f_0^\varepsilon\|$ 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, like before in the proof of the Beurling-Livingston theorem, 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 $-L \in J(f_0^\varepsilon - f_T^\varepsilon)$.
Now $\overline{f_0^\varepsilon} = f_0^\varepsilon + f_T^\varepsilon$ satisfies the interpolation constraints and there exists an $L \in J(\overline{f_0^\varepsilon})$ such that $\|L\big|_Z\| < \varepsilon$. But this means that $\mathrm{dist}(L, Z^\perp) < \varepsilon$. Since by Lemma 5.9 $\mathrm{span}\{L_i\} = Z^\perp$ this means that $\mathrm{dist}(L, \mathrm{span}\{L_i\}) < \varepsilon$ so that $\overline{f_0^\varepsilon}$ satisfies the assumptions of Definition 6.16 (ii).
It remains to show that $\overline{f_0^\varepsilon}$ indeed minimises Ω. But this follows in exactly the same way as the argument at the end of case 1. If $\overline{f_0^\varepsilon}$ is an exposed point or minimum in its face, then it satisfies the tangential bound and thus

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

If $\overline{f_0^\varepsilon}$ is not exposed or a minimum, then 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

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

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

❑

Having obtained the same statement of Ω being tangentially nondecreasing as before we now turn again to the geometric interpretation of this result. As before we need to impose some extra geometric assumptions to reduce the geometric variety in order to be able to make any general statements. With the most important examples of non-reflexive spaces being l^1 and L^1 it is now time to examine what can be said about such spaces. We choose the geometric condition so that the resulting class of function spaces in particular contains those spaces.

6.6.1 Uniformly Non-rotund Spaces

So far we have only discussed the geometric interpretations of the tangential bound for spaces that are at least strictly convex. This was because if the space is not strictly convex the ball includes straight line sections. These faces can be of any dimensionality and it is very hard to make any statements of the geometrical structure of the ball in general. Before we start the discussion of how to deal with spaces which fail to be strictly convex let us gain some more intuition by illustrating why this geometric variety makes it very hard to make any general statements.

Our first step in obtaining a geometric interpretation of the tangential bound has always been proving monotonicity along any ray. It is clear that the discussions on monotonicity in the proof of Lemma 6.14 remain true for any Banach space. In particular we know that if we fix $\hat{f} \in \mathcal{B}$ and $\lambda > 1$ and set $f = \lambda \hat{f}$ then for any $\hat{L} \in J(\hat{f})$ and every $f_T \in \ker(\hat{L})$ there exists a $t(\hat{L}, f_T) \in \mathbb{R}$ depending on the choices of $\hat{L}$ and f_T, such that for

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

there exists $L \in J(f_{t(\hat{L},f_T)})$ with $g_{t(\hat{L},f_T)} \in \ker(L)$, i.e. for any tangent at $\hat{f}$ there exits a point along the tangent so that the line back to f is tangent at this point. But as the space is not strictly convex anymore we do not know a priori whether the tangential bound of Lemma 6.18 applies to this point. What we would need to prove is that there exists a choice of $\hat{L} \in J(\hat{f})$ and $f_T \in \ker(\hat{L})$ such that the corresponding $f_{t(\hat{L},f_T)}$ is a minimum for Ω in an exposed face (or obviously in particular $f_{t(\hat{L},f_T)}$ an exposed point) and the $L \in J(f_{\hat{L},f_T})$ exposes this face. If we are able to prove this we get the same half space bound as before. But there is essentially no hope to prove that a given point is an exposed point in any fixed but unknown Banach space. Due to the huge geometric variety we have little chance to make any statement about the properties of a point based on points in its neighbourhood. We can e.g. construct a Banach space with a rotund point such that no point in its neighbourhood is rotund. Similarly a convex function on $\mathbb{R}$ may not be differentiable on a countable dense subset (c.f. e.g. [LPT12, Phe93]) so also smoothness does not allow statements about surrounding points. Worse still, there might not even be any exposed point, e.g. the space c_{00} does not contain exposed points. Moreover proving a certain point in a given face is a minimum for Ω is impossible as a statement of how Ω looks like is precisely what we are trying to obtain.

This demonstrates why we need to impose additional assumptions on the geometry of the space $\mathcal{B}$ to have any hope of obtaining a geometric result as in the previous sections. In this section we are going to discuss a class of not strictly convex spaces, namely uniformly non-rotund spaces. For the rest of the section any $\mathcal{B}$ will be assumed to be a uniformly non-rotund Banach space.

Definition 6.19 *(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.

This definition is inspired by uniform non-smoothness as defined in Definition 5.23. It is chosen to in particular include l^1 and L^1 for their use in applications. With strictly convex and uniformly non-rotund spaces every space we know of which is commonly used in applications is covered by our discussions. We conjecture, as already stated in Section 6.4, that similar arguments are possible for any Banach space once the space has been fixed. More precisely, if a space is relevant for an application it should be an easy check that the same proof strategies can be applied to obtain the analogous results. This is 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.

The reason this definition is useful is because it means that there cannot exist faces with a smooth boundary. If any part of the boundary of a face were smooth one would be able to find a two-dimensional subspace in which the smooth boundary point is rotund. If no point in the boundary of a face is smooth, the boundary has to consist of faces of a lower dimension. These faces are exposed by another functional and contain their own minimum of Ω. Since the boundary of a face is always the intersection of the face with an adjacent face we can show that we are always able to reach this minimum on the boundary and move along further from there.

It will become clear in this section that in spaces which are not strictly convex it is often convenient to think of Ω as a function of the faces of the ball. In other words we may be thinking of the faces as being collapsed to one point where Ω is minimised.

Lemma 6.20

If for every exposed face of the 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 $\hat{f} \in \mathcal{B}$, that is any $\hat{f} \in \mathcal{B}$ the above applies to, we have that

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

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

Proof:

As discussed in the introduction of this chapter the arguments for proving monotonicity along a ray in the proof of Lemma 6.12 remain true, more precisely tangents to go "out and back" always exist, but it is now not immediately clear that the tangential bound Lemma 6.18 applies to any of those tangents. As it is difficult to show that any given point f_t is admissible we are instead going to prove that we can always find an admissible point which allows us to bound points on the ray $\lambda \cdot \hat{f}$. This admissible point will be the minimum of a face of the boundary of the original face as described above. The same approach is used to extend the bound around the circle.

Before we start with the details of the proof we make a few remarks. Firstly, for this result it is at times convenient to view Ω as a function of faces of the unit ball. We saw in Section 6.4 that points within a face were equivalent for the problem and we could always move within a face to select the point for which Ω is minimised. This indicates that we can think of Ω as a map $\overline{\Omega}$, taking a face F of the norm ball in $\mathcal{B}$ and mapping it to the minimum of Ω across that face, i.e. $\overline{\Omega}(F) = \min_{f \in F} \Omega(f)$.
Secondly we note that to show monotonicity of a ray we do not necessarily need to prove it for all $\lambda > 1$ as in previous sections but we can restrict ourselves to $1 < \lambda < 1 + \varepsilon$ and as long as the ε is at least nondecreasing as a function of the norm along the ray we get monotonicity of the entire ray.

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, i.e. the minimum of Ω within a face is nondecreasing as a function of the norm. Since the 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 $\hat{L} \in J(\hat{f})$.

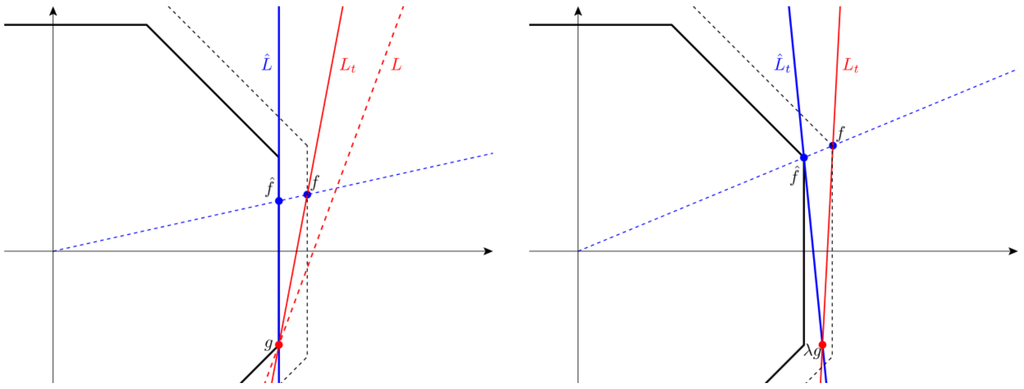
We fix an admissible $\hat{f} \in \mathcal{B}$ and let X be any 2-dimensional subspace containing $\hat{f}$. As $\mathcal{B}$ is uniformly non-rotund no point in X is rotund so its unit ball consists of straight line sections and corners as shown in Fig. 10. In particular $\hat{f}$ is within a straight line section or a corner in between two straight line sections. Then 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 the Hahn-Banach theorem there are extensions of these functionals to $\mathcal{B}$ which we will also denote by $\hat{L}$ and L, as we will only be considering the extensions. The functional $\hat{L}$ exposes a face $\hat{F}$ containing the straight line from $\hat{f}$ to g. The functional L exposes a face F which is at least the point g, possibly a face containing g, but not containing $\hat{f}$.
We now let

$$L_t = t\hat{L} + (1-t)L \qquad t \in (0,1).$$

Then 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 for the minimum across a face containing $\hat{f}$. Indeed

- 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. 10a.

- If $\hat{f}$ 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. 10b.



(a) $\hat{L}$ leads from the minimum $\hat{f}$ of a face to an exposed point g, L_t leads from g to the minimum f of the face above.

(b) $\hat{L}_t$ leads from an exposed point $\hat{f}$ to an exposed point λg, L_t leads from λg to the minimum f of the face above.

Figure 10: Admissible points bound the minima of the face above

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 space spanned by a tangent plane at the point as in the previous cases.

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 same bound as in previous cases that any admissible point bounds 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. Within X we can get from $\hat{f}$ to $\overline{f}$ along tangents by moving from $\hat{f}$ to the next exposed point towards $\overline{f}$, g_1 say. This step is either within the straight line section if $\hat{f}$ is the minimum of its face or via an arbitrarily close tangent if $\hat{f}$ is an exposed point. We then go from exposed point to exposed point within X by tangents arbitrarily close to the circle, which obviously exist. This way we obtain a sequence of points $g_i \in X$ as shown in Fig. 11.

It remains to show, similarly to the argument in part 1, that if a g_i is not admissible we can find a $\overline{g_i}$ which is admissible and allows to effectively maintain the same path from $\hat{f}$ to $\overline{f}$. Assume L_i is the functional that exposes the straight line section from g_{i-1} to g_i in X, and consider its extension to $\mathcal{B}$ by the Hahn-Banach theorem as before. Here we set $g_0 = \hat{f}$. Denote by F_i the face exposed by L_i and let

$$L_{i,t} = tL_i + (1-t)L_{i+1}, \qquad t \in (0,1).$$

Then $L_{i,t}$ exposes the face $F_{i,t} = F_i \bigcap F_{i+1}$ which in particular contains g_i and has a minimum $\overline{g_i}$. So $L_{i,t}$ for t close to 0 provides a tangent from either g_i or $\overline{g_i}$ to g_{i+1} or if necessary $\overline{g_{i+1}}$ as those are contained in $F_{i+1,t} \subset F_{i+1}$. This is illustrated in Fig. 11.

Each step includes a step away from the circle, but it can always be made arbitrarily small by varying t. For the next step we consider the corresponding tangent at this point λg_i rather than g_i. It is clear that with this process

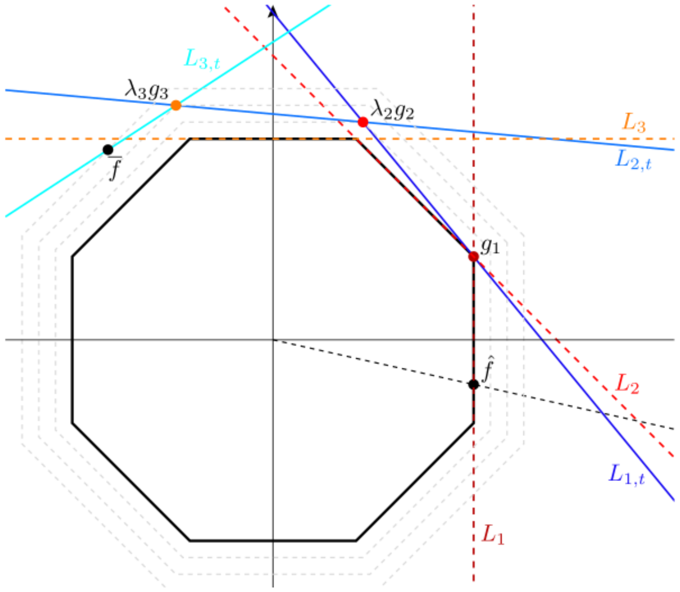


Figure 11: The bound can be extended around the circle along points which are exposed in the two dimensional subspace.

we can reach any admissible $\overline{f}$ or at least $\lambda\overline{f}$ for $\lambda < 1$ which is sufficient by monotonicity of the minima from part 1.

❑

This proof makes clear that 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 can thus only obtain a result about radial symmetry in the spirit of Theorems 6.3, 6.6, 6.8 and 6.15 by viewing Ω as a function of the faces. If we think of Ω in this way then the same intuition of almost radial symmetry as before applies.

Theorem 6.21

A function $\Omega : \mathcal{B} \to \mathbb{R}$ is admissible if and only if, viewed as a function $\overline{\Omega}$ of the faces 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 $h\colon[0,\infty) \to \mathbb{R}$ 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 between $\lim\limits_{t\nearrow r} h(t)$ and $\lim\limits_{t\searrow r} h(t)$.

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 the face F.

Proof:
The proof given in the previous cases in fact remains largely valid, only a few extra considerations are required. We are going to briefly discuss sections which remain valid and present in full any extra arguments which are required.

Part 1: *(Ω continuous in radial direction implies Ω radially symmetric)*
We can obviously only obtain radial symmetry for admissible points. Since admissible points are the minimum within their face they bound non-admissible points from below. Thus we only really need a statement about admissible points. For admissible points the previous argument that continuity in radial direction implies radial symmetry in fact remains valid.
Indeed if we assume f and g are admissible points of the same norm and $\Omega(f) > \Omega(g)$ say, then by Lemma 6.20 for all $1 < \lambda \in \mathbb{R}$ we have $\Omega(\lambda g) \geq \Omega(f)$. As discussed in Section 6.4 Ω is still minimised at 0 and thus without loss of generality non-negative. By Lemma 6.18 it is non-decreasing as a function of the face which is enough to conclude that $|\Omega(\lambda g) - \Omega(g)| \geq |\Omega(f) - \Omega(g)| > 0$ contradicting radial continuity of Ω.

Part 2: *(Radial mollification preserves being nondecreasing in tangential directions)*
For the observation in part 1 to be useful we need to verify again that the radially mollified regulariser

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

is admissible if Ω is admissible.
More precisely we check that this function is still non-decreasing along tangential directions, i.e. we need to show that for an admissible f for any $L \in J(f)$ exposing the face containing f and every $f_T \in \ker(L)$ we still have

$$\begin{aligned}\widetilde{\Omega}(f + f_T) = \int_{-1}^{0} \rho(t)\Omega\left((\|f + f_T\| - t)\frac{f + f_T}{\|f + f_T\|}\right) \mathrm{d}t \\ \geq \int_{-1}^{0} \rho(t)\Omega\left((\|f\| - t)\,\frac{f}{\|f\|}\right) \mathrm{d}t = \widetilde{\Omega}(f). \quad (57)\end{aligned}$$

The way this was previously argued is that if $\|f + f_T\| > \|f\|$ then also

$$\|(\|f + f_T\| - t)\frac{f + f_T}{\|f + f_T\|}\| > \|(\|f\| - t)\frac{f}{\|f\|}\|$$

for all $t \in [-1, 0]$. But then Lemma 6.20 implies that

$$\Omega\left((\|f + f_T\| - t)\frac{f + f_T}{\|f + f_T\|}\right) \geq \Omega\left((\|f\| - t)\frac{f}{\|f\|}\right)$$

for all $t \in [-1, 0]$.
It thus remains to check the case when $\|f + f_T\| = \|f\|$. But in this case we have that

$$\Omega\left((\|f + f_T\| - t)\frac{f + f_T}{\|f + f_T\|}\right) = \Omega\left((\|f\| - t)\frac{f + f_T}{\|f\|}\right) = \Omega\left(\frac{\|f\| - t}{\|f\|}(f + f_T)\right).$$

Since $\frac{\|f\|-t}{\|f\|} f_T \in \ker(L)$ we have that indeed

$$\Omega\left((\|f + f_T\| - t)\frac{f + f_T}{\|f + f_T\|}\right) \geq \Omega\left((\|f\| - t)\frac{f}{\|f\|}\right)$$

so that the property of being nondecreasing along all tangents is preserved.

Part 3: *(Ω is as claimed)*
Putting these two observations together we obtain the result. We know that as a function of the faces $\overline{\Omega}$ is a monotone function of the norm, so a monotone function on the real line. Part 1 shows that after mollification $\overline{\Omega}$ is in fact radially symmetric. The same considerations as before say that $\overline{\Omega}$ must have been of the claimed form.

The converse is clear again, 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 $g \neq f$ for 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. 10a, 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 $\Omega(\lambda f) \xrightarrow[\lambda\to 1]{} \Omega(f)$ 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 all Banach spaces we know of which are commonly used in applications.
It seems that for any fixed Banach space $\mathcal{B}$ which is not strictly convex but also not uniformly non-rotund a combination of previous arguments, moving to either exposed points or minima in lower dimensional faces, should allow to prove a similar radial symmetry result. This would mean that the above statement is true without assumptions on the Banach space and an admissible regulariser on any Banach space has to be essentially radially symmetric

in the appropriate sense.

We are going to close the chapter by showing that every representer theorem presented above also holds for regularisation problems and under very mild additional assumptions this can even be made an equivalence and a regulariser is admissible for regularised interpolation if and only if it is admissible for regularisation.

6.7 Regularisation and Interpolation

As we have seen in the earlier parts of this thesis we are generally interested in problems of the form

$$\min\left\{\mathcal{E}_z((L_i(f), y_i)_{i\in\mathbb{N}_m}) + \lambda\Omega(f) \,:\, f \in \mathcal{B}\right\} \tag{58}$$

where $\mathcal{B}$ is a Banach space and the L_i are continuous linear functionals on $\mathcal{B}$ with the $y_i \in Y \subseteq \mathbb{R}$ the corresponding output data.

Argyriou, Micchelli and Pontil [AMP09] show in the Hilbert space case that under very mild conditions this regularisation problem admits a linear representer theorem if and only if the regularised interpolation problem

$$\min\left\{\Omega(f) \,:\, f \in \mathcal{B}, L_i(f) = y_i \;\forall i \in \mathbb{N}_m\right\} \tag{59}$$

admits a linear representer theorem. Or in other words, the pair $(\mathcal{E}_z, \Omega)$ is admissible for (58) if and only if Ω is admissible for (59). Here by admissibility for the regularisation problem we mean the obvious analogues of Definitions 6.1, 6.7, 6.11 and 6.16 for regularisation depending on the properties of the function space $\mathcal{B}$.

This 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. More precisely they proved the following theorem for the Hilbert space setting.

Note that the assumptions on the error function and regulariser presented here are as in the paper [AMP09]. It is remarked in that paper that other conditions can also be sufficient.

Theorem 6.22

Let $\mathcal{E}_z$ 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 $\min\{\mathcal{E}_z\left((a\nu_i, y_i)_{i\in\mathbb{N}_m}\right) : a \in \mathbb{R}\}$.

Assume that the regulariser Ω is lower semicontinuous and has bounded sublevel sets.

Then Ω is admissible for the regularised interpolation problem (59) if the pair $(\mathcal{E}_z, \Omega)$ is admissible for the regularisation problem (58).

The proof of this result follows the earlier mentioned concept that one obtains a regularised interpolation problem as the limit of regularisation problems. The proof we are giving is essentially the same as the one given by Argyriou, Micchelli and Pontil in [AMP09]. The more general setting considered in this work only requires a few adjustments.

Proof:

To prove that Ω is admissible for the regularised interpolation problem (59) we are going to show that Ω is tangentially nondecreasing in the sense of Lemmas 6.2, 6.10, 6.12 and 6.18 depending on the properties of the space $\mathcal{B}$. To begin with we are going to assume that $\mathcal{B}$ is at least reflexive. Fix $0 \neq f \in \mathcal{B}$ and $L \in J(f)$. Let a_0 be the unique nonzero minimiser of $\min\{\mathcal{E}_z\left((a\nu_i, y_i)_{i\in\mathbb{N}_m}\right) : a \in \mathbb{R}\}$. For every $\lambda > 0$ consider the regularisation problem

$$\min\left\{\mathcal{E}_z\left(\left(\frac{a_0}{\|L\|^2}L(f)\nu_i, y_i\right)_{i\in\mathbb{N}_m}\right) + \lambda\Omega(f) : f \in \mathcal{B}\right\}.$$

By admissibility of the pair $(\mathcal{E}_z, \Omega)$ there exist solutions $f_\lambda \in \mathcal{B}$ such that

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

i.e. there exist $c_\lambda \in \mathbb{R}$ such that $c_\lambda L \in J(f_\lambda)$.

Now fix any $g \in \mathcal{B}$ such that $L(g) = \|L\|^2$. We then obtain

$$\begin{aligned}\mathcal{E}_z\left((a_0\nu_i, y_i)_{i\in\mathbb{N}_m}\right) + \lambda\Omega(f_\lambda) &\leq \mathcal{E}_z\left(\left(\frac{a_0}{\|L\|^2}L(f_\lambda)\nu_i, y_i\right)_{i\in\mathbb{N}_m}\right) + \lambda\Omega(f_\lambda) \\ &\leq \mathcal{E}_z\left((a_0\nu_i, y_i)_{i\in\mathbb{N}_m}\right) + \lambda\Omega(g),\end{aligned} \tag{60}$$

where the first inequality follows from a_0 minimising $\mathcal{E}_z((a\nu_i, y_i)_{i\in\mathbb{N}_m})$ and the second inequality from $L(g) = \|L\|^2$. This shows that $\Omega(f_\lambda) \leq \Omega(g)$ for all λ and so by the hypothesis that Ω has bounded sublevel sets the set $\{f_\lambda : \lambda > 0\}$ is bounded.

Hence it is weakly* compact and there exists a weakly* convergent subsequence $(f_{\lambda_l})_{l\in\mathbb{N}}$ such that $\lambda_l \xrightarrow[l\to\infty]{} 0$ and $f_{\lambda_l} \overset{*}{\rightharpoonup} \overline{f}^{**}$ as $l \to \infty$. Since $\mathcal{B}$ is reflexive the sequence also converges weakly to $\overline{f} \in \mathcal{B}$. Taking the limit inferior as $l \to \infty$ on the right-hand side of Eq. (60) we obtain

$$\mathcal{E}_z\left(\left(\frac{a_0}{\|L\|^2}L(\overline{f})\nu_i, y_i\right)_{i\in\mathbb{N}_m}\right) \leq \mathcal{E}_z((a_0\nu_i, y_i)_{i\in\mathbb{N}_m}).$$

Since a_0 is by assumption the unique, nonzero minimiser this means that

$$\frac{a_0}{\|L\|^2}L(\overline{f}) = a_0 \Leftrightarrow L(\overline{f}) = \|L\|^2.$$

But then since $L(\overline{f}) \leq \|L\| \cdot \|\overline{f}\|$ we have $\|L\| \leq \|\overline{f}\|$.

Moreover since $J(f_\lambda) \cap \operatorname{span}\{L\} \neq \emptyset$ we have $\|L\| \cdot \|f_\lambda\| = L(f_\lambda) \to \|L\|^2$ and thus $\|f_\lambda\| \to \|L\|$. Since $\|\overline{f}\| \leq \liminf \|f_\lambda\| = \|L\|$ (c.f. Brezis [Bre11] Proposition 3.5 (iii)) we have $\|\overline{f}\| = \|L\|$ and thus $L \in J(\overline{f})$.

Since the f_λ are minimisers of the regularisation problem we have for all $g \in \mathcal{B}$ such that $L(g) = \|L\|^2$

$$\mathcal{E}_z\left(\left(\frac{a_0}{\|L\|^2}L(f_\lambda)\nu_i, y_i\right)_{i\in\mathbb{N}_m}\right) + \lambda\Omega(f_\lambda) \leq \mathcal{E}_z((a_0\nu_i, y_i)_{i\in\mathbb{N}_m}) + \lambda\Omega(g).$$

Since a_0 is the minimiser this implies in particular that

$$\Omega(f_\lambda) \leq \Omega(g) \qquad \forall g \in \mathcal{B} \text{ such that } L(g) = \|L\|^2$$

and taking the limit inferior again we obtain that $\overline{f}$ is in fact a solution of the interpolation problem

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

Now this means that $\Omega(\overline{f} + f_T) \geq \Omega(\overline{f})$ for all $f_T \in \ker(L)$ and if $\overline{f} = f$ we are clearly done.

If $\overline{f} \neq f$ we know that f and $\overline{f}$ are in the same face as $L \in J(f)$ and $L \in J(\overline{f})$. They thus have the same error $\mathcal{E}_z$. If $\Omega(f) = \Omega(\overline{f})$ then both are equivalent minimisers and it is clear that both satisfy the tangential bound. If $\Omega(f) > \Omega(\overline{f})$ it is not admissible and does not need to satisfy the tangential bound.

We now discuss this proof for $\mathcal{B}$ not reflexive. Since $\ker(L)$ is proximinal by Lemma 5.28 we are in the situation of Definition 6.16 (i) and we obtain a sequence of solutions (f_λ) as before. The sequence now only converges weakly* to some $\overline{f}^{**} \in \mathcal{B}^{**}$. But by lower semicontinuity we still have that

$$\mathcal{E}_z\left(\left(\frac{a_0}{\|L\|^2}\overline{f}^{**}(L)\nu_i, y_i\right)_{i\in\mathbb{N}_m}\right) \leq \mathcal{E}_z((a_0\nu_i, y_i)_{i\in\mathbb{N}_m})$$

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 $\hat{f}$ 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 Lemma 5.28 $\ker(L)$ 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 above 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 $\mathcal{E}_z$

is independent of f and for every λ the minimiser f_λ has to be zero to satisfy $J(f_\lambda) \cap \{0\} \neq \varnothing$. This means Ω is minimised at 0.

❑

The reverse direction in fact does not require any assumptions on the error function or regulariser. This means any representer theorem obtained in the earlier sections of this chapter automatically also holds for regularisation problems. We have the following result.

Theorem 6.23

Let $\mathcal{E}, \Omega$ be an arbitrary error functional and regulariser satisfying the general assumption that minimisers always exist. Then the pair $(\mathcal{E}_z, \Omega)$ is admissible for the regularisation problem (58) if Ω is admissible for the regularised interpolation problem (59).

Proof:

Let f_0 be a solution of the regularisation problem (58). 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 $\overline{f_0}$ in the sense of the type of representer theorem valid for $\mathcal{B}$, i.e. in the sense of Definitions 6.1, 6.7, 6.11 and 6.16. 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 (58) in the sense of the representer theorem and the pair $(\mathcal{E}_z, \Omega)$ is admissible.

❑

In conclusion under the assumptions of Theorem 6.22 we have that the pair $(\mathcal{E}_z, \Omega)$ is admissible for the regularisation problem (58) if and only if Ω is admissible for the regularised interpolation problem (59). In any case all of the representer theorems presented in the earlier sections of this chapter immediately apply to regularisation problems as well.

7 Conclusions and Future Work

In this chapter we are going to discuss some consequences of the results presented in Chapter 6 and remaining open questions. We begin the chapter by presenting the last of our main results which is an easy but very important consequence of our results from Chapter 6 in Section 7.1. As it turns out we can show that if one relies on the representer theorem for learning then the solution of the regularised interpolation problem is *independent of the regulariser* but *determined by the function space* alone. This is an important result for two reasons.

Firstly it allows us to use whichever regulariser is most useful for our purpose. We can choose a regulariser we can deal with well numerically for computational purposes and use a different regulariser with useful mathematical properties, such as e.g. the duality of the norm with bounded linear functionals, for proving theoretical results. Changing the regulariser does not alter the set of solutions in the sense of the representer theorem.

Secondly it means that the choice of function space we are learning in is crucial for the solution we obtain. Thus being able to learn in a wider variety of function spaces becomes even more important. This illustrates the importance of results like the theory of RKBS presented in Chapter 4 and our results in Chapter 6 which extend well established and commonly applied methods to Banach spaces.

All results in Chapter 6 were stated for real Hilbert spaces or Banach spaces. We will briefly discuss the complex case in Section 7.2, illustrating why it is not clear if anything like the results in Chapter 6 can be said about complex vector spaces. Finally we will comment on some other open questions and directions for further research.

7.1 The Solution Is Determined by the Space

An interesting and important consequence of the results of Chapter 6 is that if one relies on the representer theorem for learning, then in most cases the solution of the regularised interpolation problem in fact does not depend on the regulariser but is determined by the function space alone. This has two important consequences.

Firstly it means we are free to work with whatever regulariser is most convenient for our purpose, whether this is computational applications or proving theoretical results. Secondly it illustrates the importance of extending well established learning methods for Hilbert spaces to Banach spaces to allow for a greater variety of spaces to learn in.

Throughout this section we say that a function f_0 is a representer theorem solution of (50) in a reflexive Banach space $\mathcal{B}$ if it is a solution of (50) in the sense of Definition 6.11, i.e. such that there exists $\hat{L} = \sum_{i=1}^{m} c_i L_i$ such that $\hat{L} \in J(f_0)$.

To prove the above claim we are going to show that in most cases a function f_0 is a representer theorem solution of (50) if and only if it is a solution of the minimal norm interpolation problem

$$\inf\{\|f\| \,:\, f \in \mathcal{B},\, L_i(f) = y_i \,\forall i \in \mathbb{N}_m\}\,. \tag{61}$$

More precisely we have the following theorem.

Theorem 7.1

Let $\mathcal{B}$ be a reflexive Banach space and Ω admissible. Then any representer theorem solution of (50) is a solution of (61).

Moreover for any solution of (61) there exists a representer theorem solution of (50) in the same face of the norm ball. Thus in particular if $\mathcal{B}$ is strictly convex then f_0 is a representer theorem solution of (50) if and only if it is a solution of (61).

Proof:

Part 1: *(A solution of (50) is a solution of (61))*

Assume that f_0 is a representer theorem solution of (50). Then since by

Lemma 5.9 $\operatorname{span}\{L_i : i \in \mathbb{N}_m\} = (\operatorname{span}\{L_i : i \in \mathbb{N}_m\}_\perp)^\perp$ we have for any $\hat{L} = \sum_{i=1}^{m} c_i L_i \in J(f_0)$ and all $f_T \in \operatorname{span}\{L_i : i \in \mathbb{N}_m\}_\perp$ that

$$\|\hat{L}\|_{\mathcal{B}^*} \cdot \|f_0\|_{\mathcal{B}} = \hat{L}(f_0) = \hat{L}(f_0 + f_T) \leq \|\hat{L}\|_{\mathcal{B}^*} \cdot \|f_0 + f_T\|$$

so $\|f_0\|_{\mathcal{B}} \leq \|f_0 + f_T\|_{\mathcal{B}}$ and f_0 is a solution of (61).

Part 2: *(For any sol. of (61) $\exists$ a sol. of (50) in the same face)*
Assume f_0 is a solution of the minimal norm interpolation problem (61). Then by Theorem 1 in [MP04] there exists an $\hat{L} = \sum_{i=1}^{m} c_i L_i$ such that $\hat{L}(f_0) = \|\hat{L}\|_{\mathcal{B}^*} \cdot \|f_0\|_{\mathcal{B}}$ and thus $\frac{\|f_0\|_{\mathcal{B}}}{\|\hat{L}\|_{\mathcal{B}^*}} \hat{L} \in J(f_0)$.
Further if f_0 is an admissible point in the sense of Definition 6.11, then the tangential bound Lemma 6.12 applies and

$$\Omega(f_0) \leq \Omega(f_0 + f_T) \qquad \forall f_T \in \operatorname{span}\{L_i : i \in \mathbb{N}_m\}_\perp,$$

so f_0 is a representer theorem solution of (50).
If f_0 is not admissible in the sense of Definition 6.11 then there exists an admissible point $\overline{f_0}$ in the same face for which above inequality holds so that $\overline{f_0}$ is a representer theorem solution of (50).

If $\mathcal{B}$ is strictly convex then every point is admissible and f_0 is a representer theorem solution of (50) if and only if it is a solution of (61).

❑

This result shows that for any admissible regulariser on a reflexive, strictly convex Banach space the set of solutions with a dual element in the linear span of the defining linear functionals is identical. 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. For a reflexive Banach space which is not strictly convex the solution is also mostly determined by the space, the regulariser only determines the point(s) within a certain face of the norm ball which is optimal.

It is clear that this result will in fact hold for non-reflexive Banach spaces in the case that $\bigcap_{i\in\mathbb{N}_m}\ker(L_i)$ is proximinal, as in this case Definitions 6.11 and 6.16 as well as Lemmas 6.12 and 6.18 agree. Moreover the minimal norm interpolation problem has a solution in this case and it is known that Theorem 1 from [MP04] holds. It seems plausible to expect that one can prove a similar result for the notion of approximate solutions proposed in Definition 6.16 and approximate minimisers of (61).

More precisely the minimal norm interpolation problem (61) is equivalent to the metric projection of 0 onto the set $\overline{f}+\bigcap_{i\in\mathbb{N}_m}\ker(L_i)$ for any $\overline{f}$ which satisfies the interpolation constraints. We may thus hope to prove a weaker version of Theorem 1 in [MP04] for points which almost attain the distance of 0 to $\overline{f}+\bigcap_{i\in\mathbb{N}_m}\ker(L_i)$. One might expect to be able to characterise such points as points for which there exists a linear functional which attains its norm at the point and which has small norm on $\bigcap_{i\in\mathbb{N}_m}\ker(L_i)$. This would be similar in spirit to the approximate representer theorem in Definition 6.16. Then one may be able to use similar arguments to the ones above to prove a analogous result for non-reflexive Banach spaces.

In Section 6.7 we saw that the regularised interpolation problem is the limit of the regularisation problem as $\lambda\to 0$ and sending λ to zero effectively drives the solution of the regularisation problem towards the solution of the regularised interpolation problem. The choice of λ is commonly a hyperparameter that is determined during training to trade the error made on the training data off against the regularity of the solution. But if all regularisers lead to the same solution for the regularised interpolation problem then for all regularisers the solution of the regularisation problem is also pushed towards the same solution as λ goes to zero. Thus also the regularisation problem is in a sense independent of the regulariser for careful tuning of the regularisation parameter λ.

We are thus free to work with whichever regulariser is most convenient in a particular 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 aforementioned paper [MP04] which heavily relies on a duality between the norm of the space and its continuous linear functionals.

7.2 Complex Vector Spaces

Throughout Chapter 6 we assumed the function space $\mathcal{H}$ or $\mathcal{B}$ to be real. This raises the question whether the results presented apply to complex vector spaces. Everything we were able to say about Ω crucially relied on the observation that being admissible is a statement about its behaviour along tangents as stated in Lemmas 6.2, 6.10, 6.12 and 6.18. But this does not allow to make statements about the behaviour of Ω on a complex vector space as there is in fact no tangent into the complex plane, not even in a Hilbert space. To see this fix $\hat{f} \in \mathcal{H}$ and consider the ray $\{t \cdot e^{i\theta} \cdot \hat{f} \,:\, t \in \mathbb{R}\}$ for some $\theta \in [0, 2\pi)$, as illustrated in Fig. 12.

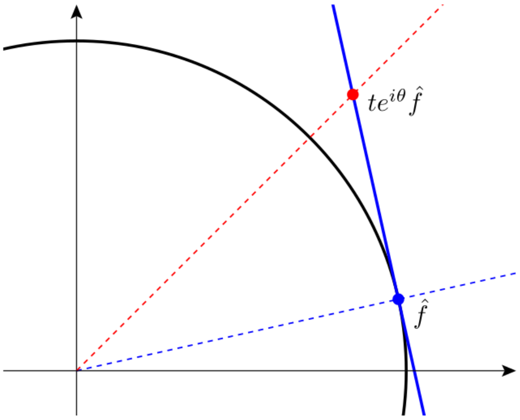


Figure 12: There cannot be a tangent into the complex plane.

Then denoting the line segment from any point along the ray to $\hat{f}$ by

$$g_t = t \cdot e^{i\theta} \cdot \hat{f} - \hat{f} = \left(t \cdot e^{i\theta} - 1\right) \hat{f}$$

this segment is tangent at $\hat{f}$ if and only if $\left\langle g_t, \hat{f} \right\rangle_{\mathcal{H}} = 0$. But for this we find that

$$\begin{aligned} &\left\langle \left(t \cdot e^{i\theta} - 1\right) \hat{f}, \hat{f} \right\rangle_{\mathcal{H}} = 0 \\ \Leftrightarrow\ &\left(t \cdot e^{i\theta} - 1\right) \|\hat{f}\|^2 = 0 \\ \Leftrightarrow\ &(t, \theta) \in \{(1, 0), (-1, \pi)\}. \end{aligned}$$

But for $\theta \in \{0, \pi\}$ the ray is just $\{t \cdot \hat{f} : t \in \mathbb{R}\}$. This means there is no tangent which intersects the ray $\{t \cdot e^{i\theta} \hat{f} : t \in \mathbb{R}\}$ whenever it has been rotated into the complex plane, i.e. the intuition from Fig. 12 is in fact not true. Likewise one can show that it is not possible to reach any point along that ray via an "out and back" argument as we used repeatedly in Chapter 6. For this reason it is currently not clear whether one can say anything about the situation in complex vector spaces.

7.3 Future work

In Section 6.6 we introduced a notion of approximate solutions and hence approximate representer theorems. While we prove the exact representer theorem for as many problems as possible and the approximate version only for problems when there is no chance of obtaining the exact version, it could be interesting to see if the approximate representer theorem can be useful for applications in general. Since computationally we usually will only aim to approximate solutions to a given ε accuracy, the notion of an approximate representer theorem may prove useful in numerical calculations. It would be interesting to see if one can derive new algorithms for solving regularised interpolation or regularisation problems based on an approximate representer theorem.

As mentioned in Section 6.6 the proofs showing that the tangential bound extends all the way around the circle given for strictly convex spaces and uniformly non-rotund spaces suggest that, once a space has been fixed, an argument along similar lines should allow to prove the same result for a space which is neither strictly convex nor uniformly non-rotund. Once a space is fixed it should be possible to always determine a path "out and back" or around the circle since the position of exposed points and lower dimensional faces is now known. If there are spaces which are relevant for applications, which have not been covered by the cases in Chapter 6, it would be worthwhile to extend the results to either those spaces in particular, or possibly even to another class of function spaces containing them. Mathematically it would be interesting to see if one can somehow characterise all possible geometries sufficiently to obtain a closed form result for arbitrary Banach spaces.

In the paper "A Unifying View of Representer Theorems" [AD14] Argyriou and Dinuzzo take a different view on generalising the concept of representer theorems to extend the range of problems the theory of representer theorems applies to. They introduce quasilinear subspace-valued maps which take a point x in a Hilbert space $\mathcal{H}$ and map it to a subspace of $\mathcal{H}$. They then prove necessary and sufficient conditions for the existence of a solution in the sum of subspaces associated with the data points. This raises the question whether our results can be extended to a similar kind of framework. We sketch briefly how this could look like.
Consider a reflexive Banach space $\mathcal{B}$. Thus all linear maps $L \in \mathcal{B}^*$ are represented by a semi-inner product $[\cdot,\cdot]_L$ and a point $x_L \in \mathcal{B}$. Assume the data points are embedded into $\mathcal{B}$, i.e. $\{x_1, \ldots, x_m\} \subset \mathcal{B}$. Consider the regularised interpolation problem

$$\min \{\Omega(f) : f \in \mathcal{B}, [f, x_i]_i = y_i\}.$$

Now rather than defining a map taking subspaces of the original space as values we define a map which maps points in $\mathcal{B}$ to cones in $\mathcal{B}^*$. This reflects that the representer theorem is in its essence a result about the dual space. Due to the lack of linearity of the duality mapping the values now have to be cones rather than subspaces. Since S in this case is a map from $\mathcal{B}$ to its dual space we also need to generalise definition 3.1 from [AD14] accordingly and make it a statement about the dual space. If we let

$$S(x) = \{\lambda \cdot J(x) : \lambda \in \mathbb{R}\}$$

then for a uniform Banach space $\mathcal{B}$, so that $J(x)$ is univocal and injective, definition 3.1 from [AD14] generalises exactly to Definition 6.7. Indeed we obtain that there exists a solution f_0 such that $f_0^* \in \sum_{i=1}^{m} S(x_i) = \sum_{i=1}^{m} c_i x_i^*$.

For a reflexive Banach space definition 3.1 from [AD14] generalises to the existence of f_0 such that $J(f_0) \cap \sum_{i=1}^{m} S(x_i) = \sum_{i=1}^{m} c_i J(x_i) \neq \varnothing$.
The concept of orthomonotonicity, defined in the paper as the essential property for admissibility of regularisers, is exactly tangential nondecreasingness from Chapter 6. In view of the similarities between the results, it seems plausible that one could prove a result in the spirit of theorem 3.1 from [AD14], saying that orthomonotonicity is a necessary and sufficient condition for a representer theorem for cone-valued maps in Banach spaces. This would generalise and unify both, the results from the paper and our results from Chapter 6, in the same way as both works generalise the classical representer theorems.
Moreover the proof of admissibility in the paper [AD14] is directly for regularisation problems rather than regularised interpolation problems. It clearly follows the same ideas from [AMP09] presented in Section 6.7 to prove equivalence of regularisation and regularised interpolation problems. This suggests that the same methods used in the proofs in [AD14] should allow to prove analogous results to the ones presented in Chapter 6 for regularisation problems directly. It would be interesting to see if this could lead to further insights about the theory.

In the paper [MP05b] Micchelli and Pontil discuss learning in Hilbert spaces of vector-valued functions, in particular in RKHS of functions which take values in a real Hilbert space. This suggests to explore whether our results can be extended from bounded linear functionals to bounded linear operators between two real normed spaces. In view of Sain in [Sai18] using the same method to characterise the norm attainment sets of bounded linear functionals on metric spaces and of bounded linear operators between two real normed spaces, one can hope that a similar extension can be done here. The books [HvNVW16, HvNVW17] deal with Banach space-valued maps and could provide a starting point for the necessary theory to develop these kinds of results.

A References

[AD14] Andreas Argyriou and Francesco Dinuzzo. A Unifying View of Representer Theorems. In *Proceedings of the 31st International Conference on Machine Learning*, Volume 32, pages 748–756, 2014.

[AGP05] Antonio Aizpuru and Francisco J García-Pacheco. Some Questions about Rotundity and Renormings in Banach Spaces. *Journal of the Australian Mathematical Society*, 79(01):131–140, 2005.

[AGP08] Antonio Aizpuru and Francisco J García-Pacheco. A Short Note about Exposed Points in Real Banach spaces. *Acta Mathematica Scientia*, 28(4):797 – 800, 2008.

[All11] Graham R. Allan. *Introduction to Banach Spaces and Algebras*, Volume 20 of *Oxford Graduate Texts in Mathematics.* Oxford University Press, 2011.

[AMP09] 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.

[Asp67] Edgar Asplund. Positivity of Duality Mappings. *Bull. Amer. Math. Soc.*, 73(2):200–203, 03 1967.

[BL62] Arne Beurling and A. E. Livingston. A Theorem on Duality Mappings in Banach Spaces. *Arkiv för Matematik*, 4(5):405–411, 1962.

[BL06] Jonathan Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization.* Springer-Verlag New York, Second Edition, 2006.

[Bla82] Jaroslav Blazek. Some Remarks on the Duality Mapping. *Acta Universitatis Carolinae. Mathematica et Physica*, 23(2):15–19, 1982.

[BP61] Errett Bishop and R. R. Phelps. A Proof that Every Banach Space is Subreflexive. *Bull. Amer. Math. Soc.*, 67(1):97–98, 1961.

[Bre11] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* Universitext. Springer-Verlag New York, 2011.

[Bro65a] Felix E. Browder. Multi-Valued Monotone Nonlinear Mappings and Duality Mappings in Banach Spaces. *Transactions of the American Mathematical Society*, 118:338–351, 1965.

[Bro65b] Felix E. Browder. On a Theorem of Beurling and Livingston. *Canadian Journal of Mathematics*, 17:367–372, 1965.

[Bro69] F.E. Browder. Nonlinear Variational Inequalities and Maximal Monotone Mappings in Banach Spaces. *Mathematische Annalen*, 183:213–231, 1969.

[BV10] Jonathan M. Borwein and J.D. Vanderwerff. *Convex Functions: Constructions, Characterizations and Counterexamples.* Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2010.

[CO90] Dennis D. Cox and Finbarr O'Sullivan. Asymptotic Analysis of Penalized Likelihood and Related Estimators. *Ann. Statist.*, 18(4):1676–1695, 1990.

[Con94] J.B. Conway. *A Course in Functional Analysis.* Graduate Texts in Mathematics. Springer New York, 1994.

[CS01] Felipe Cucker and Steve Smale. On the Mathematical Foundations of Learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2001.

[Day55] Mahlon M. Day. Strict Convexity and Smoothness of Normed Spaces. *Transactions of the American Mathematical Society*, 78(2):516–528, 1955.

[Dra04] Sever Silvestru Dragomir. *Semi-inner Products and Applications*. Nova Science Publishers, 2004.

[Eke74] Ivar Ekeland. On the Variational Principle. *Journal of Mathematical Analysis and Applications*, 47:324–353, 1974.

[EPP00] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13(1), 2000.

[Fau77] Gard D. Faulkner. Representation of Linear Functionals in a Banach Space. *Rocky Mountain Journal of Mathematics*, 7(4):789–792, 1977.

[Gil67] J. R. Giles. Classes of Semi-Inner-Product Spaces. *Transactions of the American Mathematical Society*, 129(3):436–446, 1967.

[GSGP14] Pando G. Georgiev, Luis Sánchez-González, and Panos M. Pardalos. *Construction of Pairs of Reproducing Kernel Banach Spaces*, pages 39–57. Springer New York, 2014.

[Hol75] R.B. Holmes. *Geometric Functional Analysis and its Applications*. Graduate Texts in Mathematics. Springer-Verlag, 1975.

[HUL01] J.B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer-Verlag Berlin Heidelberg, 2001.

[HvNVW16] Tuomas Hytönen, Jan van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach Spaces, Volume I: Martingales and Littlewood-Paley Theory*, Volume 63 of *Ergebnisse der Mathematik und ihrer Grenzgebiete.* Springer International Publishing, 2016.

[HvNVW17] Tuomas Hytönen, Jan van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach Spaces, Volume II: Probabilistic Methods and Operator Theory*, Volume 67 of *Ergebnisse der Mathematik und ihrer Grenzgebiete.* Springer International Publishing, 2017.

[Jam47] Robert C. James. Orthogonality and Linear Functionals in Normed Linear Spaces. *Transactions of the American Mathematical Society*, 61(2):265–292, 1947.

[Jam64] Robert C. James. Weak Compactness and Reflexivity. *Israel Journal of Mathematics*, 2:101–119, 1964.

[KLMW18] Vladimir Kadets, Ginés López, Miguel Martín, and Dirk Werner. Equivalent Norms with an Extremely Nonlineable Set of Norm Attaining Functionals. *Journal of the Institute of Mathematics of Jussieu*, page 1–21, 2018.

[Köt83] Gottfried Köthe. *Topological Vectorspaces I*, Volume 159 of *Grundlehren der Mathematischen Wissenschaften.* Springer-Verlag Berlin Heidelberg, 1983.

[KW71] George Kimeldorf and Grace Wahba. Some Results on Tchebycheffian Spline Functions. *Journal of Mathematical Analysis and Applications*, 33(1):82 – 95, 1971.

[LPT12] Joram Lindenstrauss, David Preiss, and Jaroslav Tišer. *Frechet Differentiability of Lipschitz Functions and Porous Sets in Banach Spaces.* Annals of Mathematics Studies. Princeton University Press, 2012.

[LT79] Joram Lindenstrauss and Lior Tzafriri. *Classical Banach Spaces II: Function Spaces*, Volume 97 of *Ergebnisse der Mathematik und ihrer Grenzgebiete.* Springer-Verlag Berlin Heidelberg, 1979.

[Lum61] G. Lumer. Semi-Inner-Product Spaces. *Transactions of the American Mathematical Society*, 100(1):29–43, 1961.

[Meg12] R.E. Megginson. *An Introduction to Banach Space Theory.* Graduate Texts in Mathematics. Springer-Verlag New York, 2012.

[MP04] 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.

[MP05a] Charles A. Micchelli and Massimiliano Pontil. Learning the Kernel Function via Regularization. *Journal of Machine Learning Research*, 6:1099–1125, Jul 2005.

[MP05b] Charles A. Micchelli and Massimiliano Pontil. On Learning Vector-Valued Functions. *Neural Computation*, 17(1):177–204, 2005.

[Phe93] R.R. Phelps. *Convex Functions, Monotone Operators and Differentiability.* Lecture Notes in Mathematics. Springer Berlin Heidelberg, 1993.

[Rea18] Charles J Read. Banach Spaces with no Proximinal Subspaces of Codimension 2. *Israel Journal of Mathematics*, 223(1):493–504, 2018.

[Rud91] Walter Rudin. *Functional Analysis.* International Series in Pure and Applied Mathematics. McGraw-Hill, 1991.

[Sai05] Fathi B. Saidi. On the Proximinality of the Unit Ball of Proximinal Subspaces in Banach Spaces: A Counterexample. *Proceedings of the American Mathematical Society*, 133(9):2697–2703, 2005.

[Sai18] Debmalya Sain. On the Norm Attainment Set of a Bounded Linear Operator and Semi-Inner-Products in Normed Spaces. *arXiv*, 1802.10439v2, 2018.

[Sch18] Kevin Schlegel. When is there a Representer Theorem? Nondifferentiable Regularisers and Banach spaces. *arXiv*, 1804.09605, April 2018.

[Sch19a] Kevin Schlegel. Approximate Representer Theorems in Non-reflexive Banach spaces. *arXiv*, 1911.00433, November 2019.

[Sch19b] Kevin Schlegel. When is there a representer theorem? Nondifferentiable regularisers and Banach spaces. *Journal of Global Optimization*, April 2019.

[Sch19c] Kevin Schlegel. When is there a Representer Theorem? Reflexive Banach spaces. *arXiv*, 1809.10284v2, May 2019.

[Sch20] Kevin Schlegel. Approximate Representer Theorems in Non-reflexive Banach Spaces. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, Volume 117 of *Proceedings of Machine Learning Research*, pages 827–844. PMLR, 08 Feb–11 Feb 2020.

[Sch21] Kevin Schlegel. When is there a representer theorem? Reflexive Banach Spaces. *Advances in Computational Mathematics*, 47, July 2021.

[SHS01] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A Generalized Representer Theorem. In *Computational Learning Theory*, pages 416–426. Springer Berlin Heidelberg, 2001.

[Sim08] S. Simons. *From Hahn-Banach to Monotonicity.* Lecture Notes in Mathematics. Springer Netherlands, 2008.

[Sin70] Ivan Singer. *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces.* Grundlehren der Mathematischen Wissenschaften. Springer Berlin Heidelberg, 1970.

[SS98] J. A. Smola and B. Schölkopf. On a Kernel-Based Method for Pattern Recognition, Regression, Approximation, and Operator Inversion. *Algorithmica*, 22(1):211–231, 1998.

[SS02] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels.* MIT Press, 2002.

[STC04] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

[SZH13] Guohui Song, Haizhang Zhang, and Fred J. Hickernell. Reproducing Kernel Banach Spaces with the l^1 Norm. *Applied and Computational Harmonic Analysis*, 34(1):96 – 116, 2013.

[Tho96] A. C. Thompson. *Minkowski Geometry.* Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1996.

[Wen04] Holger Wendland. *Scattered Data Approximation.* Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

[ZXZ09] Haizhang Zhang, Yuesheng Xu, and Jun Zhang. Reproducing Kernel Banach Spaces for Machine Learning. *Journal of Machine Learning Research*, 10:2741–2775, 2009.

[ZZ12] 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.