# Neural Networks as Functions Parameterized by Measures
## Representer Theorems and Approximation Benefits

Sanders, Koen

*Award date:*
2020


# Neural Networks as Functions Parameterized by Measures: Representer Theorems and Approximation Benefits

Master Thesis


| | |
|---|---|
| Master: | Industrial & Applied Mathematics |
| Department: | Mathematics & Computer Science |
| Student: | Koen Sanders |
| Identity number: | 0818388 |
| Supervisors: | Professor Lorenzo Rosasco |
| | Dr. Ir. Rui M. Castro |


Eindhoven, October 16, 2020

# Acknowledgement

This thesis was preceded by an internship at the Laboratory for Computational and Statistical Learning (LCSL) in the Machine Learning Centre Genua (MaLGa), located at the University of Genua (UniGe). I would like to thank everyone from MaLGa for providing a welcome and educational environment. I learned a lot this first period, and enjoyed myself. Also the first few months of my master project I spend at MaLGa, it was a great experience to see how daily life is in a research group. This has inspired me to continue in academics and in time pursue a PhD.

Due to the COVID-19 pandemic I had to continue the last few months from the Netherlands. Even though all the communication was online, I could still participate in the lab, received weekly guidance and could always ask questions if I needed to. I want to thank everyone at MaLGa, for giving me this opportunity.

Especially I would like to thank my first supervisor, Lorenzo Rosasco, for his personal guidance and help through each stage of the process. Also my second supervisor Rui M. Castro for the helpful remarks and support. I also want to express my gratitude towards Jaouad Mourtada and Ernesto de Vito, a lot of this work was done in collaboration with them.

**Abstract**

Two-layer neural networks can be represented as functions parameterizated by a measure, first mentioned in [Bar93] and elaborated on in [Bac14]. In this thesis we formally state the functional analysis framework to substantiate this statement. The main benefit of the function space of two-layer neural networks represented in this way, is that it becomes a linear space and thus allows for the construction of different mathematical structures such as norms and inner products. This mathematical structure, that is lacking in the original function space of two-layer neural networks, is exploited to investigate the connection to Reproducing Kernel Hilbert and Banach Spaces for which the corresponding reproducing kernels and representer theorems are derived. These suggest that the heavy overparameterization used in practise for neural networks, is not necessary to represent the optimal model when training with a finite number of training data points. Also, recent approximation results for two-layer neural networks in this representation are surveyed. The emphasis is put on approximation results of functions with some notion of sparsity. When approximating such functions neural network are proven to outperform, and thus to be different from, kernel methods.

# Contents

# 1 Introduction

The general topic of this thesis is the theory behind the success of machine learning. Machine learning is an application of Artificial Intelligence, and defined to be the study of algorithms that improve automatically through experience ([Mit97]). Lately there has been a resurgence of interest in machine learning with the successes booked by deep learning, mostly fueled by the increase of computing power and available data. Deep learning is characterized by the use of artificial neural networks. Examples of impressive results that have been achieved are in the fields of computer vision ([KSH12a]), speech recognition ([GMH13]) and game playing ([Sil+16]).
In this thesis we focus on the machine learning-task of supervised learning. In supervised learning the goal is to learn a function that maps an input to an output based on training pairs of input and output. How well a function maps the input to the output is evaluated through a functional $L$ returning the error, the objective is to find a function $f$ for which $L(f)$ is as low as possible. The performance is measured in terms of the generalization error, how accurate the trained model predicts on test-data it has not seen before. Classically, the generalization error can be decomposed into three parts. Consider $f_0$ the best possible function that exists to map the input to the output, and $L(f_0)$ the corresponding lowest possible error. Similarly $L(f_{\mathcal{H}})$ is the lowest error you can achieve when approximating only with functions $f$ from a function space $\mathcal{H}$. $L(\widehat{f}_{\text{opt}})$ is the best error obtained by exactly solving a certain optimization problem, such as Empirical Risk Minimization, and $L(\widehat{f})$ is the function obtained by a practical optimization routine.

$$\underbrace{L(\widehat{f})}_{\text{Generalization Error}} = \underbrace{(L(\widehat{f}) - L(\widehat{f}_{\text{opt}}))}_{\text{Optimization error}} + \underbrace{(L(\widehat{f}_{\text{opt}}) - L(f_{\mathcal{H}}))}_{\text{Estimation error}} + \underbrace{(L(f_{\mathcal{H}}) - L(f_0))}_{\text{Approximation Error}}$$

The approximation error is the accuracy you lose by not considering all possible functions, it is dependent on the choice for the space of approximating functions $\mathcal{H}$. The estimation error is the accuracy you lose by only having a limited amount of data-points to train your model. The optimization error is the accuracy you lose while looking for the optimal function, it is dependent on the choice of optimization method.
The space of functions that received the most attention and gained the most success in supervised learning problems is the space of artificial neural networks. They have been the state of the art model for the benchmark supervised learning data-tasks, such as classifying ImageNet, for the last few years ([KSH12b]-[Tou+20]). However when analyzing these models, a lot of classical machine learning theory such as the Bias-Variance Trade-off ([Bel+18]), does not apply and needs to be rethought. There are numerous open questions regarding the lack of theory for understanding the generalization success of neural network. The inductive bias is the set of assumptions made by the model used to predict outputs for test data, it is the bias towards certain (type of) functions. Finding the optimal weights in a neural network is a strongly non convex optimization problem, it is unclear how optimizing with simple gradient methods converges to a minimum with the right inductive bias ([Gun+18],[LZB20]). Another open question is in approximation theory, which classes of functions can be approximated well by neural networks ([MP16],[Sch17])? Also, the role of overfitting and regularization is not well understood ([Zha+16],[Bel+18]). So there is a clear gap between the practical success and theoretical explanation of the generalization error of neural networks.
Normally in machine learning, the space of approximating functions is well understood. It has a linear structure and natural mathematical structures, such as a norm or an inner product. One of the main obstructions to bridge the gap between theory and practise for neural networks is that, so far, there is little understanding of the mathematical structure of the function space of neural networks. The function space is non-linear and does not possess any natural mathematical structures. In this thesis we discuss a way to overcome this problem for shallow networks, especially two-layer neural networks. With $m$ neurons in the hidden layer, two-layer neural networks are formally defined as:

$$f_m(\mathbf{x}) = \sum_{i=1}^{m} a_i \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i), \tag{1}$$

where $a_i \in \mathbb{R}$, $\mathbf{x} \in \mathcal{X}$, $(\mathbf{w}_i, b_i) \in \Theta$ for $i = 1, ..., m$, with $\mathcal{X} \subset \mathbb{R}^d$ the input domain and $\Theta \subset \mathbb{R}^{d+1}$ the parameter space, and the activation function $\sigma : \mathbb{R} \to \mathbb{R}$. The space of functions of the form (1) is non-linear. A visualization of a two-layer neural network with $m$ neurons is displayed in figure 1.

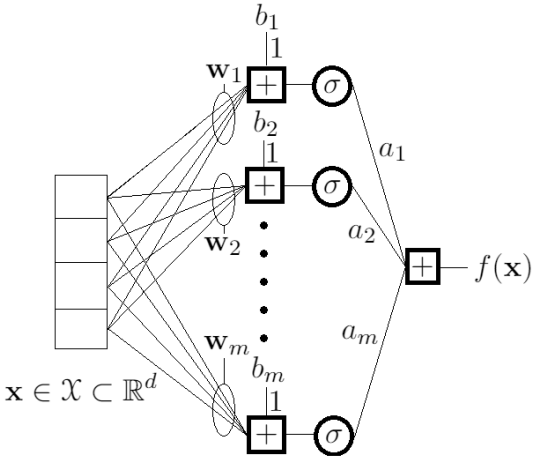


Figure 1: Two-layer Neural Network with $m$ neurons in the hidden layer

Letting $(\mathbf{w}_i, b_i) = \theta_i$ and $\sigma(\mathbf{w}_i^\top \mathbf{x} + b_i) = \varphi(\mathbf{x}, \theta_i)$ the neural network can equivalently be denoted as $f(\mathbf{x}) = \sum_{i=1}^m a_i \varphi(\mathbf{x}, \theta_i)$. The key observation by [Bar93] and [Bac14], is that by taking an atomic measure $\mu = \sum_{i=1}^m a_i \delta_{\theta_i}$, the function of a two-layer neural network can be rewritten as:

$$f(\mathbf{x}) = \int_\Theta \varphi(\mathbf{x}, \theta) \mu(d\theta). \tag{2}$$

The general idea is that when considering not only atomic measures, but all possible measures on the parameter space $\Theta$, the space of functions of the form (2) becomes a linear function space. From this space various interesting linear sub-spaces of two-layer neural networks can be defined, each with their own norm and some having an inner product as well. For instance the particular case of functions of the form (2), parameterized by the set of absolutely continuous measures with full support, corresponds to considering the space of possibly infinitely wide two-layer neural networks. Two main contributions are made, with the use of these mathematical structures. Firstly, representer theorems are derived for several of the above mentioned sub-spaces of two-layer neural networks. A representer theorem states that the optimal solution of regularized risk minimization, over a finite set of training data, can be written as an expansion in terms of the training data. This give the advantage that the problem of finding the optimal solution can be reduced from a problem in an infinite dimensional space to a finite dimensional problem. The most significant representer theorem derived is for possibly infinitely wide two-layer neural networks. From this we find that, when trained with a finite number $n$ of training samples, the optimal neural network uses at most $n$ neurons. This is in great contradiction with recent empirical findings on overparametrization ([Zha+16]), which suggest that adding more parameters increases the generalization performance.
Secondly, known approximation results for functions parametrized by measures are reviewed ([Bar93], [Pin99],[Bac14]). The main contribution here is to prove approximation benefits for neural networks in comparison to alternatives such as kernels methods, when approximating functions with some notion of sparsity. Such sparse functions, for instance functions of projections, are only dependent on a subset of variables. The rate at which you can approximate these functions with kernel methods, scales exponentially with the dimension. This exponential dependence on the dimension is known as the 'Curse of dimensionality'. For neural networks however, we prove adaptivity to the underlying structure of such sparse functions, resulting in a rate that is only dependent on the dimension of the subset of variables. So in this particular case, the curse of dimensionality is broken. The other thing we can take from this, is that neural networks are significantly different from kernel methods.

## 2 Overview of Results and Discussion

In this section further discussion of the results mentioned in the introduction is given. This discussion is divided up into three parts, first treating the results concerning the representer theorems and the approximation properties. Lastly we give further perspectives and some open questions arising from these results.
The starting point is the non-linear function space of two-layer neural networks with $m$ neurons in the hidden layer. We repeat the key observation by [Bar93] and [Bac14], namely that a two layer neural network with $m$ neurons can be rewritten as a function parameterized by an atomic measure $\mu = \sum_{i=1}^{m} a_i \delta_{\theta_i}$:

$$f_m(\mathbf{x}) = \sum_{i=1}^{m} a_i \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i) = \sum_{i=1}^{m} a_i \varphi(\mathbf{x}, \theta_i) = \int_\Theta \varphi(\mathbf{x}, \theta) \mu(d\theta),$$

where $\mathbf{x} \in \mathcal{X}$, $(\mathbf{w}_i, b_i) = \theta_i \subset \Theta$ and $\sigma(\mathbf{w}_i^\top \mathbf{x} + b_i) = \varphi(\mathbf{x}, \theta_i)$. Examples of commonly used activation functions $\sigma : \mathbb{R} \to \mathbb{R}$ are the ReLU $\sigma(u) = (u)_+$ and the sigmoid $\sigma(u) = 1/(1 + e^{-u})$. By not only considering atomic measures, but the space of all possible measures over the parameter space $\Theta$ we can define the function space:

$$\mathbf{B} = \left\{ f \in \mathbb{R}^{\mathcal{X}} \;\middle|\; f(\mathbf{x}) = \int_\Theta \varphi(\mathbf{x}, \theta) \mu(d\theta), \text{ where } \mu \text{ is an arbitrary measure over } \Theta \right\}. \tag{3}$$

This function space $\mathbf{B}$ is linear, that is for all $f, g \in \mathcal{F}$ and all $C \in \mathbb{R}$ holds:

- $(f + g) \in \mathbf{B}$
- $C \cdot f \in \mathbf{B}$

On this linear space certain mathematical structures can be put. An inner product and a norm induce respectively a Hilbert and a Banach space. A norm can be put on this space of functions in (3) by considering the variation norm of the measures, we proof $\mathbf{B}$ to be a Banach space in section 3.2.3.
We specify the discussion by considering a subspace of $\mathbf{B}$, of functions parameterized by only measures that are absolutely continuous with respect to some reference measure $\tau$. In this case we can further rewrite the two-layer neural network as:

$$f(\mathbf{x}) = \int_\Theta \varphi(\mathbf{x}, \theta) \mu(d\theta) = \int_\Theta \varphi(\mathbf{x}, \theta) p(\theta) \tau(d\theta). \tag{4}$$

Here the density $p$ is given by the Radon-Nikodym derivative $p(\theta) = \mu(d\theta)/\tau(d\theta)$. We consider a reference measure $\tau$ that is a probability measure with full support on $\Theta$. For example if $\Theta = \mathbb{R}^d$, the reference measure $\tau$ could be the $d$-dimensional Lebesgue measure. Considering a reference measure with full support on $\Theta$ corresponds to introducing a continuous limit, which can be seen as considering infinitely wide two-layer neural networks. Structure can be put on this space of functions as in (4), by either taking density functions $p$ in $L_1(\Theta, \tau)$ or in $L_2(\Theta, \tau)$. Restricting $p \in L_1(\Theta, \tau)$ induces a Banach space $\mathcal{F}_1$ and $p \in L_2(\Theta, \tau)$ a Hilbert space $\mathcal{F}_2$. $\mathcal{F}_2$ is a, significantly smaller, subspace of $\mathcal{F}_1$.

**Representer Theorems** Consider training data, given by the pairs $(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n) \in \mathcal{X} \times \mathcal{Y}$ and a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$. For a function space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, with corresponding norm $||\cdot||_{\mathcal{H}}$ a representer theorem exists if the optimal solution of the risk minimization, for every $\lambda > 0$:

$$f^* = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{y}_i, f(\mathbf{x}_i)) + \lambda ||f||_{\mathcal{H}}^2 \right\}, \tag{5}$$

can be written as a linear combination of at most $n$ of some sort of representers of the training data. It is known from [Bac14] that the Hilbert space $\mathcal{F}_2$ is a Reproducing Kernel Hilbert Space

(RKHS). For a RKHS a representer theorem is well established ([Sch01]), the optimal solution of the risk minimization in (5) is a linear combination of the reproducing kernel functions, evaluated at the training points. So for $\mathcal{F}_2$, the optimal solution is of the form:

$$f^*_{\mathcal{F}_2}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K_{\mathcal{F}_2}(\mathbf{x}, \mathbf{x}_i) = \sum_{i=1}^{n} \alpha_i \int_{\Theta} \varphi(\mathbf{x}, \theta)\varphi(\mathbf{x}_i, \theta)\tau(d\theta).$$

This reproducing kernel with integral representation can be approximated by random sampling methods [RR08]. So finding the optimal solution amounts to finding the optimal coefficients $\alpha_i$ for $1 \leq i \leq n$.

We prove that the Banach space $\mathcal{F}_1$ is a Reproducing Kernel Banach Space (RKBS), see lemma 3.7, and the corresponding reproducing kernel is derived in section 3.4. In a recent line of work ([Sch18],[Uns19],[LZZ19],[PN20]) the existence of representer theorems in more general spaces than only in RKHS's, such as RKBS's, is researched. Without getting into much detail, a more general notion states that the dual element of the optimal function is in the linear span of the dual elements of the training data. In a RKHS the dual element is the element itself, in a RKBS a second space and a more complicated construction is necessary to define the representer theorem. In this line of work the connection with two-layer neural network was not made explicitly before. The representer theorem for $\mathcal{F}_1$, see theorem 5.4, has been derived using Carathéodory's theorem for the conical hull ([Roc97]). For $\mathcal{F}_1$, the optimal solution is of the form:

$$f^*_{\mathcal{F}_1}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \varphi(\mathbf{x}, \theta_i) = \sum_{i=1}^{n} \alpha_i \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i).$$

Note that the points $\theta_i$ for $1 \leq i \leq n$ in the parameter space $\Theta$ are unknown. This representer theorem for $\mathcal{F}_1$ only gives insight on the form of the optimal solution, not the identity. It sheds light on the inductive bias of neural networks, note that the optimal solution is a two-layer neural network with $n$ neurons. So the important point here is that, even when we consider possibly infinitely wide neural networks, to capture the optimal function at most $n$ neurons are needed. This is not an obvious result, it contradicts recent empirical results on overparametrization ([Zha+16]), where adding more neurons is shown to keep improving the generalization performance. A possible explanation on the benefits of overparameterization, is that this optimal neural network with at most $n$ neurons is hard to find by the local search optimization methods used. Possibly, increasing the number of parameters, would make the loss landscape more suitable to end up in a local optimum close to the global one. This is a hypothesis that would need further research.

**Approximation Results** The approximation error $\varepsilon$ for approximating the function space $\mathcal{G} \subset \mathbb{R}^{\mathcal{X}}$ with another function space $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ is defined:

$$\forall g \in \mathcal{G}: \quad \exists f \in \mathcal{F} \text{ s.t. } \sup_{\mathbf{x} \in \mathcal{X}} \left\{ \left| g(\mathbf{x}) - f(\mathbf{x}) \right| \right\} \leq \varepsilon.$$

When approximating $d$-dimensional smooth functions up to a $\varepsilon$-accuracy with two-layer neural networks, at least $\Omega(\varepsilon^{-d})$ training samples are needed ([Bar93],[Pin99]). This exponential dependence on the dimension is the curse of dimensionality, it is similar for other approximation methods such as splines and kernel methods. The emphasis here is put on approximating functions with some notion of sparsity. As a particular example, we consider functions that only depend on projections on a $s < d$ dimensional subspace, of the form:

$$h(\mathbf{x}) = g(\mathbf{W}^\top \mathbf{x}) \quad \text{with: } \mathbf{W} \in \mathbb{R}^{d \times s} \quad \text{and } g : \mathbb{R}^s \to \mathbb{R} \text{ a smooth function.}$$

We show that when approximating with $\mathbf{B}$, one only needs $\Omega(\varepsilon^{-s})$ training samples to obtain a $\varepsilon$-accuracy. This also holds for the RKBS $\mathcal{F}_1$ but not for RKHS $\mathcal{F}_2$. These results were implicitly mentioned in [Bar93] and [Bac14] but we explicitly proof them in respectively lemma's 6.7 and 6.8. The curse of dimensionality is broken because of the adaptivity neural networks show to these underlying structures. This partially answers the open question of approximation theory of which classes can be approximated well by neural networks. Alternative approximation methods, such as kernel methods, do not have this property and would still need $\Omega(\varepsilon^{-d})$ samples.

## 2.1 Further Perspective & Open Questions

Here we give some further perspective on the above results. Combining both main results also sheds light on another recent line of work ([BMM18],[AL19],[DL19]), where the similarities and differences of neural networks and kernels are researched. With the introduction of the Neural Tangent Kernel ([JGH18]) was initially even suggested that neural networks completely behave the same way as kernels, and thus that the function space of neural networks is always a Reproducing Kernel Hilbert Space. Here we show that a completely different structure can be put on the function space of neural networks by changing the norm, to obtain a Reproducing Kernel Banach Space. This RKBS shows adaptivity to underlying structures while the RKHS does not, so dependent on the norm you put on the space the approximation properties are completely different. Thus there are significant differences in structure and approximation performance between neural networks and kernels.
Furthermore we thought about expanding the approximation result on the adaptivity to underlying structures. In [Sch17] is suggested that in a similar way multi-layer neural networks show adaptivity to composite functions with multiple levels. If our approximation analysis, for neural networks rewritten as a function parameterized by a measure, can be extended to multi-layer neural network is an open question.
In this thesis the focus is mainly on representation and approximation results. Also the optimization part would be interesting to research for this representation of two-layer neural networks, because we can study convex optimization schemes in measure spaces. In [Bac14] an attempt was done to optimize convex functions in the infinite dimensional space $\mathcal{F}_1$, with the use of the Frank-Wolfe algorithm ([FW56]), but this was found to be NP-hard. A possible interesting direction would be to find the optimal measure by regularizing through the relative entropy of this measure with respect to another fixed measure. With this possible computational and statistical advantages, and more insight in the behaviour of neural networks during optimization, could be obtained.

# 3 Spaces of functions parameterized by measures

In this technical section we formally introduce the general framework of spaces of functions parameterized by measures and discuss the mathematical structures that can be put on these spaces. We start with some preliminaries on the fields of functional analysis and measure theory in section 3.1. Secondly the construction of the most general linear space of functions parametrized over all possible measures is presented in section 3.2. This is done by first presenting 3 function spaces that are necessary to define the general framework in section 3.2.1. Subsequently we will introduce a feature map and a linear operator in section 3.2.2 that are key for the construction of the space in section 3.2.3. Then we specify the discussion to fixing a certain reference measure in section 3.3, we consider function spaces parameterized by absolutely continuous measures with respect to this reference measure. Finally we present the connection to Reproducing Kernel Banach and Hilbert spaces in section 3.4. For the RKBS and RKHS's defined, we derive the reproducing kernels. To emphasize, the link with two-layer neural network is not made until the subsequent section 4. We keep the input domain $\mathcal{X}$, the parameter space $\Theta$, the reference measure $\tau$ on $\Theta$ and the basis function $\varphi$ unspecified.

## 3.1 Preliminaries

We start by introducing certain concepts to the reader from the fields of functional analysis and measure theory. In this thesis function spaces are considered, which are either linear or non-linear. A function space $\mathcal{F} \subset K^{\mathcal{X}}$ is a space of functions from a set $\mathcal{X}$, interpreted as the input domain, to a field of numbers $K$, for example the real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$. It is linear if the properties of addition and scalar multiplication holds. That is, if for all functions $f, g \in \mathcal{F}$ and all numbers $\alpha \in K$ holds:

- $(f + g) \in \mathcal{F}$
- $\alpha \cdot f \in \mathcal{F}$.

Certain mathematical structures can be put on these linear function spaces. Examples of such structures are either a norm or a inner product. The norm of a function space $\mathcal{F}$ is a nonnegative function $||\cdot|| : \mathcal{F} \to K$ such that $\forall f, g \in \mathcal{F}$ and $\forall \alpha \in K$:

- $||f + g|| \leq ||f|| + ||g||$
- $||\alpha f|| = |\alpha| ||f||$.
- $||f|| \geq 0$ and $||f|| = 0$ if and only if $f = 0$

The inner product is a bilinear form $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to K$ such that $\forall f, g, h \in \mathcal{F}$ and $\forall \alpha \in K$:

- $\langle f + g, h \rangle = \langle f, h \rangle + \langle g, h \rangle$
- $\langle \alpha f, g \rangle = \alpha \langle f, g \rangle$
- $\langle f, f \rangle \geq 0$ and 0 if and only if $f = 0$
- $\langle f, g \rangle = \overline{\langle g, f \rangle}$.

A function space with a norm is complete, if every Cauchy sequence converges with respect to this norm to an element in the space itself. A Banach space $\mathcal{B}$ is a complete linear space with a norm $||\cdot||_{\mathcal{B}}$. An example of a Banach space is $L_1([a, b])$, the space of all real-valued functions whose absolute value is integrable in the interval $[a, b]$. $L_1([a, b])$ is complete with respect to the $L_1([a, b])$-norm given by $||f||_{L_1([a,b])} = \int_a^b |f(x)| dx$. A Hilbert space $\mathcal{H}$ is a complete linear space that, apart from a norm, also has an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. An example of a Hilbert space is $L_2([a, b])$, the space of all real-valued functions that are square integrable in the interval $[a, b]$. The inner product of $L_2([a, b])$ is given by $\langle f, g \rangle_{L_2([a,b])} = \int_a^b f(x)g(x)dx$ and its norm $||f||_{L_2([a,b])} = \sqrt{\langle f, f \rangle_{L_2([a,b])}}$.
We investigate functions parameterized by measures, a brief introduction in measure theory. Consider a set $X$ with a $\sigma$-algebra $\Sigma(X)$. The tuple $(X, \Sigma(X))$ is a measurable space, and every subset $E \in \Sigma(X)$ is a measurable set. A measure on $(X, \Sigma(X))$ is a function $\mu : \Sigma(X) \to K$ satisfying:

- Null empty set: $\mu(\emptyset) = 0$, where $\emptyset$ denotes an empty set
- Countable additivity: $\mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$ holds for all countable collections of pairwise disjoint sets $\{E_i\}_{i=1}^{\infty}$ in $\Sigma(X)$.

A measure is real-valued if the field is $\mathbb{R}$ and complex-valued if the field is $\mathbb{C}$. The triple $(X, \Sigma(X), \mu)$ forms a measure space. A function $f : X \to K$ is measurable in $(X, \Sigma(X), \mu)$ when it is integrable with respect to $\mu$, i.e. if $\int_X f(x)\mu(dx)$ is finite. The most common example is the $d$-dimensional Lebesgue measure $\lambda$, assigning a positive number from the field $\mathbb{R}^+$ to subsets of the Euclidean space $\mathbb{R}^d$. For $d = 1$, $\mathbb{R}$ is simply a line and every measurable subset of it is a segment or a union of segments on the line, e.g. the segment $[a, b]$ from point $a$ to point $b$. The Lebesgue measure assigns the length of the segment to it, so $\lambda([a, b]) = b - a$. For the case $d = 1$, the Lebesgue measure is often just denoted as $dx$. Lebesgue measurable functions on the line $f : \mathbb{R} \to \mathbb{R}$ are functions that can be integrated on the real line, i.e. functions for which $\int_{\mathbb{R}} f(x)dx$ is finite. Examples of such functions are continuous functions $\mathcal{C}(\mathbb{R})$.
There are lots of different types of measures, we consider here in particular discrete and absolutely continuous measures. Given a measurable space $(X, \Sigma(X))$ and two measures $\mu$ and $\tau$ on this space, the measure $\mu$ is discrete with respect to $\tau$ if there exists an either finite or countably infinite subset $S \subset X$ such that:

- $\tau(E) = 0 \quad \forall E \in \Sigma(S)$
- $\mu(E) = 0 \quad \forall E \in \Sigma(X \setminus S)$
- All sets existing out of a single element $s \in S$ are measurable.

An example of a discrete measure, on a Lebesgue measurable space $(X, \Sigma(X))$, is a sum of Dirac Measures $\mu = \sum_{i=1}^{n} a_i \delta_{s_i}$. For all elements $\{s_1, ..., s_n\} \in S$, $\mu$ assigns a nonzero value and 0 is assigned to all elements in $X \setminus S$.
The measure $\mu$ is absolutely continuous with respect to $\tau$, if $\mu$ is dominated by $\tau$, i.e.:

- $\tau(E) = 0 \implies \mu(E) = 0 \quad \forall E \in \Sigma(X)$
- $\mu(E) = \int_E p(x)\tau(dx) \quad \forall E \in \Sigma(X)$
- Support of $\mu$ is equal to the support $\tau$.

Where the density $p$ is given by the Radon-Nikodym derivative $\mu(dx)/\tau(dx)$. An example is the probability density function of a continuous random variable, where the reference measure is given by the 1-dimensional Lebesgue measure $dx$. The probability the random variable is between the segment $a$ and $b$ is given by $\mathbb{P}(a \leq X \leq b) = \mu([a,b]) = \int_a^b f(x)dx$.

## 3.2 Linear spaces of functions parameterized by measures

In this subsection the most general linear space of functions parameterized by all possible measures is constructed, the construction is divided into 3 parts. First we present the background material, introducing 3 function spaces that are necessary for the construction. Secondly we introduce a feature map and a linear operator, which are key elements in the construction. Lastly we construct the function space itself.

### 3.2.1 Background

To improve readability the definitions of relevant concepts in this section have been moved to appendix A, use the pointers or the hyperlinks to be directed. Let $\mathcal{X}$ be a set, and $\Theta$ a locally compact second countable space (7.2) with a Borel sigma algebra (7.4) $\Sigma(\Theta)$ . The set $\mathcal{X}$ is an input domain for functions $f : \mathcal{X} \to \mathbb{C}$ and $\Theta$ a parameter space used to parameterize these functions. We are going to introduce several function spaces to be able to define these functions.

- $C_0(\Theta, \mathbb{C})$ is the space of all continuous functions $g : \Theta \to \mathbb{C}$ that vanish at infinity, i.e. for each $\varepsilon > 0$ there exists a compact subset $K_\varepsilon \subset \Theta$ such that $\sup_{\theta \in (\Theta \setminus K_\varepsilon)} |g(\theta)| < \varepsilon$.

$C_0(\Theta, \mathbb{C})$ is a Banach space (7.22) with respect to the infinity norm $||g||_{C_0(\Theta,\mathbb{C})} = ||g||_\infty = \sup_{\theta \in \Theta} |g(\theta)|$.

- $(C_0(\Theta, \mathbb{C}))^*$ is the topological dual (7.25) of $C_0(\Theta, \mathbb{C})$, i.e. this is the space of all continuous linear maps $\omega : C_0(\Theta, \mathbb{C}) \to \mathbb{C}$,

The bilinear form between $C_0(\Theta, \tau)$ and its topological dual $(C_0(\Theta, \tau))^*$ is the canonical pairing (7.27) $\langle \cdot, \cdot \rangle$ [1] : $(C_0(\Theta, \mathbb{C}))^* \times C_0(\Theta, \mathbb{C}) \to \mathbb{C}$. Here this canonical pairing is the action of a continuous linear map $\omega \in (C_0(\Theta, \mathbb{C}))^*$ on a continuous function $g \in C_0(\Theta, \mathbb{C})$, and equal to $\omega(g)$. $(C_0(\Theta, \mathbb{C}))^*$ is a Banach space with respect to the norm $||\omega||_{(C_0(\Theta,\mathbb{C}))^*} = \sup_{||g||_\infty \leq 1} | \langle \omega, g \rangle |$. We can show the latter space is equivalent to a suitable space of measures.

- $\mathcal{M}(\Theta)$ is the space of all complex-valued regular Borel measures (7.10) on $\Theta$.

$\mathcal{M}(\Theta)$ is a Banach Space with respect to the variation norm (7.11) $||\mu||_{\mathcal{M}(\Theta)} = |\mu|(\Theta)$.

**Theorem 3.1.** *(Appendix C of [Con94]) Riesz Representation theorem: If $\Theta$ is a locally compact space, for any $\omega \in (C_0(\Theta, \mathbb{C}))^*$ there exists a unique $\mu \in \mathcal{M}(\Theta)$ such that:*

$$\langle \omega, g \rangle = \int_\Theta g(\theta)\mu(d\theta)$$

*The map $\omega \mapsto \mu$ is a linear bijection (7.34) and $||\mu||_{\mathcal{M}(\Theta)} = ||\omega||_{(C_0(\Theta,\mathbb{C}))^*}$.*

By theorem 3.1 we get that there is an isometric isomorphism between the space of measures $\mathcal{M}(\Theta)$ and the dual of $C_0(\Theta, \mathbb{C})$. Strictly speaking we should always write this explicitly, but in this thesis we will often identify the dual of $C_0(\Theta, \mathbb{C})$ as $\mathcal{M}(\Theta)$. We will use the following abuse of notation, denoting the action of a measure $\mu \in \mathcal{M}(\Theta)$ on a function $g \in C_0(\Theta, \mathbb{C})$ as:

$$\langle \mu, g \rangle = \int_\Theta g(\theta)\mu(d\theta).$$

[1] Note that the same notation is used for all bilinear forms, such as inner products and canonical pairings

Since $||\mu||_{\mathcal{M}(\Theta)} = ||\omega||_{(C_0(\Theta,\mathbb{C}))^*}$, the variance norm of a measure $\mu \in \mathcal{M}(\Theta)$ can be rewritten as:

$$||\mu||_{\mathcal{M}(\Theta)} = \sup_{||g||_\infty \leq 1} |\langle \mu, g \rangle|.$$

### 3.2.2 Feature Map & Linear Operator

Now we are going to introduce a feature map and a linear operator, which are both fundamental in defining function spaces of functions parametrized by measures. We consider the family of functions:

$$\varphi : \mathcal{X} \times \Theta \to \mathbb{C} \text{ s.t. } \varphi(\mathbf{x}, \cdot) \in C_0(\Theta, \mathbb{C}) \ \forall \mathbf{x} \in \mathcal{X}.$$

We define a feature map $\phi : \mathcal{X} \to C_0(\Theta, \mathbb{C})$, mapping $\mathbf{x} \in \mathcal{X}$ to a function $g \in C_0(\Theta, \mathbb{C})$, which is given by:

$$\phi(\mathbf{x}) = \varphi(\mathbf{x}, \cdot).$$

We also define an operator $A : (C_0(\Theta, \mathbb{C}))^* \to \mathbb{C}^{\mathcal{X}}$ or rather $A : \mathcal{M}(\Theta) \to \mathbb{C}^{\mathcal{X}}$, which is given by

$$A(\mu)(\mathbf{x}) = \langle \mu, \phi(\mathbf{x}) \rangle = \langle \mu, \varphi(\mathbf{x}, \cdot) \rangle = \int_\Theta \varphi(\mathbf{x}, \theta) \mu(d\theta).$$

This operator is linear, mapping the sum of measures $A(\mu_1 + \mu_2)$ is equal to summing the map of two measures $A(\mu_1) + A(\mu_2)$ for all $\mu_1, \mu_2 \in \mathcal{M}(\Theta)$. And for any $C \in \mathbb{C}$, $A(C \cdot \mu) = C \cdot A(\mu)$ for all $\mu \in \mathcal{M}(\Theta)$.

### 3.2.3 Constructing the Space

With this linear operator $A$, a linear space of functions $f : \mathcal{X} \to \mathbb{C}$ can be defined by taking its image (7.31), $\text{Im}(A) = \{A(\mu) | \mu \in \mathcal{M}(\Theta)\}$. Note that $A$ might not be injective (7.32), i.e. it might be that two measures $\mu_1, \mu_2 \in \mathcal{M}(\Theta)$ map to the same function. The kernel (7.33) of $A$ is given by $\text{Ker}(A) = \{\mu \in \mathcal{M}(\Theta) | A(\mu) = 0\}$. When one fixes a function $f \in \text{Im}(A)$ and find a $\mu \in \mathcal{M}(\Theta)$ such that $f = A(\mu)$, adding every $\nu \in \text{Ker}(A)$ to $\mu$ amounts to the same fixed function $f$. So for all $\nu \in \text{Ker}(A)$, the function $f$ can also be written as $A(\mu + \nu)$, which we can rewrite as $A(\mu) + A(\nu)$ because of the linearity of $A$, and this equals $A(\mu)$ because $A(\nu) = 0 \ \forall \nu \in \text{Ker}(A)$. To still ensure completeness of $\text{Im}(A)$, we need $\text{Ker}(A)$ to be a closed subspace of $\mathcal{M}(\Theta)$.

**Lemma 3.2.** *$Ker(A) \subset \mathcal{M}(\Theta)$ is closed.*

*Proof.* It is well known from functional analysis that a linear operator such as $A : \mathcal{M}(\Theta) \to \mathbb{C}^{\mathcal{X}}$ is continuous if and only if the kernel of $A$ is a closed subspace of $\mathcal{M}(\Theta)$ ([Con94]). Because $A$ maps to a space of functions in $\mathbb{C}^{\mathcal{X}}$, the function $A(\mu)$ is 0 if it is 0 for every evaluation map i.e. $\delta_{\mathbf{x}}(A(\mu)) = A(\mu)(\mathbf{x}) = 0 \ \forall \mathbf{x} \in \mathcal{X}$. The kernel of $\text{Ker}(A)$ can thus be written as the intersection of the kernels of all evaluation maps $\bigcap_{\mathbf{x} \in \mathcal{X}} \{\mu \in \mathcal{M}(\Theta) | A(\mu)(\mathbf{x}) = 0\}$. These evaluation maps are by definition continuous so $A$ is also continuous. Because $A$ is continuous we can conclude that $\text{Ker}(A)$ is a closed subspace of $\mathcal{M}(\Theta)$. ■

**Theorem 3.3.** *[Con94] If $\mathcal{M}(\Theta)$ is a Banach space and $Ker(A)$ is a closed subspace of $\mathcal{M}(\Theta)$ than the quotient space (7.35) $\mathcal{M}(\Theta)/Ker(A)$ is also a Banach space with elements $[\mu]$, which are named equivalence classes. A norm on $\mathcal{M}(\Theta)/Ker(A)$ can be defined:*

$$||\,[\mu]\,||_{\mathcal{M}(\Theta)/Ker(A)} = \inf_{\nu \in Ker(A)} ||\mu + \nu||_{\mathcal{M}(\Theta)}.$$

*The quotient space $\mathcal{M}(\Theta)/Ker(A)$ is complete with respect to the norm, so it is a Banach space.*

We define the map $j : \mathcal{M}(\Theta)/\mathrm{Ker}(A) \to \mathrm{Im}(A)$. For every $[\mu] \in \mathcal{M}(\Theta)/\mathrm{Ker}(A)$, $j$ maps to a unique $A(\mu) \in \mathrm{Im}(A)$ and the inverse mapping $j^{-1} : \mathrm{Im}(A) \to \mathcal{M}(\Theta)/\mathrm{Ker}(A)$ on this function $A(\mu)$ returns the same $[\mu]$ again, it is by definition a linear bijection. So the $\mathrm{Im}(A)$ becomes a Banach space by setting the norm $||A(\mu)||_{\mathrm{Im}(A)} = ||j^{-1}(A(\mu))||_{\mathcal{M}(\Theta)/\mathrm{Ker}(A)}$. Therefore the Banach norm for a function $f \in \mathrm{Im}(A)$ is the infimum of the norm of all $\mu \in \mathcal{M}(\Theta)$ for which $f(\mathbf{x}) = A(\mu)(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. Through the rest of this paper, we denote the Banach space $\mathrm{Im}(A)$ as $\mathbf{B}$:

$$\mathbf{B} = \left\{ f \in \mathbb{C}^{\mathcal{X}} \,\middle|\, \exists \mu \in \mathcal{M}(\Theta) \text{ s.t. } f(\mathbf{x}) = \int_{\Theta} \varphi(\mathbf{x}, \theta) \mu(d\theta) \ \forall \mathbf{x} \in \mathcal{X} \right\}, \tag{6}$$

which is complete with respect to the norm:

$$\gamma_{\mathbf{B}}(f) \coloneqq \inf_{\mu \in \mathcal{M}(\Theta)} \left\{ |\mu|(\Theta) \middle| f(\mathbf{x}) = \int_{\Theta} \varphi(\mathbf{x}, \theta) \mu(d\theta) \ \forall \mathbf{x} \in \mathcal{X} \right\}.$$

**Making the step from complex-valued functions to real-valued functions.** From now on we have assumed the measures $\mu \in \mathcal{M}(\Theta)$ to be complex-valued, in order to construct a function space with functions $f \in \mathbb{C}^{\mathcal{X}}$, this is the most general setting. In applying this framework to real world problems we might want to consider functions $f \in \mathbb{R}^{\mathcal{X}}$ instead. If $\mu : \Sigma(\Theta) \to \mathbb{C}$ is a complex measure, its absolute value is $|\mu| : \Sigma(\Theta) \to [0, +\infty) \cup \{+\infty\}$. From [Con94], we know there exists a measurable (7.9) function $\beta$ on the measure space (7.8) $(\Theta, \Sigma(\Theta), \mu)$ such that $|\beta(\theta)| = 1$ for almost every $\theta \in \Theta$ and $\mu(E) = \int_E \beta(\theta)|\mu|(d\theta)$ for all measurable sets (7.6) $E \in \Sigma(\Theta)$. So if $\mu$ is real-valued, $\beta(\theta) = \{-1, 1\}$ for almost every $\theta \in \Theta$. We could thus consider $f \in \mathbb{R}^{\mathcal{X}}$, parametrized by real-valued regular Borel measures through the framework defined above, without running into trouble.

### 3.3 Fixing the reference measure to be a probability measure with full support on $\Theta$

In this subsection two sub-spaces of the big Banach space $\mathbf{B}$ are derived. We specialize the discussion by fixing a reference measure. This reference measure is fixed to be a probability measure (7.13) $\tau$ that has full support (7.14) on $\Theta$. We consider all measures $\mu \in \mathcal{M}(\Theta)$ that are absolutely continuous (7.17) with respect to this fixed probability measure $\tau$ and thus have a density (7.18) $p(\theta) = \mu(d\theta)/\tau(d\theta)$. Note that every $\mu$ that satisfies the above also, just as $\tau$, has full support on $\Theta$. The representation in equation (6) can be further rewritten as $f(\mathbf{x}) = \int_{\Theta} \varphi(\mathbf{x}, \theta) p(\theta) \tau(d\theta)$. We denote $\mathcal{M}_\tau(\Theta) \subset \mathcal{M}(\Theta)$ the set of complex-valued regular Borel measures that are absolutely continuous with respect to $\tau$. In a similar way as for $\mathbf{B}$ we can define the space $\mathcal{F}_1$.

$$\mathcal{F}_1 = \left\{ f \in \mathbb{C}^{\mathcal{X}} \,\middle|\, \exists \mu \in \mathcal{M}_\tau(\Theta) \text{ s.t. } f(\mathbf{x}) = \int_{\Theta} \varphi(\mathbf{x}, \theta) \mu(d\theta) = \int_{\Theta} \varphi(\mathbf{x}, \theta) p(\theta) \tau(d\theta) \ \forall \mathbf{x} \in \mathcal{X} \right\}. \tag{7}$$

This is also a Banach space complete with respect to a similar norm as $\mathbf{B}$ but with the infimum over measures in the space $\mathcal{M}_\tau(\Theta)$. If a measure $\mu$ on a measurable space (7.5) $(\Theta, \Sigma(\Theta))$ has a density with respect to a countable additive (7.7), non-negative measure $\tau$, e.g. a probability measure, the total variation of the measure $\mu$ equals the $L_1(\Theta, \tau)$-norm of the density $p(\theta) = \mu(d\theta)/\tau(d\theta)$ [Pol15]. Thus the Banach norm of $\mathcal{F}_1$ is given by:

$$\gamma_1(f) \coloneqq \inf_{p} \left\{ \int_{\Theta} |p(\theta)| \tau(d\theta) \ \middle|\ f(\mathbf{x}) = \int_{\Theta} \varphi(\mathbf{x}, \theta) p(\theta) \tau(d\theta) \ \forall \mathbf{x} \in \mathcal{X} \right\}.$$

To emphasize, the norms $\gamma_{\mathbf{B}}$ and $\gamma_1$ are only equal when the complex-valued regular Borel measure $\mu$ is absolutely continuous with respect to a fixed probability measure $\tau$ that has full support on $\Theta$. So the norm $\gamma_1$ is always larger or equal to $\gamma_{\mathbf{B}}$ and the Banach space $\mathcal{F}_1$ is included in the Banach space $\mathbf{B}$. $\mathcal{F}_1$ is defined as all functions $f \in \mathbb{C}^{\mathcal{X}}$ that have finite $\gamma_1$-norm, which is expressed as the infimum of the $L_1(\Theta, \tau)$-norm of the density $p$ over all decompositions of $f(\mathbf{x}) = \int_{\Theta} \varphi(\mathbf{x}, \theta) p(\theta) \tau(d\theta)$. In the same way we can also define a Hilbert space (7.23) $\mathcal{F}_2$ with functions $f \in \mathbb{C}^{\mathcal{X}}$ that have a

finite $\gamma_2$-norm, the infimum of the $L_2(\Theta, \tau)$-norm of the density $p$ over the same decompositions of $f$.

$$\gamma_2(f) \coloneqq \inf_p \left\{ \left( \int_\Theta |p(\theta)|^2 \tau(d\theta) \right)^{1/2} \;\middle|\; f(\mathbf{x}) = \int_\Theta \varphi(\mathbf{x}, \theta) p(\theta) \tau(d\theta) \; \forall \mathbf{x} \in \mathcal{X} \right\}.$$

**Connection $\mathcal{F}_2$ with $\mathcal{F}_1$** The $\gamma_2$-norm of a function $f \in \mathbb{C}^{\mathcal{X}}$ is always larger than its $\gamma_1$-norm. This can be seen through the Jensen's Inequality (7.40), namely $\left(\int_\Theta |p(\theta)| \tau(d\theta)\right)^2 \leq \int_\Theta |p(\theta)|^2 \tau(d\theta)$. Thus the function space $\mathcal{F}_2$ is included in $\mathcal{F}_1$. Since $\mathcal{F}_2 \subset \mathcal{F}_1$, approximation properties proven for $\mathcal{F}_2$ can be transferred to $\mathcal{F}_1$, but these transferred properties will be sub-optimal for $\mathcal{F}_1$. The function space $\mathcal{F}_2$ is often too small to obtain good approximation results in comparison to approximating with functions from $\mathcal{F}_1$. There are functions with a small number of variables, that are captured by $\mathcal{F}_1$ but not by $\mathcal{F}_2$. An illustrative example is a single real-valued function from the family of functions $\varphi$ with both the input domain $\mathcal{X}$ and the parameter space $\Theta$ equal to $\mathbb{R}$, namely $f^*(\cdot) = \varphi(\cdot, \theta^*) : \mathbb{R} \to \mathbb{R}$. The distribution of this real valued function is a point mass of 1 at $\theta^*$, singular with respect to the 1-dimensional Lebesgue measure $d\theta$. This singular distribution can be approximated with a density function of the form, with $\varepsilon$ tending to $0^+$:

$$p_\varepsilon(\theta) = \begin{cases} \frac{1}{\varepsilon} & \text{if } \theta \in [\theta^*, \theta^* + \varepsilon] \\ 0 & \text{otherwise.} \end{cases}$$

The $L_1$- and squared $L_2$-norms of this density:

$$\gamma_1(f) = \int_{\mathbb{R}} |p_\varepsilon(\theta)| d\theta = \int_{\theta^*}^{\theta^*+\epsilon} \left|\frac{1}{\varepsilon}\right| d\theta = \left[\frac{1}{\varepsilon}\theta\right]_{\theta^*}^{\theta^*+\epsilon} = 1 = 1 \text{ as } \varepsilon \to 0^+$$

$$\gamma_2^2(f) = \int_{\mathbb{R}} |p_\varepsilon(\theta)|^2 d\theta = \int_{\theta^*}^{\theta^*+\varepsilon} \left|\frac{1}{\varepsilon}\right|^2 d\theta = \left[\left(\frac{1}{\varepsilon}\right)^2 \theta\right]_{\theta^*}^{\theta^*+\varepsilon} = \frac{1}{\varepsilon} = \infty \text{ as } \varepsilon \to 0^+.$$

So when $\varepsilon \to 0^+$, for the $L_1$-norm of the density function tends to a singular distribution with bounded norm and hence $f^* \in \mathcal{F}_1$. But for the $L_2$-norm the norm of the density function goes to infinity and thus $f^* \notin \mathcal{F}_2$.

**Recap of defined spaces** To summarize the above results, we have defined three spaces of interest: the Banach spaces $\mathbf{B}$, $\mathcal{F}_1$ and the Hilbert space $\mathcal{F}_2$. Corresponding norms relate to each other as $\gamma_{\mathbf{B}}(f) \leq \gamma_1(f) \leq \gamma_2(f)$ for all $f \in \mathbf{B}$. The spaces are visualized in figure 2.

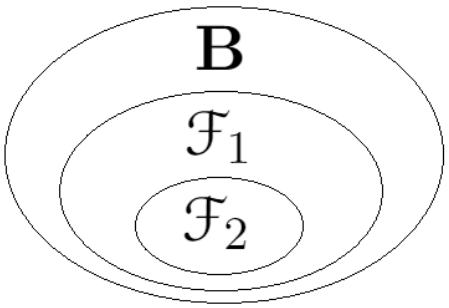


Figure 2: Function spaces $\mathbf{B}$, $\mathcal{F}_1$ & $\mathcal{F}_2$

## 3.4 Connection to RKHS & RKBS

In this subsection the connection of the above introduced spaces with Reproducing Kernel Hilbert and Banach Spaces is presented. First we introduce the more well known definition and properties of

RKHS's and present how this applies to $\mathcal{F}_2$. Subsequently we explain how the properties of a RKBS are different from those of a RKHS and we prove that both $\mathbf{B}$ and $\mathcal{F}_1$ are RKBS's. A framework for constructing reproducing kernels of RKBS's from [LZZ19] is presented and applied to derive the reproducing kernels of $\mathbf{B}$ and $\mathcal{F}_1$.
It is found and proven in [Bac14] that the Hilbert space $\mathcal{F}_2$ is a Reproducing Kernel Hilbert Space (3.4) with corresponding RKHS-norm $\gamma_2$, for which a proof sketch is given in Appendix B.

**Definition 3.4.** *[CT08] A Reproducing Kernel Hilbert Space (RKHS) on a prescribed non-empty set $\mathcal{X}$ is a Hilbert space(7.23) $\mathcal{H}$ of certain functions on $\mathcal{X}$ such that point evaluation functionals are continuous on $\mathcal{H}$. I.e. for all $\mathbf{x} \in \mathcal{X}$ there exists a constant $C_{\mathbf{x}} \in \mathbb{R}^+$ such that:*

$$|\delta_{\mathbf{x}}(f)| = |f(\mathbf{x})| \leq C_{\mathbf{x}}||f||_{\mathcal{H}} \ \forall f \in \mathcal{H},$$

*where the inner product of the Hilbert space $\mathcal{H}$ is denoted as $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the norm $|| \cdot ||_{\mathcal{H}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$.*

Intuitively, this means that if two functions $f, g \in \mathcal{H}$ are close in RKHS-norm, i.e. $||f-g||_{\mathcal{H}}$ is small, than for all points $\mathbf{x} \in \mathcal{X}$ the pointwise evaluation $|f(\mathbf{x}) - g(\mathbf{x})|$ is also small. For a Hilbert space to have the property of continuous evaluation functionals, and thus to be a RKHS, is known to be equivalent to a number of things, visualized in figure 3.

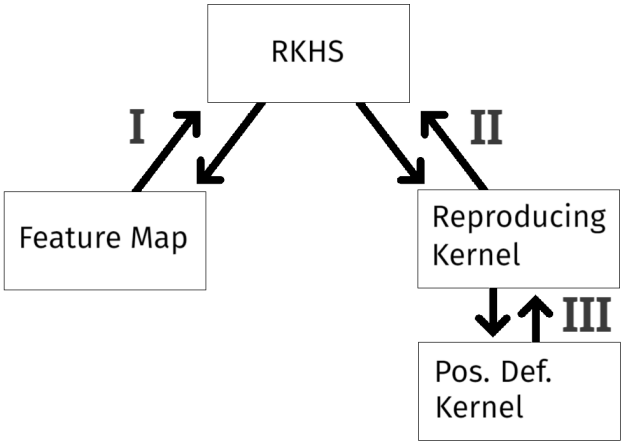


Figure 3: Equivalence relations Reproducing Kernel Hilbert Space

First of all, any RKHS on $\mathcal{X}$ can be realized as the closed subspace of a feature space $\mathcal{W}$, through a suitable feature map $\Phi : \mathcal{X} \to \mathcal{W}$. The feature space $\mathcal{W}$ is also a Hilbert Space with corresponding inner product $\langle \cdot, \cdot \rangle_{\mathcal{W}}$. Functions $f$ in a RKHS $\mathcal{H}$ can be represented as $f(\mathbf{x}) = \langle w, \Phi(\mathbf{x}) \rangle_{\mathcal{W}}$. For the RKHS $\mathcal{F}_2$ the feature space is $L_2(\Theta, \tau)$ with inner product $\langle \cdot, \cdot \rangle_{L_2(\Theta,\tau)}$. The feature map $\Phi : \mathcal{X} \to L_2(\Theta, \tau)$ is given by $\Phi(\mathbf{x}) = \varphi(\mathbf{x}, \cdot)$ and defines functions $f(\mathbf{x}) = \langle p, \Phi(\mathbf{x}) \rangle_{L_2(\Theta,\tau)} = \langle p, \varphi(\mathbf{x}, \cdot) \rangle_{L_2(\Theta,\tau)} = \int_\Theta p(\theta)\varphi(\mathbf{x}, \theta)\tau(d\theta)$.
Secondly, for every RKHS there exists a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ called a Reproducing Kernel, such that:

- $K(\mathbf{x}, \cdot) \in \mathcal{H} \quad \forall \mathbf{x} \in \mathcal{X}$
- $\langle f, K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} = f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{X}, \forall f \in \mathcal{H}.$

The equivalence between the existence of the reproducing kernel and the feature map follows automatically. The reproducing kernel of a RKHS $\mathcal{H}$ can be constructed by taking the inner product in the feature space $\mathcal{W}$ of two feature maps:

- $K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{W}} \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$

Thirdly, for every RKHS the reproducing kernel is positive definite:

- $\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j)\geq 0 \quad \forall \mathbf{x}_1,...,\mathbf{x}_n\in\mathcal{X},\ n\in\mathbb{N},\ \alpha_1,...,\alpha_n\in\mathbb{R}.$

And conversely, for every positive definite kernel $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$, a unique RKHS on $\mathcal{X}$ exists for which $K$ is the reproducing kernel. For $\mathcal{F}_2$ the positive definite reproducing kernel is given by, for all $\mathbf{x},\mathbf{x}'\in\mathcal{X}$:

$$K_{\mathcal{F}_2}(\mathbf{x},\mathbf{x}')=\langle\Phi(\mathbf{x}),\Phi(\mathbf{x}')\rangle_{L_2(\Theta,\tau)}=\langle\varphi(\mathbf{x},\cdot),\varphi(\mathbf{x}',\cdot)\rangle_{L_2(\Theta,\tau)}=\int_\Theta\varphi(\mathbf{x},\theta)\varphi(\mathbf{x}',\theta)\tau(d\theta). \tag{8}$$

I.e. the kernel $K_{\mathcal{F}_2}(\mathbf{x},\mathbf{x}')$ is the expectation, for $\theta$ following the probability distribution $\tau$, of $\varphi(\mathbf{x},\theta)\varphi(\mathbf{x}',\theta)$. Reproducing Kernel Hilbert Spaces are ideal for kernel based learning algorithms. Because of the use of the inner product, the functional analysis of an RKHS is well-known and understood. But as shown in section 3.3 often Hilbert spaces are not large enough and Banach space have richer geometrical structures and norms. That's why Reproducing Kernel Banach Spaces are also investigated, starting with Pre-Reproducing Kernel Banach Spaces.

**Definition 3.5.** *[LZZ19] A Pre-Reproducing Kernel Banach Space (P-RKBS) on a prescribed non-empty set $\mathcal{X}$ is a Banach Space (7.22) $\mathcal{B}$ of certain functions on $\mathcal{X}$ such that every point evaluation function $\delta_\mathbf{x}$, $\forall\mathbf{x}\in\mathcal{X}$ on $\mathcal{B}$ is continuous. I.e. for all $\mathbf{x}\in\mathcal{X}$ there exists a constant $C_\mathbf{x}\in\mathbb{R}^+$ such that:*

$$|\delta_\mathbf{x}(f)|=|f(\mathbf{x})|\leq C_\mathbf{x}||f||_\mathcal{B}\ \forall f\in\mathcal{B}.$$

*In the above the norm of the Banach space $\mathcal{B}$ is denoted as $||\cdot||_\mathcal{B}$.*

Both the spaces $\mathbf{B}$ and $\mathcal{F}_1$ defined in subsection 3.2 are Pre-Reproducing Kernel Banach Spaces.

**Lemma 3.6.** $\mathbf{B}$ *is a P-RKBS*

*Proof.* For every function $\varphi(\mathbf{x},\cdot)\in C_0(\Theta,\mathbb{C})$ we know there for all $\theta\in\Theta$ there always exists a positive constant $C\in\mathbb{R}^+$ that bounds $|\varphi(\mathbf{x},\theta)|$, and this holds for every $\mathbf{x}$ in $\mathcal{X}$. Thus for all $f\in\mathbf{B}$ and $\mathbf{x}\in\mathcal{X}$ the evaluation functionals:

$$|\delta_\mathbf{x}(f)|=|f(\mathbf{x})|\leq\int_\Theta|\varphi(\mathbf{x},\theta)|\cdot|\mu|(d\theta)\leq C|\mu|(\Theta)\leq C\gamma_\mathbf{B}(f),$$

where the last inequality is because $\gamma_\mathbf{B}$ is defined as the infimum over all $\mu$ for which $f(\mathbf{x})=\int_\Theta\varphi(\mathbf{x},\theta)\mu(d\theta)$ for all $\mathbf{x}\in\mathcal{X}$. ■

**Lemma 3.7.** $\mathcal{F}_1$ *is a P-RKBS*

*Proof.* For all $f\in\mathcal{F}_1$ and $\mathbf{x}\in\mathcal{X}$ the evaluation functionals:

$$|\delta_\mathbf{x}(f)|=|f(\mathbf{x})|\leq\int_\Theta|\varphi(\mathbf{x},\theta)|\cdot|\mu|(d\theta)\leq C|\mu|(\Theta)=C||p||_{L_1(\Theta,\tau)}\leq C\gamma_1(f),$$

since the total variation $|\mu|(\Theta)$ for $\mu\in\mathcal{M}_\tau(\Theta)$ is equal to $||p||_{L_1(\Theta,\tau)}$ with $p=\mu(d\theta)/\tau(d\theta)$ the density. The last inequality holds because $\gamma_1$ is defined as the infimum over all $p\in L_1(\Theta,\tau)$ for which $f(\mathbf{x})=\int_\Theta\varphi(\mathbf{x},\theta)p(\theta)\tau(d\theta)$ for all $\mathbf{x}\in\mathcal{X}$. ■

Because of the lack of in an inner product the analysis of a P-RKBS is not as straightforward as for a RKHS. The three equivalences that hold for RKHS's, visualized in figure 3, do not automatically hold for P-RKBS's. The first equivalence, between the property of continuous evaluation functionals and the existence of a feature map, is similar for a P-RKBS as for a RKHS. The story becomes more complicated for the second equivalance relation. A P-RKBS for which a reproducing kernel can be constructed, is a RKBS, but if the construction is simple or complicated depends on the structure of the P-RKBS.

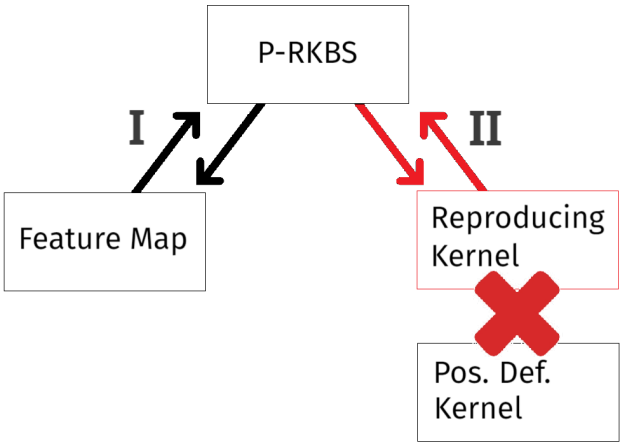


Figure 4: Equivalence relations Pre-Reproducing Kernel Banach Space

Again, any P-RKBS can be realized through a suitable feature map $\Phi : \mathcal{X} \to \mathcal{W}^*$. Here $\mathcal{W}^*$ is a Banach space with norm $||\cdot||_{\mathcal{W}^*}$. $\mathcal{W}^*$ is the dual of a feature space $\mathcal{W}$, which is also a Banach space with norm $||\cdot||_{\mathcal{W}}$. The canonical pairing between $\mathcal{W}$ and its dual $\mathcal{W}^*$ is denoted as $\langle\cdot,\cdot\rangle_{\mathcal{W}\times\mathcal{W}^*}$. Functions $f$ in a P-RKBS can be represented as $\langle w, \Phi(\mathbf{x})\rangle_{\mathcal{W}\times\mathcal{W}^*}$. This equivalence is stated formally in proposition 3.1 of [CSV16]. For the P-RKBS $\mathbf{B}$ the feature space is $\mathcal{M}(\Theta)$, with topological dual $C_0(\Theta,\mathbb{C})$ and canonical pairing $\langle\cdot,\cdot\rangle_{\mathcal{M}(\Theta)\times C_0(\Theta,\mathbb{C})}$. The feature map $\Phi : \mathcal{X} \to C_0(\Theta,\mathbb{C})$ is given by $\Phi(\mathbf{x}) = \varphi(\mathbf{x},\cdot)$, and it defines functions $f(\mathbf{x}) = \langle\mu, \Phi(\mathbf{x})\rangle_{\mathcal{M}(\Theta)\times C_0(\Theta,\mathbb{C})} = \int_\Theta \varphi(\mathbf{x},\theta)\mu(d\theta)$.
For the P-RKBS $\mathcal{F}_1$ the feature space is $\mathcal{M}_\tau$. $L_\infty(\Theta,\tau)$ is a Banach space with functions $g : \Theta \to \mathbb{C}$ with finite $L_\infty(\Theta,\tau)$-norm, it is complete with respect to the essential supremum norm defined as $||g||_{L_\infty(\Theta,\tau)} = \inf\{C \geq 0 \mid \tau(\{\theta| \ |g(\theta)| > C\}) = 0\}$. $\mathcal{M}_\tau$ is isometric to the topological dual of $L_\infty(\Theta,\tau)$ [DS58]. So for the dual of the feature space we take $L_\infty(\Theta,\tau)$ and the canonical pairing is $\langle\cdot,\cdot\rangle_{\mathcal{M}_\tau(\Theta)\times L_\infty(\Theta,\tau)}$. The feature map $\Phi : \mathcal{X} \to L_\infty(\Theta,\tau)$ is given by $\Phi(\mathbf{x}) = \varphi(\mathbf{x},\cdot)$, and it defines functions $f(\mathbf{x}) = \langle\mu, \Phi(\mathbf{x})\rangle_{\mathcal{M}_\tau(\Theta)\times L_\infty(\Theta,\tau)} = \int_\Theta \varphi(\mathbf{x},\theta)\mu(d\theta) = \int_\Theta \varphi(\mathbf{x},\theta)p(\theta)\tau(d\theta)$ with $p \in L_1(\Theta,\tau)$.
For the second equivalence relation in figure 4, between the property of continuous evaluation functionals and the existence of a reproducing kernel, the story splits. For a P-RKBS that is a reflexive Banach space (7.28), this equivalence is similar to the RKHS one. Intuitively, a reflexive Banach space is isometric to the dual of its dual space $(\mathcal{B}^*)^*$. For every reflexive P-RKBS $\mathcal{B}$, with dual space $\mathcal{B}^*$, there exists a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that:

- $K(\mathbf{x},\cdot) \in \mathcal{B}^* \quad \forall \mathbf{x} \in \mathcal{X}$
- $\langle f, K(\mathbf{x},\cdot)\rangle_{\mathcal{B}\times\mathcal{B}^*} = f(\mathbf{x})$.

When the P-RKBS is a non-reflexive Banach Space, the derivation of the reproducing kernel becomes more complicated. A more general definition for the reproducing kernel of P-RKBS's, that applies not only to reflexive P-RKBS's, is given by:

**Definition 3.8.** *[LZZ19] For the definition of the Reproducing Kernel of a P-RKBS we consider the following setting:*
*Let $\mathcal{B}_1$ be an P-RKBS on a set $\Omega_1$. If there exists a Banach Space $\mathcal{B}_2$ of functions on a another set $\Omega_2$, a continuous bilinear form (7.38) $\langle\cdot,\cdot\rangle_{\mathcal{B}_1\times\mathcal{B}_2}$ and a function $K : \Omega_1 \times \Omega_2 \to \mathbb{C}$ such that the following holds:*

- $K(\mathbf{x},\cdot) \in \mathcal{B}_2 \quad \forall \mathbf{x} \in \Omega_1$
- $\langle f, K(\mathbf{x},\cdot)\rangle_{\mathcal{B}_1\times\mathcal{B}_2} = f(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega_1, \forall f \in \mathcal{B}_1,$

*then $K$ is a reproducing kernel for $\mathcal{B}_1$ and $\mathcal{B}_1$ a RKBS. Additionally, if $\mathcal{B}_2$ is a P-RKBS on the set $\Omega_2$ and the following holds:*

- $K(\cdot, \mathbf{y}) \in \mathcal{B}_1 \quad \forall \mathbf{y} \in \Omega_2$
- $\langle K(\cdot, \mathbf{x}), h \rangle_{\mathcal{B}_1 \times \mathcal{B}_2} = h(\mathbf{y}) \quad \forall \mathbf{y} \in \Omega_2, \forall h \in \mathcal{B}_2,$

*then* $\tilde{K}(\mathbf{x}, \mathbf{y}) \coloneqq K(\mathbf{y}, \mathbf{x})$, *with* $\mathbf{x} \in \Omega_2$ *and* $\mathbf{y} \in \Omega_1$ *is a reproducing kernel for* $\mathcal{B}_2$. *If all the above holds we call* $\mathcal{B}_2$ *an adjoint RKBS of* $\mathcal{B}_1$ *and we call* $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *a pair of RKBS's.*

A general framework for the construction of RKBS's and their corresponding reproducing kernel, through the use of a pair of feature maps, is given by:

**Definition 3.9.** *[LZZ19] For the Construction of a RKBS with corresponding Reproducing Kernel we consider* $\mathcal{W}_1, \mathcal{W}_2$ *to be two Banach spaces and* $\langle \cdot, \cdot \rangle_{\mathcal{W}_1 \times \mathcal{W}_2}$ *a continuous bilinear form on* $\mathcal{W}_1 \times \mathcal{W}_2$. *Suppose there exist two non-empty sets* $\Omega_1$, $\Omega_2$ *and feature mappings* $\Phi_1 : \Omega_1 \to \mathcal{W}_1$, $\Phi_2 : \Omega_2 \to \mathcal{W}_2$ *such that span*$\{\Phi_1(\Omega_1)\}$ *is dense in* $\mathcal{W}_1$ *with respect to the bilinear form* $\langle \cdot, \cdot \rangle_{\mathcal{W}_1 \times \mathcal{W}_2}$ *(7.39) and span*$\{\Phi_2(\Omega_2)\}$ *is dense in* $\mathcal{W}_2$ *with respect to the bilinear form* $\langle \cdot, \cdot \rangle_{\mathcal{W}_1 \times \mathcal{W}_2}$, *and can construct:*

$$\mathcal{B}_1 \coloneqq \left\{ f_v(\mathbf{x}) \coloneqq \langle \Phi_1(\mathbf{x}), v \rangle_{\mathcal{W}_1 \times \mathcal{W}_2} \, | v \in \mathcal{W}_2, \mathbf{x} \in \Omega_1 \right\} \text{ with norm } ||f_v||_{\mathcal{B}_1} \coloneqq ||v||_{\mathcal{W}_2}$$
$$\mathcal{B}_2 \coloneqq \left\{ h_u(\mathbf{y}) \coloneqq \langle u, \Phi_2(\mathbf{y}) \rangle_{\mathcal{W}_1 \times \mathcal{W}_2} \, | u \in \mathcal{W}_1, \mathbf{y} \in \Omega_2 \right\} \text{ with norm } ||h_u||_{\mathcal{B}_2} \coloneqq ||u||_{\mathcal{W}_1},$$

*then* $\mathcal{B}_1$ *is an RKBS on* $\Omega_1$ *with the adjoint RKBS* $\mathcal{B}_2$ *on* $\Omega_2$, *and the Reproducing Kernel for* $\mathcal{B}_1$*:*

$$K(\mathbf{x}, \mathbf{y}) \coloneqq \langle \Phi_1(\mathbf{x}), \Phi_2(\mathbf{y}) \rangle_{\mathcal{W}_1 \times \mathcal{W}_2}.$$

*If the Banach Space* $\mathcal{W}_2$ *is a closed subspace of* $\mathcal{W}_1^*$, *the dual space of continuous linear functions on the Banach Space* $\mathcal{W}_1$, *the construction of the pair of RKBS's can be simplified. Than, we can reduce the continuous bilinear form* $\langle u, v \rangle_{\mathcal{W}_1 \times \mathcal{W}_2}$ *by the canonical pairing* $\langle \cdot, \cdot \rangle : \mathcal{W}_1 \times \mathcal{W}_1^* \to \mathbb{C}$ *given by* $\langle u, v \rangle$ *for all* $u \in \mathcal{W}_1, v \in \mathcal{W}_2$. *In this case the density condition is satisfied if* $\mathcal{W}_1 = \overline{span}\{\Phi_1(\Omega_1)\}$ *and span*$\{\Phi_2(\Omega_2)\}$ *is dense in* $\mathcal{W}_1^*$ *under the weak* topology (7.30).*

With the use of this framework, we are now going to construct the kernels for the P-RKBS's $\mathbf{B}$ and $\mathcal{F}_1$, which are both non-reflexive.

**Construction of RKBS B when linear operator $A$ is injective** The Reproducing Kernel and the adjoint RKBS of the RKBS $\mathbf{B}$ are first constructed through the framework defined above, under the assumption that the linear operator $A : \mathcal{M}(\Theta) \to \mathbb{C}^{\mathcal{X}}$ defined to construct the space $\mathbf{B}$ in subsection 3.2.2, is injective. I.e. we assume that every $\mu \in \mathcal{M}(\Theta)$ is mapped to a unique $f \in \mathbb{C}^{\mathcal{X}}$. This is generally not the case but allows for a more simple and insightful construction. We consider the non-empty set $\Omega_1$ to be the input domain $\mathcal{X}$ and $\Omega_2$ to be the parameter space $\Theta$. The Banach Space $\mathcal{W}_1$ is $C_0(\Theta, \mathbb{C})$ and for $\mathcal{W}_2$ we take $\mathcal{M}(\Theta)$. Since $\mathcal{W}_2$ is isometric to and thus a closed subspace of $\mathcal{W}_1^*$, we take the continuous bilinear to be the canonical pairing $\langle g, \mu \rangle = \int_\Theta g(\theta)\mu(d\theta)$. The feature mappings $\Phi_1 : \mathcal{X} \to C_0(\Theta, \mathbb{C})$ and $\Phi_2 : \Theta \to \mathcal{M}(\Theta)$ are given by:

$$\Phi_1(\mathbf{x}) \coloneqq \varphi(\mathbf{x}, \cdot), \text{ with } \forall \mathbf{x} \in \mathcal{X}, \, \varphi : \mathcal{X} \times \Theta \to \mathbb{C}$$
$$\Phi_2(\psi) \coloneqq \delta_\psi(\cdot), \, \forall \psi \in \Theta.$$

The pair of RKBS's $\mathcal{B}_{1,\mathbf{B}}$ and $\mathcal{B}_{2,\mathbf{B}}$ can be constructed as follows:

$$\mathcal{B}_{1,\mathbf{B}} \coloneqq \left\{ f_\mu(\mathbf{x}) \coloneqq \langle \Phi_1(\mathbf{x}), \mu \rangle = \int_\Theta \varphi(\mathbf{x}, \theta)\mu(d\theta) \big| \mathbf{x} \in \mathcal{X}, \mu \in \mathcal{M}(\Theta) \right\}$$
$$\mathcal{B}_{2,\mathbf{B}} \coloneqq \left\{ h_g(\psi) \coloneqq \langle g, \Phi_2(\psi) \rangle = \int_\Theta g(\theta)\delta_\psi(d\theta) = g(\psi) \big| g \in C_0(\Theta, \mathbb{C}), \psi \in \Theta \right\},$$

with corresponding norms $||f_\mu||_{\mathcal{B}_{1,\mathbf{B}}} = ||\mu||_{\mathcal{M}(\Theta)}$ and $||h_g||_{\mathcal{B}_{2,\mathbf{B}}} \coloneqq ||g||_{C_0(\Theta,\mathbb{C})}$. The definition of the Reproducing Kernel Banach norm $||\cdot||_{\mathcal{B}_{1,\mathbf{B}}}$ is not equal to $\gamma_{\mathbf{B}}$, since in the construction of $\mathbf{B}$ the map $A$ is not assumed to be injective. The Reproducing Kernel of $\mathcal{B}_{1,\mathbf{B}}$ through this construction, for all $\mathbf{x} \in \mathcal{X}$ and $\psi \in \Theta$:

$$K_{\mathcal{B}_{1,\mathbf{B}}}(\mathbf{x}, \psi) = \langle \Phi_1(\mathbf{x}), \Phi_2(\psi) \rangle = \int_\Theta \varphi(\mathbf{x}, \theta)\delta_\psi(d\theta) = \varphi(\mathbf{x}, \psi).$$

Note that this Reproducing Kernel is not positive definite. The density conditions for the construction to hold are:

1. $C_0(\Theta, \mathbb{C}) = \overline{\text{span}}\{\varphi(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\}$ or equivalently the $\text{span}\{\varphi(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\}$ is dense in $C_0(\Theta, \mathbb{C})$ under the weak topology (7.29) meaning $\forall g \in C_0(\Theta, \mathbb{C})$ there exists a sequence $(g_n) \in \text{span}\{\varphi(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\}$ such that $\langle g_n, \mu\rangle \to \langle g, \mu\rangle$ as $n$ goes to infinity $\forall \mu \in \mathcal{M}(\Theta)$.

2. $\text{span}\{\delta_\psi|\psi \in \Theta\}$ is dense in $\mathcal{M}(\Theta)$ with respect to the weak$^*$ topology.

It is well known that the space of linear combinations of Dirac measures supported on different point in a space $\Theta$ is dense in $\mathcal{M}(\Theta)$ with respect to the weak$^*$ topology ([Par67]), so condition 2 holds. If the condition 1 above is satisfied is dependent on the choice of feature map $\Phi_1$. If one would take for example a function $\varphi$ independent of $\mathbf{x}$ it is clear this condition is not satisfied. Condition 1 needs to be satisfied to assure the definition of the norm, that every function $f_\mu \in \mathcal{B}_{1,\mathbf{B}}$ is defined through a unique measure $\mu \in \mathcal{M}(\Theta)$. If and only if the closed span of $\{\varphi(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\}$ is not equal to $C_0(\Theta, \mathbb{C})$ than one can find a measure which is 0 on this smaller subset $\overline{\text{span}}\{\varphi(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\}$, and not 0 in $C_0(\Theta, \mathbb{C}) \setminus \overline{\text{span}}\{\varphi(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\}$. Condition 1 is thus equivalent to the condition of the operator $A : \mathcal{M}(\Theta) \to \mathbb{C}^{\mathcal{X}}$ defined in subsection 3.2 to be injective. We assume this to be true, so the construction holds.

**Construction of RKBS B when linear operator $A$ is not injective** The construction of the Reproducing Kernel can be extended when not assuming the linear operator $A$ to be injective. For $\mathcal{W}_2$ than, instead of $\mathcal{M}(\Theta)$, the quotient space $\mathcal{M}(\Theta)/\text{Ker}(A)$ with elements $[\mu]$ is considered. In this way the norm $||f_{[\mu]}||_{\mathcal{B}_{1,\mathbf{B}}} = ||\,[\mu]\,||_{\mathcal{M}(\Theta)/\text{Ker}(A)} = \inf_{\nu \in \text{Ker}(A)} ||\mu + \nu||_{\mathcal{M}(\Theta)} = \inf_{\mu \in \mathcal{M}(\Theta)} ||\mu||_{\mathcal{M}(\Theta)} = \gamma_{\mathbf{B}}(f)$. For the density conditions to hold $\mathcal{W}_1$ is taken $\overline{\text{span}}\{\varphi(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\} \subset C_0(\Theta, \mathbb{C})$, the closure of the linear span of the functions $\varphi(\mathbf{x}, \cdot) : \Theta \to \mathbb{C}$ that are continuous for all $\mathbf{x} \in \mathcal{X}$. The same non-positive definite reproducing kernel, $K_{\mathbf{B}}(\mathbf{x}, \psi) = \varphi(\mathbf{x}, \psi)$ for all $\mathbf{x} \in \mathcal{X}$ and $\psi \in \Theta$, is obtained.

**Construction of RKBS $\mathcal{F}_1$ when linear operator $A$ is injective** Also for the RKBS $\mathcal{F}_1$ we first construct the reproducing kernel $K_{\mathcal{F}_1}$ under the assumption hat every measure $\mu \in \mathcal{M}_\tau(\Theta)$ is mapped to a unique $f \in \mathbb{C}^{\mathcal{X}}$. We consider the non-empty sets $\Omega_1$ and $\Omega_2$ to be equal to the input domain $\mathcal{X}$ and the parameter space $\Theta$ respectively. The Banach Space $\mathcal{W}_1$ is equal to $L_\infty(\Theta, \tau)$and for the Banach Space $\mathcal{W}_2$ we take $\mathcal{M}_\tau(\Theta)$. Since $\mathcal{W}_2$ is isometric to and thus a closed subspace of $\mathcal{W}_1^*$, we take the continuous bilinear form to be the canonical pairing $\langle g, \mu\rangle = \int_\Theta g(\theta)\mu(d\theta)$. The feature mapping $\Phi_1 : \mathcal{X} \to L_\infty(\Theta, \tau)$ is equal to $\Phi_1(\mathbf{x}) \coloneqq \varphi(\mathbf{x}, \cdot)$, with $\forall \mathbf{x} \in \mathcal{X}$, $\varphi : \mathcal{X} \times \Theta \to \mathbb{C}$ and $\Phi_2 : \Theta \to \mathcal{M}_\tau(\Theta)$ is given by $\Phi_2(\psi) \coloneqq \delta_\psi(\cdot)$, $\forall \psi \in \Theta$. The pair of RKBS's $\mathcal{B}_{1,\mathcal{F}_1}$ and $\mathcal{B}_{1,\mathcal{F}_1}$ can be constructed as follows:

$$\mathcal{B}_{1,\mathcal{F}_1} \coloneqq \left\{ f_\mu(\mathbf{x}) \coloneqq \langle \Phi_1(\mathbf{x}), \mu\rangle = \int_\Theta \varphi(\mathbf{x}, \theta)\mu(d\theta) \middle| \mathbf{x} \in \mathcal{X}, \mu \in \mathcal{M}_\tau(\Theta) \right\}$$
$$\mathcal{B}_{2,\mathcal{F}_1} \coloneqq \left\{ h_g(\psi) \coloneqq \langle g, \Phi_2(\psi)\rangle = \int_\Theta g(\theta)\delta_\psi(d\theta) = g(\psi) \middle| g \in L_\infty(\Theta, \tau), \psi \in \Theta \right\},$$

with corresponding norms $||f_\mu||_{\mathcal{B}_{1,\mathcal{F}_1}} \coloneqq ||\mu||_{\mathcal{M}_\tau(\Theta)}$ and $||h_g||_{\mathcal{B}_{2,\mathcal{F}_1}} \coloneqq ||g||_{L_\infty(\Theta,\tau))}$. Again, note here that the norm $||\cdot||_{\mathcal{B}_{1,\mathcal{F}_1}} = ||\mu||_{\mathcal{M}_\tau(\Theta)} = ||p||_{L_1(\Theta,\tau)}$ is not equal to $\gamma_1$. The non-positive definite Reproducing Kernel of $\mathcal{B}_{1,\mathcal{F}_1}$ through this construction, for all $\mathbf{x} \in \mathcal{X}$ and $\psi \in \Theta$:

$$K_{\mathcal{B}_{1,\mathcal{F}_1}}(\mathbf{x}, \psi) = \langle \Phi_1(\mathbf{x}), \Phi_2(\psi)\rangle = \int_\Theta \varphi(\mathbf{x}, \theta)\delta_\psi(d\theta) = \varphi(\mathbf{x}, \psi).$$

The density conditions for the construction to hold are:

1. $L_\infty(\Theta, \tau) = \overline{\text{span}}\{\varphi(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\}$ or equivalently the $\text{span}\{\varphi(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\}$ is dense in $L_\infty(\Theta, \tau)$ under the weak topology.

2. $\delta_\Theta(\cdot)$ is dense in $\mathcal{M}_\tau(\Theta)$ with respect to the weak$^*$ topology.

Since the space of linear combinations of Dirac measures supported on different points in a space $\Theta$ is dense in $\mathcal{M}(\Theta)$ with respect to the weak* topology, it is also dense in $\mathcal{M}_\tau \subset \mathcal{M}(\Theta)$. Condition 1 does only hold, since every measure $\mu \in \mathcal{M}_\tau(\Theta)$ maps to a unique function $f \in \mathbb{C}^{\mathcal{X}}$

**Construction of RKBS $\mathcal{F}_1$ when linear operator $A$ is not injective** Again the construction of the reproducing kernel of $\mathcal{F}_1$ can be extended to when the linear operator $A : \mathcal{M}_\tau(\Theta) \to \mathbb{C}^{\mathcal{X}}$ defining the space is not injective. $\mathcal{W}_1$ is taken to be $\overline{\text{span}}\{\varphi(\mathbf{x}, \cdot)|\mathbf{x} \in \mathcal{X}\} \subset C_0(\Theta, \mathbb{C})$ and $\mathcal{W}_2 = \mathcal{M}_\tau(\Theta)/\text{Ker}(A)$. Also here the reproducing kernel obtained is the same, $K_{\mathcal{F}_1}(\mathbf{x}, \psi) = \varphi(\mathbf{x}, \psi)$ for all $\mathbf{x} \in \mathcal{X}$ and $\psi \in \Theta$.

# 4 Two-layer Neural Networks as functions parameterized by measures

In the setting described above the specific function from the family of functions $\varphi$ remains unspecified. By considering a fixed basis function from this family of functions we can make the connection to two-layer neural networks. The general framework from section 3 is applied to this setting with a fixed basis function, for several commonly used activation functions, and several function spaces of two-layer neural networks are constructed. Furthermore we describe the advantages and what changes in the analysis of neural networks in this representation. In the final part of this section 4.2 a way of rewriting the neural networks is presented, such that it is fitting for the analysis in chapter 6.

## 4.1 Spaces of two-layer neural networks parameterized by measures

If we consider a certain fixed basis function from this family of functions $\varphi$ we can make the connection to two-layer neural networks. This basis function is the function of a single neuron in the hidden layer. For the $i$-th neuron in the hidden layer this function is given by $\varphi(\mathbf{x}, \theta_i) = \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i)$, with activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and where $(\mathbf{w}_i, b_i) = \theta_i \in \Theta \subseteq \mathbb{R}^{d+1}$. The space of two-layer neural networks with $m$ neurons is given by:

$$\mathcal{F}_\sigma^m = \left\{ f : \mathcal{X} \to \mathbb{R} \,\middle|\, \exists a_1, ..., a_m \in \mathbb{R}, \theta_1, ..., \theta_m \in \Theta \text{ s.t. } f(\mathbf{x}) = \sum_{i=1}^m a_i \sigma(\mathbf{w_i}^\top \mathbf{x} + b_i) \ \forall \mathbf{x} \in \mathcal{X} \right\}. \quad (9)$$

For every $m \in \mathbb{N}$, $\mathcal{F}_\sigma^m$ is a non-linear function space, since the property of addition (7.19) does not hold. For example for $f, g \in \mathcal{F}_\sigma^1$, given by $f(\mathbf{x}) = a_1\varphi(\mathbf{x}, \theta_1)$ and $g(\mathbf{x}) = a_2\varphi(\mathbf{x}, \theta_2)$, their sum is not a function in $\mathcal{F}_\sigma^1$ but in $\mathcal{F}_\sigma^2$.

**Definition 4.1.** *[Pin99] A set of functions $\mathcal{G}$ is Dense in $\mathcal{C}(\mathbb{R}^d)$, if it is true that for any function $f \in \mathcal{C}(\mathbb{R}^d)$, any compact subset $K \subset \mathbb{R}^d$, and any $\varepsilon > 0$, there exists a $g \in \mathcal{G}$ such that: $\max_{\mathbf{x}\in K} |f(\mathbf{x}) - g(\mathbf{x})| < \varepsilon$. I.e. all functions $f \in \mathcal{C}(\mathbb{R}^d)$ can be approximated arbitrarily well by the function class $\mathcal{G}$.*

$\mathcal{F}_\sigma = \bigcup_{m\geq 1} \mathcal{F}_\sigma^m$ is the union of all functions space $\mathcal{F}^m$ with $m \geq 1$, thus consists of all functions that can be represented as a linear combination of the fixed basis function. For the activation function $\sigma$ in $\mathcal{F}_\sigma$ the following density result is derived:

**Theorem 4.2.** *[Pin99] Let $\sigma \in \mathcal{C}(\mathbb{R})$ or Riemann-integrable (continuous almost everywhere). Then $\mathcal{F}_\sigma$ is dense in $\mathcal{C}(\mathbb{R}^d)$, if and only if $\sigma$ is not a polynomial.*

Thus with a non-polynomial fixed basis function from the family of function $\varphi$, any continuous function in $\mathbb{R}^d$ can be approximated arbitrarily well, by a single hidden layer neural network, when the number of neurons is allowed to be infinite. Note that all standard activation functions, like sigmoid or ReLU, would suffice. But according to what is known in machine learning as the 'No Free Lunch Theorem', there is no function class that can be used to approximate all functions. The function class $\mathcal{F}_\sigma$ is too large and thus needs a bound on the complexity to reduce the size. For the

representation in (9) this bound can for example be on the number of neurons $m$, or on the $\ell_1$- or $\ell_2$-norm of the parameter vector $\theta$. Here the parameter vector is the concatenation of all parameters $a_i$ and $\theta_i$ for $1 \leq i \leq m$.
Using the framework from subsection 3.2, we can define the function space $\mathbf{B}_\sigma$. This space consists out of functions that can be represented as the parametrization of the basis function $\varphi(\mathbf{x}, (\mathbf{w}, b)) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$, over a real-valued regular Borel measure $\mu$ with finite variation:

$$\mathbf{B}_\sigma = \left\{ f : \mathcal{X} \to \mathbb{R} \middle| \ \exists \mu \in \mathcal{M}(\Theta) \text{ s.t. } f(\mathbf{x}) = \int_\Theta \sigma(\mathbf{w}^\top \mathbf{x} + b) \mu(d(\mathbf{w}, b)) \ \forall \mathbf{x} \in \mathcal{X} \right\}. \tag{10}$$

$\mathbf{B}_\sigma$ is also a Banach Space and linear around the parameters of the space, which are the measures $\mu$. It is complete with respect to the norm $\gamma_{\mathbf{B}_\sigma} = \inf_{\mu \in \mathcal{M}(\Theta)} \{|\mu|(\Theta) | f(\mathbf{x}) = \int_\Theta \sigma(\mathbf{w}^\top \mathbf{x} + b) \mu(d(\mathbf{w}, b)) \ \forall \mathbf{x} \in \mathcal{X}\}$. This function class is also too large, and for the representation in (10) a complexity bound on the measure $\mu$ can be considered. This would mean putting a bound $C$ on the norm $\gamma_{\mathbf{B}_\sigma}$ and only considering functions $f \in \mathbf{B}_\sigma$ with $\gamma_{\mathbf{B}_\sigma} \leq C$.
We can obtain subspaces $\mathcal{F}_{2,\sigma} \subset \mathcal{F}_{1,\sigma} \subset \mathbf{B}_\sigma$ by only considering measures $\mu \in \mathcal{M}(\Theta)$ that are absolutely continuous with respect to a fixed probability measure $\tau$ that has full support on $\Theta$. Taking the infimum of either the $L_1(\Theta, \tau)$- or $L_2(\Theta, \tau)$-norm of the resulting density $p(\mathbf{w}, b) = \mu(d(\mathbf{w}, b))/\tau(d(\mathbf{w}, b))$ gives us respectively the $\gamma_{1,\sigma}$- and $\gamma_{2,\sigma}$-norms. Because the reference measure $\tau$ has full support on $\Theta$ we have potentially an infinite number of neurons, and by bounding either the $L_1(\Theta, \tau)$- or $L_2(\Theta, \tau)$-norm one lets a sparsity inducing complexity norm select the number of neurons automatically.
Note that the spaces $\mathbf{B}_\sigma$, $\mathcal{F}_{1,\sigma}$ and $\mathcal{F}_{2,\sigma}$ inherit the topology and mathematical structures of respectively $\mathbf{B}$, $\mathcal{F}_1$ and $\mathcal{F}_2$ as long as the basis function of a single neuron in the hidden layer $\sigma(\mathbf{w}^\top \mathbf{x} + b)$ with $(\mathbf{w}, b) = \theta \in \Theta$, is from the family of functions $\varphi$ as defined in subsection 4.1. For commonly used activation functions such as the ReLU $\varphi(\mathbf{x}, (\mathbf{w}, b)) = (\mathbf{w}^\top \mathbf{x} + b)_+$ and the sigmoid $\varphi(\mathbf{x}, (\mathbf{w}, b)) = 1/(1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)})$ this is the case, i.e. for both holds that $\forall \mathbf{x} \in \mathcal{X}$, $\varphi(\mathbf{x}, \cdot) \in C_0(\Theta, \mathbb{R})$. So $\mathbf{B}_\sigma$ and $\mathcal{F}_{1,\sigma}$ are also RKBS's and $\mathcal{F}_{2,\sigma}$ is also a RKHS for common choices of activation function $\sigma$.
The measure $\mu$ can be interpreted as 'the distribution of the neurons'. An advantage of the representation in (10) is that the optimization would be over the measure $\mu$ instead of all the separate parameters $\theta_i$ for $i = 1, ..., m$. When minimizing a functional $J = \mathbb{E}[\ell(\mathbf{y}, f(\mathbf{x}))]$ with $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$ a loss function and $f : \mathcal{X} \to \mathbb{R}$ such as in the representation in (9), the functional $J$ is non-convex since it has multiple unique minima. This due to the fact that in the hidden layer there are multiple neurons which are assigned different parameter values in order to minimize the loss. Apart from the parameter values, these neurons are the same. So one could swap any two neurons in the hidden layer and account for the change by adjusting the values of the weights in the output layer. With this the set of parameters would change, but the value of the loss function would not. For a two-layer neural network with $m$ neurons in its hidden layer, there thus are $m!$ equivalent minima, corresponding to the $m!$ different allocations of the neurons. When optimizing over the measure $\mu$ there is only 1 optimal measure $\mu$ the optimization can converge towards. So by considering the representation in (10) minimizing the functional $J = \mathbb{E}[\ell(\mathbf{y}, f(\mathbf{x}))]$ becomes a convex optimization problem.

**Mean Field perspective.** For $m$ large enough there has been recent work (e.g. [Son18], [MMM19], [DL19]) on descripting the evolution of the measure $\mu$ over time $t$, when the network is trained with Stochastic Gradient Descent, by a PDE. This PDE is known as the mean field description or distributional dynamics. These distributional dynamics have been proven effective to show convergence of two layer neural networks trained with SGD.

## 4.2 Rewriting the two-layer neural network

In chapter 6, when analyzing the approximation properties of two-layer neural networks, we use the following rewriting/assumption steps introduced by [Bac14], unless explicitly mentioned otherwise:

1. We assume the input vector $\mathbf{x} \in \mathbb{R}^d$ is bounded in $\ell_q$-norm by a scalar $R \in \mathbb{R}^+$. We denote this ball as $B_{R,q} = \{\mathbf{x} \mid ||\mathbf{x}||_q \leq R\} \subset \mathbb{R}^d$

2. We append this scalar $R$ to the input vector $\mathbf{x} \in \mathbb{R}^d$ to create a new variable $\mathbf{z} = [\mathbf{x}^\top, R]^\top \in \mathbb{R}^{d+1}$. For all $q \in [2, \infty]$ the $\ell_q$-norm of $\mathbf{z}$ is bounded by $\sqrt{2}R$. The input domain $\mathcal{X}$ of the two-layer neural network is thus given by the ball $B_{\sqrt{2}R,q} = \{\mathbf{z} \mid ||\mathbf{z}||_q \leq \sqrt{2}R\} \subset \mathbb{R}^{d+1}$

3. We define the parameter vector $\theta = [\mathbf{w}^\top, b/R]^\top \in \mathbb{R}^{d+1}$, so in terms of the new variable $\mathbf{z}$ a basis function of a two-layer neural network is $\sigma(\mathbf{w}^\top \mathbf{x} + b) = \sigma(\theta^\top \mathbf{z})$. The parameter space $\Theta$ is thus given by $\mathbb{R}^{d+1}$

For the above steps, the activation function can remain undefined. When we specify the activation functions as homogeneous the rewriting can be extended. The homogeneous activation functions considered are $\sigma(u) = (u)_+^\alpha$, for which $\alpha = 1$ gives the ReLU activation function. If mentioned that the activation functions are homogeneous, the following assumptions hold on top of the previous ones:

4. Because of this homogeneity we can assume w.l.o.g. that the parameters of the basis functions of a two-layer neural network $\theta = [\mathbf{w}^\top, b/R]^\top$ are equal to $1/R$ in $\ell_p$-norm, with $1/p + 1/q = 1$. So if $q = 1$ then $p = \infty$ and if $q = 2$ then $p = 2$. Thus for two-layer neural networks with homogeneous activation the space $\Theta$ is a sphere denoted as $S^d_{1/R,q} = \{\theta \mid ||\theta||_q = 1/R\} \subset \mathbb{R}^{d+1}$. This means $R \cdot \theta = [R\mathbf{w}^\top, b]^\top$ is defined on the unit sphere with $\ell_q$-norm $S^d_{1,q} \subset \mathbb{R}^{d+1}$.

5. We can take $p = q = 2$, because for $p \in [1, \infty]$ the $||\cdot||_p$-norm are equivalent to $C||\cdot||_2$ with the constant $C$ bounded by $d^{\alpha/2}$ ([Bac14]) and for the results we consider unspecified constant only dependent on $d$.

6. For $q = 2$, adding the bounding parameter $R$ and the input vector $\mathbf{x}$ to create a new variable $\mathbf{z}$, is a trick to send an $\ell_2$-ball of radius $R$ in $\mathbb{R}^d$ to a spherical cap in $\mathbb{R}^{d+1}$. Because of the homogeniety, the activation functions are invariant to changing of the data scale of the input data. We thus can consider functions on the spherical cap of the unit sphere with radius 1 and Euclidean norm, denoted as $S^d_{1,2} = \{\mathbf{z} \mid ||\mathbf{z}||_2 = 1\} \subset \mathbb{R}^{d+1}$.

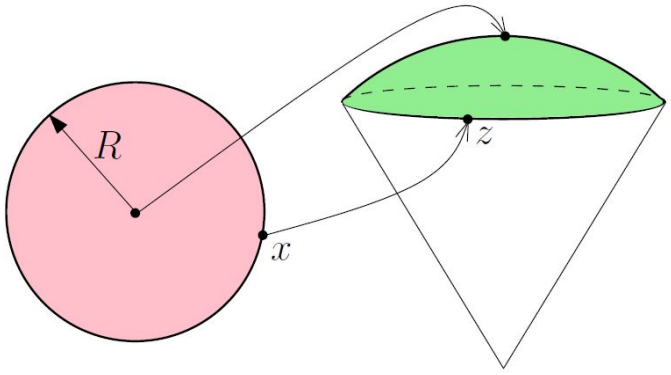


Figure 5: Sending a ball of radius $R$ in $\mathbb{R}^d$ to a spherical cap in $\mathbb{R}^{d+1}$

7. On this spherical cap the Lipschitz-continuity and differentiability properties of the original function are completely transfered. Describing how to extend to the full sphere, with the preservation of properties, is a classical result and described in [H92].

We started with functions defined within a ball of radius $R$ in $\mathbb{R}^d$, meaning such a function can assign a number to each point within the ball. By rewriting the neural network and making assumptions on the activation function we can consider functions on the unit sphere of Euclidean norm in $\mathbb{R}^{d+1}$. Functions defined on the unit sphere are easier to analyse. After analysing functions on the sphere $\mathbb{S}^d \subset \mathbb{R}^{d+1}$, results proven can be translated back to the entire space $\mathbb{R}^{d+1}$ ([Bac14]).

# 5 Representer Theorems

In this section we will consider the setting of minimizing a functional that only depends on a subset of values in the input domain $\widehat{\mathcal{X}} \subset \mathcal{X}$, consisting out of training samples $\mathbf{x}_1, ..., \mathbf{x}_n$. The functional to be minimized is the empirical risk $\widehat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{y}_i, f(\mathbf{x}_i))$. In section 3.4 we have shown that the functions in the spaces $\mathbf{B}$, $\mathcal{F}_1$ and $\mathcal{F}_2$ can be written in the form $f(\mathbf{x}) = \langle w, \Phi(\mathbf{x})\rangle$ through the use of a feature map. For the RKHS $\mathcal{F}_2$ this inner product is in the feature space $L_2(\Theta, \tau)$ and $w$ are elements from this feature space. For the RKBS's $\mathbf{B}$ and $\mathcal{F}_1$, $w$ are elements from the measure spaces $\mathcal{M}(\Theta)$ and $\mathcal{M}_\tau(\Theta)$ respectively and $\langle \cdot, \cdot \rangle$ denotes a canonical pairing. A general representer theorem states that when considering functions of the form $f(\mathbf{x}) = \langle w, \Phi(\mathbf{x})\rangle$ the solution of, for all $\lambda > 0$:

$$w_\lambda^* = \underset{w \in \mathcal{W}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(\mathbf{y}_i, \langle w, \Phi(\mathbf{x}_i)\rangle\right) + \lambda ||w||^2 \right\},$$

is always of the form, with coefficients $\alpha_1, ..., \alpha_n \in \mathbb{R}$:

$$w_\lambda^* = \sum_{i=1}^{n} \alpha_i \Phi(\mathbf{x}_i).$$

This implies:

$$f_\lambda^*(\mathbf{x}) = \langle w_\lambda^*, \Phi(\mathbf{x})\rangle = \sum_{i=1}^{n} \alpha_i \left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x})\right\rangle.$$

Note that to derive a representer theorem the reproducing kernel is not necessary, knowing the form of the reproducing kernel does give more insight in the form of the solution. We derive representer theorems for the general spaces $\mathbf{B}$, $\mathcal{F}_1$ and $\mathcal{F}_2$ with the basis function unspecified, but these equivalently hold for $\mathbf{B}_\sigma$, $\mathcal{F}_{1,\sigma}$ and $\mathcal{F}_{2,\sigma}$, for any activation function $\sigma$ within the family of functions $\varphi$.

## 5.1 Representer Theorem RKHS $\mathcal{F}_2$

Functions $f \in \mathcal{F}_2$ can be represented as $f(\mathbf{x}) = \langle p, \Phi(\mathbf{x})\rangle_{L_2(\Theta,\tau)}$ with $\Phi(\mathbf{x}) = \varphi(\mathbf{x}, \cdot)$. Considering functions of that form, the solution of:

$$p_\lambda^* = \underset{w \in \mathcal{W}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(\mathbf{y}_i, \langle p, \Phi(\mathbf{x}_i)\rangle_{L_2(\Theta,\tau)}\right) + \lambda ||p||^2_{L_2(\Theta,\tau)} \right\},$$

is always of the form, with coefficients $\alpha_1, ..., \alpha_n \in \mathbb{R}$:

$$p_\lambda^* = \sum_{i=1}^{n} \alpha_i \Phi(\mathbf{x}_i).$$

This implies:

$$f_\lambda^*(\mathbf{x}) = \langle p_\lambda^*, \Phi(\mathbf{x})\rangle_{L_2(\Theta,\tau)} = \sum_{i=1}^{n} \alpha_i \left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x})\right\rangle_{L_2(\Theta,\tau)} = \sum_{i=1}^{n} \alpha_i \left\langle \varphi(\mathbf{x}_i, \cdot), \varphi(\mathbf{x}, \cdot)\right\rangle_{L_2(\Theta,\tau)}. \tag{11}$$

As described in section 3.4 the RKHS $\mathcal{F}_2$ of functions with a finite RKHS-norm $\gamma_2$ has a corresponding positive definite reproducing kernel $K_{\mathcal{F}_2}(\mathbf{x}, \mathbf{x}') = \langle \varphi(\mathbf{x}, \cdot), \varphi(\mathbf{x}', \cdot)\rangle_{L_2(\Theta,\tau)} = \int_\Theta \varphi(\mathbf{x}, \theta)\varphi(\mathbf{x}', \theta)\tau(d\theta)$. Note that the optimal solution is of the form $f_\lambda^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K_{\mathcal{F}_2}(\mathbf{x}_i, \mathbf{x})$. This confirms the well known result for representer theorems in RKHS:

**Theorem 5.1.** *[Sch01] Consider training data* $(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n) \in \mathcal{X} \times \mathcal{Y}$*, a loss function* $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$*, a kernel* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *and* $\mathcal{H}$ *the corresponding RKHS with inner product* $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ *and norm* $|| \cdot ||_{\mathcal{H}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$*. The Representer Theorem for a RKHS states that for the optimization problem, for every* $\lambda > 0$*:*

$$f^* = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i) + \lambda ||f||_{\mathcal{H}} \right\},$$

*the optimal function* $f^*$ *has a representation of the form:*

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i),$$

*with* $\alpha_i \in \mathbb{R}\ \forall\ 1 \leq i \leq n$*.*

Explicit regularization in $\mathcal{F}_2$ can be done by adding a penalizing term $\lambda ||f||_{\mathcal{F}_2}$ or constraining $f$ to be in a ball defined as $\mathcal{F}_{2,C} = \{f \in \mathcal{F}_2 \mid \gamma_2(f) \leq C\}$. It is shown in appendix C that the set of solutions $\{f^*_\lambda \mid \lambda > 0\}$, corresponding to the penalty term regularization method, is a subset of $\{f^*_C \mid C > 0\}$, the set of solutions for regularizing by constraining in a ball. Thus for every value of $\lambda \in \mathbb{R}^+$ there exists a value $C \in \mathbb{R}^+$ such that $f^*_\lambda = f^*_C$. Thus the representer theorem for the RKHS defined above also holds when regularizing by constraining $f$ to be in a ball.
Since we only have the subset of values $\widehat{\mathcal{X}} \subset \mathcal{X}$ we only consider functions $f_{|\widehat{\mathcal{X}}} : \widehat{\mathcal{X}} \to \mathbb{R}$ instead of functions $f : \mathcal{X} \to \mathbb{R}$. The space $\mathcal{F}_{2|\widehat{\mathcal{X}}}$ consists of all these functions $f_{|\widehat{\mathcal{X}}}$ with finite norm $\gamma_{2|\widehat{\mathcal{X}}}(f_{|\widehat{\mathcal{X}}})$ which is:

$$\gamma_{2|\widehat{\mathcal{X}}}(f_{|\widehat{\mathcal{X}}}) = \inf_p \left\{ ||p||_{L_2(\Theta,\tau)} \middle| f_{|\widehat{\mathcal{X}}}(\mathbf{x}) = \int_\Theta \varphi(\mathbf{x}, \theta) p(\theta) \tau(d\theta)\ \forall \mathbf{x} \in \widehat{\mathcal{X}} \right\}.$$

When minimizing the empirical risk by constraining the norm $\gamma_{2|\widehat{\mathcal{X}}}(f_{|\widehat{\mathcal{X}}}) \leq C$, the optimal solution is defined as:

$$f^*_{\mathcal{F}_2|\widehat{\mathcal{X}},C} = \arg\min_{f \in \mathcal{F}_{2|\widehat{\mathcal{X}}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{y}_i, f(\mathbf{x}_i)) \;\middle|\; \gamma_{2|\widehat{\mathcal{X}}}(f) \leq C \right\}.$$

Learning with the RKHS $\mathcal{F}_2$ can be done through Kernel Methods (7.1), where the kernel is approximated with Random Sampling. In [RR08] is shown that kernels with integral representation such as $K_{\mathcal{F}_2}(\mathbf{x}, \mathbf{x}')$, can be approximated by randomly sampling $m$ times the parameters $\theta$ from the fixed probability distribution $\tau$, to obtain the approximate kernel $\widehat{K}_{\mathcal{F}_2}(\mathbf{x}, \mathbf{x}') = \frac{1}{m} \sum_{i=1}^{m} \varphi(\mathbf{x}, \theta_i) \varphi(\mathbf{x}', \theta_i)$. When the number of random samples $m \to \infty$ the approximate kernel $\widehat{K}_{\mathcal{F}_2}(\mathbf{x}, \mathbf{x}')$ approaches $K_{\mathcal{F}_2}(\mathbf{x}, \mathbf{x}')$. They show that $\sup_{\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d} |\widehat{K}_{\mathcal{F}_2}(\mathbf{x}, \mathbf{x}') - K_{\mathcal{F}_2}(\mathbf{x}, \mathbf{x}')| \leq \varepsilon$ with a high probability, if the number of random samples $m$ is $\mathcal{O}(1/\varepsilon^2)$. This random sampling method gives the opportunity to work in an explicit $m$-dimensional feature space instead of an infinite dimensional space.
Combining the optimal solution from the representer theorem in formula 5 rewritten into ball regularization and the Random Sampling result, the optimal predictor has an explicit representation of the form:

$$f^*_{\mathcal{F}_2|\widehat{\mathcal{X}},C}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K_{\mathcal{F}_2}(\mathbf{x}, \mathbf{x}_i) = \sum_{i=1}^{n} \alpha_i \int_\Theta \varphi(\mathbf{x}, \theta) \varphi(\mathbf{x}_i, \theta) \tau(d\theta) \approx \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m} \varphi(\mathbf{x}, \theta_j) \varphi(\mathbf{x}_i, \theta_j),$$

with $\alpha_i \in \mathbb{R}\ \forall 1 \leq i \leq n$ and $\theta_j \in \Theta\ \forall 1 \leq j \leq m$, sampled from $\tau$. This can be further rewritten into $f^*_{\mathcal{F}_2|\widehat{\mathcal{X}},C}(\mathbf{x}) = \sum_{j=1}^{m} \alpha_j \varphi(\mathbf{x}, \theta_j)$ so for finding the optimal solution, one only needs to find the optimal coefficients $\alpha_j \in \mathbb{R}\ \forall 1 \leq j \leq m$.

**Link to Random Features** In the analysis of two-layer neural networks as Random Features, the input layer weights ($\theta_j = (\mathbf{w}_i, b_j)$ for $1 \leq j \leq m$) are assumed to be fixed at random initialization. Then a linear predictor, parameterized by the output layer weights $a_1, ..., a_m$, is learned over a set of $m$ random features $\mathbf{x} \mapsto \sigma(\mathbf{w}_j^\top \mathbf{x} + b)$. The optimal predictor is the minimizer of:

$$f^*_{\mathrm{RF}|\widehat{\mathcal{X}},\lambda} = \underset{\mathbf{a}\in\mathbb{R}^m}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathbf{y}_i, \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\theta_j^\top \mathbf{x}_i)\right) + \lambda ||\mathbf{a}||_2 \right\},$$

where $\theta_1, ..., \theta_m$ are independent random samples from some probability measure $\tau$. This minimizer is equivalent to $f^*_{\mathcal{F}_{2,\sigma}|\widehat{\mathcal{X}},C}$. They are both of the form $f(\mathbf{x}) = \sum_{j=1}^m \alpha_j \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j)$, with $\theta_j = (\mathbf{w}_j, b_j)$ for $1 \leq j \leq m$ fixed, and optimizing is equal to finding the optimal coefficients $\alpha_j \in \mathbb{R}$ for $1 \leq j \leq m$. In [YS19] is proven that random features can not be used to learn a single ReLU neuron. This link to Random Features confirms both results on $\mathcal{F}_2$, namely that it is easy to optimize and that the function class is too small to capture a lot of functions.

## 5.2 Representer Theorem RKBS B

We again only consider functions $f_{|\widehat{\mathcal{X}}} : \widehat{\mathcal{X}} \to \mathbb{R}$, the space $\mathbf{B}_{|\widehat{\mathcal{X}}}$ consists of all these functions $f_{|\widehat{\mathcal{X}}}$ with finite norm $\gamma_{\mathbf{B}|\widehat{\mathcal{X}}}(f_{|\widehat{\mathcal{X}}})$ which is:

$$\gamma_{\mathbf{B}|\widehat{\mathcal{X}}}(f_{|\widehat{\mathcal{X}}}) = \inf_{\mu\in\mathcal{M}(\Theta)} \left\{ |\mu|(\Theta) \middle| f_{|\widehat{\mathcal{X}}}(\mathbf{x}) = \int_\Theta \varphi(\mathbf{x},\theta)\mu(d\theta) \ \forall \mathbf{x} \in \widehat{\mathcal{X}} \right\}.$$

Similarly as for $\mathcal{F}_2$ in the previous subsection, here for $\mathbf{B}$ we regularize by constraining in a ball, instead of with a penalty term. The empirical risk is minimized by constraining the norm $\gamma_{\mathbf{B}|\widehat{\mathcal{X}}}(f_{|\widehat{\mathcal{X}}}) \leq C$. For the derivation of the representer theorem for $\mathbf{B}$ we use Carathéodory's theorem for the conical hull (5.2).

**Theorem 5.2.** *[Roc97] The conical hull (also known as the cone) of a set $P$ is the set of all its conical combinations (weighted sums), that is cone$(P) = \{\sum_{i=1}^n \alpha_i p_i | p_i \in P, \alpha_i \in \mathbb{R}_{\geq 0}, n \in \mathbb{N}\}$. Carathéodory's theorem for the conical hull states that if a point $\mathbf{a} \in \mathbb{R}^n$ lies in the conical of a set $P$, then $\mathbf{a}$ can be written as the conical combination of at most $n$ points in $P$.*

We have derived the representer theorem for $\mathbf{B}$, given by:

**Theorem 5.3.** *Consider training data $(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n) \in \mathcal{X} \times \mathcal{Y}$ and a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$. The Representer Theorem for $\mathbf{B}$ states that the optimization problem:*

$$f^*_{\mathbf{B}|\widehat{\mathcal{X}},C} = \underset{f_{|\widehat{\mathcal{X}}}\in\mathbf{B}_{|\widehat{\mathcal{X}}}}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(f_{|\widehat{\mathcal{X}}}(\mathbf{x}_i), \mathbf{y}_i) \middle| \gamma_{\mathbf{B}|\widehat{\mathcal{X}}}(f_{|\widehat{\mathcal{X}}}) \leq C \right\},$$

*for any $C > 0$, has a minimizer of the form:*

$$f^*_{\mathbf{B}|\widehat{\mathcal{X}},C}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \varphi(\mathbf{x}, \theta_i).$$

*Proof.* Every function $f_{|\widehat{\mathcal{X}}} \in \mathbf{B}_{|\widehat{\mathcal{X}}}$ lies in the conical hull of $\mathbf{B}$, so is a conical combination of points in $\mathbf{B}$. Points in $\mathbf{B}$ are the basis functions $\varphi(\cdot, \theta) : \mathcal{X} \to \mathbb{C}$ with $\theta \in \Theta$. Since $\widehat{\mathcal{X}}$ is composed of $n$ elements, by Carathéodory's theorem for the conical hull, every function $f_{|\widehat{\mathcal{X}}} \in \mathbf{B}_{|\widehat{\mathcal{X}}}$ can be represented as the conical combination of at most $n$ basis functions $\varphi(\cdot, \theta)$. So also the optimal function $f^*_{\mathbf{B}|\widehat{\mathcal{X}},C} \in \mathbf{B}_{|\widehat{\mathcal{X}}}$ can be represented as the conical combination of at most $n$ basis functions $\varphi(\cdot, \theta)$. We thus also have a representer theorem for $\mathbf{B}$. ∎

This is comparable to the representer theorem for RKHS but the identity of the $n$ basis functions, in our case the parameters $\theta_i$ with $1 \leq i \leq n$, is unknown beforehand. For the RKHS the identity of the $n$ basis functions is the kernel functions evaluated at the training points $\mathbf{x}_1, ..., \mathbf{x}_n$. The only unknowns in the representation of the minimizer $f^*_{\mathcal{F}_2,C}$ are the coefficients $\alpha_1, ..., \alpha_n$. Here the unknowns are both the coefficient and the $n$ functions $\varphi(\cdot, \theta)$, these can be any $n$ points in $\mathbf{B}$. Since functions $f_{|\widehat{\mathcal{X}}}$ can be decomposed in at most basis functions $\varphi(\cdot, \theta)$, the measures $\mu_{|\widehat{\mathcal{X}}}$ parametrizing these basis functions are atomic measures (7.16) and supported on at most $n$ points $\theta$ in $\Theta$. The set of these atomic measures on $\Theta$ is denoted as $\mathcal{M}_{|\widehat{\mathcal{X}}}(\Theta)$. Since the functions $f \in \mathbf{B}$ are completely determined by their measure $\mu$, finding an optimal function $f^*_{\mathbf{B}|\widehat{\mathcal{X}},C}$ is equivalent to finding an optimal measure $\mu^*_{\mathbf{B}|\widehat{\mathcal{X}},C}$ given by:

$$\mu^*_{\mathbf{B}|\widehat{\mathcal{X}},C} = \underset{\mu_{|\widehat{\mathcal{X}}} \in \mathcal{M}_{|\widehat{\mathcal{X}}}(\Theta)}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell \left( \int_\Theta \varphi(\theta, \mathbf{x}_i) \mu_{|\widehat{\mathcal{X}}}(d\theta), \mathbf{y}_i \right) \middle| \inf_{\mu_{|\widehat{\mathcal{X}}} \in \mathcal{M}_{|\widehat{\mathcal{X}}}(\Theta)} \{ |\mu_{|\widehat{\mathcal{X}}}|(\Theta) \} \leq C \right\}.$$

Similarly, every measure $\mu_{|\widehat{\mathcal{X}}} \in \mathcal{M}_{\widehat{\mathcal{X}}}$ lies within the conical hull of $\mathcal{M}(\Theta)$, so is a conical combination of points in $\mathcal{M}(\Theta)$. Points in $\mathcal{M}(\Theta)$ are Dirac measures (7.12) $\delta_\theta$ supported on a single point $\theta \in \Theta$. By Caratheodory's theorem for the conical hull, all measures $\mu_{|\widehat{\mathcal{X}}} \in \mathcal{M}_{\widehat{\mathcal{X}}}$, so also the optimal measure $\mu^*_{\mathbf{B}|\widehat{\mathcal{X}},C}$, can be represented as the conical combination of at most $n$ Dirac measures $\delta_\theta$:

$$\mu^*_{\mathbf{B}|\widehat{\mathcal{X}},C} = \sum_{i=1}^{n} \alpha_i \delta_{\theta_i}.$$

## 5.3 Representer Theorem RKBS $\mathcal{F}_1$

The same representer theorem as for the RKBS $\mathbf{B}$ can be derived for the RKBS $\mathcal{F}_1$. The only difference would be that we consider the smaller space of measures $\mathcal{M}_\tau(\Theta) \subset \mathcal{M}(\Theta)$ to define the space $\mathcal{F}_{1|\widehat{\mathcal{X}}}$ of functions $f_{|\widehat{\mathcal{X}}}$ with finite $\gamma_{1|\widehat{\mathcal{X}}}$-norm:

$$\begin{aligned}
\gamma_{1|\widehat{\mathcal{X}}}(f_{|\widehat{\mathcal{X}}}) &= \inf_{\mu \in \mathcal{M}_\tau(\Theta)} \left\{ |\mu|(\Theta) \middle| f_{|\widehat{\mathcal{X}}}(\mathbf{x}) = \int_\Theta \varphi(\mathbf{x}, \theta) \mu(d\theta) \ \forall \mathbf{x} \in \widehat{\mathcal{X}} \right\} \\
&= \inf_{p} \left\{ ||p||_{L_1(\Theta,\tau)} \middle| f_{|\widehat{\mathcal{X}}}(\mathbf{x}) = \int_\Theta \varphi(\mathbf{x}, \theta) p(\theta) \tau(d\theta) \ \forall \mathbf{x} \in \widehat{\mathcal{X}} \right\}.
\end{aligned}$$

Following the same steps as for $\mathbf{B}$, we have also derived a representer theorem for $\mathcal{F}_1$:

**Theorem 5.4.** *Consider training data $(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n) \in \mathcal{X} \times \mathcal{Y}$ and a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$. The Representer Theorem for $\mathcal{F}_1$ states that the optimization problem:*

$$f^*_{\mathcal{F}_1|\widehat{\mathcal{X}},C} = \underset{f_{|\widehat{\mathcal{X}}} \in \mathcal{F}_{|\widehat{\mathcal{X}}}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f_{|\widehat{\mathcal{X}}}(\mathbf{x}_i), \mathbf{y}_i) \middle| \gamma_{1|\widehat{\mathcal{X}}}(f_{|\widehat{\mathcal{X}}}) \leq C \right\},$$

*for any $C > 0$, has a minimizer of the form:*

$$f^*_{\mathcal{F}_1|\widehat{\mathcal{X}},C}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \varphi(\mathbf{x}, \theta_i).$$

*Proof.* Applying Carathéodory's theorem we have that all measures $\mu_{|\widehat{\mathcal{X}}} \in \mathcal{M}_\tau(\Theta)$, so also the optimal measure $\mu^*_{\mathcal{F}_1|\widehat{\mathcal{X}},C}$, can be represented as the conical combination of at most $n$ Dirac measures $\delta_\theta$. The minimizer $f^*_{\mathcal{F}_1|\widehat{\mathcal{X}},C}$ also has a representation of the form:

$$f^*_{\mathcal{F}_1|\widehat{\mathcal{X}},C}(\mathbf{x}) = \left\langle \sum_{i=1}^{n} \alpha_i \delta_{\theta_i}, \varphi(\mathbf{x}, \cdot) \right\rangle = \sum_{i=1}^{n} \alpha_i \int_\Theta \varphi(\mathbf{x}, \theta) \delta_{\theta_i}(d\theta) = \sum_{i=1}^{n} \alpha_i \varphi(\mathbf{x}, \theta_i).$$

■

For considering a fixed basis function of the form $\varphi(\mathbf{x}, \theta) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$, with $\theta = (\mathbf{w}, b)$, we thus have an optimal solution of the form $f^*_{\mathcal{F}_{1,\sigma}|\widehat{\mathcal{X}},C}(\mathbf{x}) = \sum_{i=1}^n \alpha_i \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i)$. This is the regular form of a two-layer neural network with $n$ neurons. This is not an obvious result. By fixing the reference measure to be a probability measure $\tau$ with full support on $\Theta$ and only considering measures $\mu \in \mathcal{M}_\tau(\Theta)$ that are absolutely continuous with respect to $\tau$, we allow ourself to optimize over a space of strictly absolutely continuous measures. But when considering a finite number of training samples $n$, the optimal measure is a discrete sum of $n$ Dirac measures. And the optimal function is recovered to be of the form of a regular two-layer neural network with $n$ neurons.
Note that the same representation of the optimal function can be obtained by considering a linear combination of the evaluation function of the kernel $K_{\mathcal{F}_1}$

$$f^*_{\mathcal{F}_1|\widehat{\mathcal{X}},C}(\mathbf{x}) = \sum_{i=1}^n \alpha_i K_{\mathcal{F}_1}(\mathbf{x}, \theta_i) = \sum_{i=1}^n \alpha_i \varphi(\mathbf{x}, \theta_i).$$

But here the kernel is not evaluated at $n$ known training points in the input domain but at $n$ unknown points in the parameter space.

# 6 Approximation results

In this section we survey and derive several approximation results for function spaces defined in the sections 3 and 4. We start off with a basic example introducing approximation theory in subsection 6.1. Subsequently we consider known results for approximating functions $f \in \mathbf{B}$ with a bounded $\gamma_{\mathbf{B}}$-norm by functions from $\mathcal{F}_\sigma^m$ and approximating Lipschitz Continuous functions by functions from $\mathbf{B}_\sigma$ in respectively subsections 6.2 and 6.3. In both these sections approximation bounds are given dependent on one of the norms defined in the previous sections 3 and 4. Subsequently we look closer to approximating functions with a known underlying structure, namely functions that depend on projections on a smaller subspace, in 6.4. The Banach spaces $\mathbf{B}_\sigma$ or $\mathcal{F}_{1,\sigma}$ show adaptivity to these underlying structures, the RKHS $\mathcal{F}_{2,\sigma}$ does not. This adaptivity to underlying structures has been mentioned implicitly in [Bar93] and [Bac14] but never explicitly proven with the use of the representation as functions parameterized by measures.

## 6.1 Basic ideas: Approximation Theory and Fourier Analysis

The idea of approximation theory is to approximate functions from a certain function space $\mathcal{G}$ by functions from an often simpler or smaller function space $\mathcal{F}$. The approximation error is the accuracy you lose because you consider functions from a less complex function space to approximate with. When you approximate functions from a function space by functions from the space itself, the approximation error would be 0. The approximation error $\varepsilon$ for approximating $\mathcal{G} \subset \mathbb{R}^{\mathcal{X}}$ with $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ is defined:

$$\forall g \in \mathcal{G}: \quad \exists f \in \mathcal{F} \text{ s.t. } \sup_{\mathbf{x} \in \mathcal{X}} \left\{ \left| g(\mathbf{x}) - f(\mathbf{x}) \right| \right\} \leq \varepsilon.$$

To introduce the concepts we start with an illustrating basic example, the functions we want to approximate are from the Sobolev Space $W^{s,2}(\mathbb{R})$.

**Definition 6.1.** *[Maz85] The One Dimensional Sobolev Space $W^{s,p}(\mathbb{R})$ with $1 \leq p \leq \infty$ is the subset of functions $f \in L_p(\mathbb{R})$ for which the weak derivatives (7.41) up to order $s$ have a finite $L_p$-norm, I.e.:*

$$W^{s,p}(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{R} \;\middle|\; ||f||_{s,p} = \left( \sum_{j=0}^{s} \int_{\mathbb{R}} \left| f^{(j)}(x) \right|^p dx \right)^{1/p} < \infty \right\}.$$

$W^{s,p}(\mathbb{R})$ is a Banach Space, and is complete with respect to the norm $||\cdot||_{s,p}$. For $p=2$, $W^{s,2}(\mathbb{R})$ is a Hilbert Space with inner product $\langle f,g\rangle_{W^{s,2}(\mathbb{R})} = \sum_{i=1}^{s}\langle f^{(i)},g^{(i)}\rangle_{L_2(\mathbb{R})} = \sum_{i=1}^{s}\int_{\mathbb{R}} f^{(i)}(x)g^{(i)}(x)dx$. Functions $f$ in this space can be defined in terms of their Fourier Transform (7.37) $\widehat{f}$, based on the following facts:

1. Parseval's Identity: $||f||_{L_2(\mathbb{R})} = ||\widehat{f}||_{L_2(\mathbb{R})}$, i.e. $\int_{\mathbb{R}}|f(x)|^2\,dx = \int_{\mathbb{R}}|\widehat{f}(\omega)|^2 d\omega$.
2. $\widehat{f'}(\omega) = -i\omega\widehat{f}(\omega)$, and thus $\widehat{f^{(s)}}(\omega) = (-i\omega)^s\widehat{f}(\omega)$.

With this the squared norm $||f||_{s,2}^2$ can be rewritten as:

$$\begin{aligned}||f||_{s,2}^2 &= \sum_{j=0}^{s}\int_{\mathbb{R}}\left|f^{(j)}(x)\right|^2 dx \overset{1}{=} \sum_{j=0}^{s}\int_{\mathbb{R}}\left|\widehat{f^{(j)}}(\omega)\right|^2 d\omega \overset{2}{=} \sum_{j=0}^{s}\int_{\mathbb{R}}\left|(-i\omega)^j\widehat{f}(\omega)\right|^2 d\omega \\ &= \int_{\mathbb{R}}\left(1+|\omega|^2+|\omega|^4+...+|\omega|^{2s}\right)\left|\widehat{f}(\omega)\right|^2 d\omega = \int_{\mathbb{R}}\left(1+|\omega|^2\right)^s\left|\widehat{f}(\omega)\right|^2 d\omega.\end{aligned}$$

The functions we are approximating with are from the space of continuous band-limited functions.

**Definition 6.2.** *[Coh03] The Space of Continuous Band-Limited functions with band-limit $\omega_0$ is given by*

$$\mathcal{H}_{\omega_0}(\mathbb{R}) = \left\{g\in\mathcal{C}(\mathbb{R})\middle|\, support\,(\widehat{g})\in[-\omega_0,\omega_0]\right\}.$$

*This space is a Reproducing Kernel Hilbert Space with Reproducing Kernel $K(x,x') = \frac{\omega_0}{\pi}sinc(\omega_0(x'-x)) = \frac{\sin(a(x'-x))}{\pi(x'-x)}$. Since $K(x,x') = K(x-x')$, it is a Translation Invariant Kernel (7.42).*

Since the Reproducing Kernel $K(x,x')$, with Fourier Transform $\widehat{K}(\omega) = \int_{\mathbb{R}}K(x,x')e^{i\omega x'}dx' = \mathbb{1}_{[-\omega_0,\omega_0]}e^{i\omega x}$ is translation invariant, we will consider Fourier hypothesis spaces. Functions $g\in\mathcal{H}_{\omega_0}(\mathbb{R})$ have a Fourier Transform $\widehat{g}(\omega) = \int_{\mathbb{R}}g(x)e^{-i\omega x}dx$ and by use of the Fourier Inversion theorem from this Fourier Transform the original function can be retrieved as $g(x) = \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0}\widehat{g}(\omega)e^{i\omega x}d\omega$. The reproducing property of the kernel $K(x,x')$ can be obtained as follows:

$$\begin{aligned}g(x) &= \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0}\widehat{g}(\omega)e^{i\omega x}d\omega = \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0}\left(\int_{\mathbb{R}}g(x')e^{-i\omega x'}dx'\right)e^{i\omega x}d\omega = \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0}\left(\int_{\mathbb{R}}g(x')e^{i\omega(x-x')}dx'\right) \\ &= \int_{\mathbb{R}}g(x')\left(\frac{1}{2\pi}\int_{\omega_0}^{\omega_0}e^{i\omega(x-x')}d\omega\right)dx' = \int_{\mathbb{R}}g(x')\left(\frac{1}{2\pi}\left[\frac{1}{i(x-x')}e^{iw(x-x')}\right]_{-\omega_0}^{\omega_0}\right)dx' \\ &= \int_{\mathbb{R}}g(x')\left(\frac{1}{2\pi}\frac{2}{x-x'}\left(\frac{e^{i\omega_0(x-x')}-e^{i\omega_0(x-x')}}{2i}\right)\right)dx' = \int_{\mathbb{R}}g(x')\left(\frac{1}{\pi(x-x')}\sin(\omega_0(x-x'))\right) \\ &= \int_{\mathbb{R}}g(x')K(x,x')dx' = \langle g(\cdot),K(x,\cdot)\rangle_{L^2(\mathbb{R})}.\end{aligned}$$

When we assume the squared Sobolev norm $||f||_{s,2}^2$ of a function $f\in W^{s,2}(\mathbb{R})$ is bounded by a finite constant $C\in\mathbb{R}^+<\infty$, we can derive the approximation error $\varepsilon$ when approximating with functions $g\in\mathcal{H}_{\omega_0}(\mathbb{R})$ as follows:

$$\begin{aligned}\varepsilon &= \underset{g\in\mathcal{H}_{\omega_0}(\mathbb{R})}{\arg\min}\,||f-g||_{L^2}^2 = \underset{g\in\mathcal{H}_{\omega_0}(\mathbb{R})}{\arg\min}\,||\widehat{f}-\widehat{g}||_{L^2}^2 = \underset{g\in\mathcal{H}_{\omega_0}(\mathbb{R})}{\arg\min}\int_{\mathbb{R}}\left|\widehat{f}(\omega)-\widehat{g}(\omega)\right|^2 d\omega \\ &= \int_{\mathbb{R}}\left|\widehat{f}(\omega)-\mathbb{1}_{[-\omega_0,\omega_0]}\widehat{f}(\omega)\right|^2 d\omega = \int_{-\infty}^{-\omega_0}\left|\widehat{f}(\omega)\right|d\omega + \int_{\omega_0}^{\infty}\left|\widehat{f}(\omega)\right|d\omega = 2\int_{\omega_0}^{\infty}\left|\widehat{f}(\omega)\right|d\omega \\ &= 2\int_{\omega_0}^{\infty}\left(1+|\omega|^2\right)^s\left|\widehat{f}(\omega)\right|\frac{1}{(1+|\omega|^2)^s}d\omega \le \frac{2}{(1+|\omega_0|^2)^s}\int_{\omega_0}^{\infty}\left(1+|\omega|^2\right)^s\left|\widehat{f}(\omega)\right|d\omega \le \frac{2\cdot C}{(1+|\omega_0|^2)^s} \\ &= \mathcal{O}\left(\frac{C}{\omega_0^{2s}}\right).\end{aligned}$$

So for an $\varepsilon$-approximation one needs the band-limit $\omega_0 \gtrsim \left(\frac{C}{\varepsilon}\right)^{\frac{1}{2s}}$. A very high $s$ defining functions $f \in W^{s,2}(\mathbb{R})$ would indicate a space of very smooth function, and smooth functions disappear more quickly in the Fourier domain. So as expected, for a higher $s$, a smaller band-limit is necessary to obtain an accurate approximation.

## 6.2 Approximating B by $\mathcal{F}_\sigma^m$

In this section we consider approximating functions $f \in \mathbf{B}$ with a finite $\gamma_{\mathbf{B}}$-norm by functions from $\mathcal{F}_\sigma^m$, specifically a two-layer neural network with $m$ neurons rewritten as in steps 1 to 3 in subsection 4.2, $f_m(\mathbf{z}) = \sum_{i=1}^m a_i \sigma(\theta_i^\top \mathbf{z})$. Throughout this subsection we consider both the input domain $\mathcal{X}$ and the parameter space $\Theta$ to be $\mathbb{R}^{d+1}$. We first present some implicit result mentioned from [Bar93] to illustrate that such results are not new, but were not explicitly noted. Subsequently we present more recent and general approximation results.
In [Bar93] these kind of approximation results were first implicitly mentioned. Functions $f \in L_1(\mathbb{R}^{d+1}) \cap L_2(\mathbb{R}^{d+1})$ were considered to be approximated by finite width sigmoidal activated two-layer neural networks $f_{m,\text{sigmoid}} = \sum_{i=1}^m a_i/(1+e^{-\theta_i^\top \mathbf{z}})$. The Levy Continuity theorem is used to prove uniqueness of the mesaures.

**Theorem 6.3.** *[Wil91] Consider a series of random variables $\{\theta_n\}_{n=1}^\infty$ and the sequence of corresponding characteristic functions $\{\psi\}_{n=1}^\infty$ defined as $\psi_n(\mathbf{z}) = \mathbb{E}[e^{i\theta_n^\top \mathbf{z}}]\ \forall\ \mathbf{z} \in \mathcal{X},\ \forall\ n \in \mathbb{N}$. Then the Levy Continuity Theorem states, if the sequence of characteristic functions converges point-wise to some function $\psi$ as $\lim_{n\to\infty} \psi_n(\mathbf{z}) \to \psi(\mathbf{z})\ \forall\ \mathbf{z} \in \mathcal{X}$ then the following statements are equivalent:*

- *$\theta_n$ converges in some distribution $\mathcal{D}$ to some random variable $\theta$: $\theta_n \xrightarrow{\mathcal{D}} \theta$*
- *$\{\theta_n\}$ is tight: $\lim_{a\to\infty}(\sup_n \mathbb{P}(|\theta_n| > a)) = 0$*
- *$\psi(\mathbf{z})$ is a characteristic function of some random variable $\theta$*
- *$\psi(\mathbf{z})$ is a continuous function of $\mathbf{z}$*
- *$\psi(\mathbf{z})$ is continuous at $\mathbf{z} = 0$*

We consider the Banach space $\mathbf{B}_{c.e.}$, where $c.e.$ indicates the complex exponential fixed basis function $\sigma(u) = e^{iu}$. It is complete with respect to $\gamma_{\mathbf{B}_{\text{c.e.}}}$-norm, which for a function $f$ is given by $\gamma_{\mathbf{B}_{\text{c.e.}}}(f) := \inf_\mu\{|\mu|(\Theta)\big|f(\mathbf{z}) = \int_\Theta e^{i\theta^\top \mathbf{z}}\mu(d\theta)\forall \mathbf{z} \in \mathcal{X}\}$.

**Lemma 6.4.** *For functions $f \in L_1(\mathbb{R}^{d+1})$ the norm $\gamma_{\mathbf{B}_{\text{c.e.}}}(f)$ is finite*

*Proof.* Since $f \in L_1(\mathbb{R}^{d+1})$, they have a Fourier transform and thus can be represented as a linear combination of the fixed basis function $e^{i\theta^\top \mathbf{z}}$ which is from the family of functions $\varphi$. That is, $f(\mathbf{z}) = \int_\Theta e^{i\theta^\top \mathbf{z}}\widehat{f}(\theta)d\theta = \int_\Theta e^{i\theta^\top \mathbf{z}}\mu(d\theta)$ for all $\mathbf{z} \in \mathcal{X}$, where $\widehat{f}$ is the Fourier Transform of $f$ because of the Fourier Inversion Theorem. Here the measure $\mu$ is necessarily unique due to the Levy Continuity Theorem (6.3). Since $f(\mathbf{z}) = \int_\Theta e^{i\theta^\top \mathbf{z}}\mu(d\theta) = \mathbb{E}_{\theta\sim\mu}[e^{i\theta^T \mathbf{z}}]$ we know that $\theta_n$ converges in $\mu$ to the random variable $\theta$. Functions $f$ that can be written in this Fourier representation thus have a finite $\gamma_{\mathbf{B}_{\text{c.e.}}}$-norm that is given by $\gamma_{\mathbf{B}_{\text{c.e.}}} = \int_\Theta |\widehat{f}(\theta)|d\theta$. ■

**Lemma 6.5.** *For functions $f \in L_1(\mathbb{R}^{d+1}) \cap L_2(\mathbb{R}^{d+1})$ the norm $\gamma_{\mathbf{B}_{\text{c.e.}}}(\nabla f)$ is finite*

*Proof.* Since $f \in L_1(\mathbb{R}^{d+1}) \cap L_2(\mathbb{R}^{d+1})$, its gradient has a Fourier representation given by $\nabla f(\mathbf{z}) = \int_\Theta e^{i\theta^\top \mathbf{z}}\widehat{\nabla f}(\theta)d\theta = \int_\Theta e^{i\theta^\top \mathbf{z}} i\theta \widehat{f}(\theta)d\theta = \int_\Theta e^{i\theta^\top \mathbf{z}} i\theta\mu(d\theta)$, where the second equality follows from Parseval's Identity. So the existence of the Fourier representation of $\nabla f$ is equivalent to having a finite $\gamma_{\mathbf{B}_{\text{c.e.}}}(\nabla f)$-norm, which is given by $\gamma_{\mathbf{B}_{\text{c.e.}}}(\nabla f) = \{|i\theta\mu|(\Theta)\big|\nabla f(\mathbf{z}) = \int_\Theta e^{i\theta^\top \mathbf{z}} i\theta\mu(d\theta)\forall \mathbf{z} \in \mathcal{X}\} = \int_\Theta |i\theta\mu|(d\theta) = \int_\Theta ||\theta||\ |\widehat{f}(\theta)|d\theta$. ■

Such functions $f$ with finite $\gamma_{\mathbf{B}_{\text{c.e.}}}(\nabla f)$ norm can be approximated by a finite width sigmoidal activated two-layer neural network $f_{m,\text{sigmoid}}$ such that $||f - f_{m,\text{sigmoid}}||^2_{L_2(\mathbb{R}^{d+1})} = \mathcal{O}\left(\gamma_{\mathbf{B}_{\text{c.e.}}}(\nabla f)^2/m\right)$ [Bar93]. The proof is based on convex combinations in a Hilbert Space. This was the introduction in the literature of such approximation results of the very specific case of sigmoidal activated two-layer neural networks approximating functions with a finite norm on the measure parameterizing the fixed complex exponential basis function.
In [Bac14] these results were generalized by considering arbitrarily activated two-layer neural networks, as long as they are not polynomial as shown in theorem 4.2, approximating functions with a finite norm on the measure parameterizing a non-fixed basis function from the family of functions $\varphi$. Functions $f$ with finite $\gamma_{\mathbf{B}}$-norm can be approximated by a finite width arbitrary activated two-layer neural network $f_m$ such that $||f - f_m||^2_{L_2(\mathbb{R}^{d+1})} = \mathcal{O}\left(\gamma_{\mathbf{B}}(f)^2/m\right)$. The bound is obtained through the conditional gradient algorithm, but has been proved in several other ways ([KS01],[Mha04]). If the activation function $\sigma$ is specified to be ReLU, $\sigma(u) = (u)_+$, this general bound can be slightly improved. Functions $f$ with either a bounded $\gamma_1$- or $\gamma_2$-norm can be approximated by finite width ReLU-activated two-layer neural network $f_{m,\text{ReLU}}$ such that $||f - f_{m,\text{ReLU}}||^2_{L_\infty(\mathbb{R}^{d+1})} = \mathcal{O}\left(\gamma_1^2(f) m^{\frac{-(d+3)}{2d}}\right)$ [Bac14] and $||f - f_{m,\text{ReLU}}||^2_{L_2(\mathbb{R}^{d+1})} = \mathcal{O}\left(\gamma_2^2(f) m^{\frac{-(d+3)}{2d}}\right)$ [Bac15] respectively.

## 6.3 Approximating Lipschitz Continuous functions by $\mathbf{B}_\sigma$

In this section we consider approximating $L$-Lipschitz continuous functions on the Euclidean sphere $g : S^d_{1,2} \to \mathbb{R}$ by functions from $\mathbf{B}_\sigma$, specifically a homogeneously activated two-layer neural network with a bounded $\gamma_{2,\text{homo}}$-norm, rewritten as in steps 1 to 7 in subsection 4.2.
A function $g : S^d_{1,2} \to \mathbb{R}$ is $L$-Lipschitz continuous if for all $\mathbf{x}, \mathbf{x}' \in S^d_{1,2}$ holds: $g(\mathbf{x}) \leq L$, $|g(\mathbf{x}) - g(\mathbf{x}')| \leq L||\mathbf{x} - \mathbf{x}'||_2$. As described in subsection 4.1 functions $f \in \mathbb{C}^{\mathcal{X}}$ with finite $\gamma_{2,\sigma}(f)$ form the RKHS $\mathcal{F}_{2,\sigma}$. We consider the fixed homogeneous basis function $(\theta^\top \mathbf{z})^\alpha_+$ from the family of functions $\varphi$. In this case the reproducing kernel of the RKHS $\mathcal{F}_{2,\text{homo}}$ is given by $K_{\mathcal{F}_{2,\text{homo}}}(\mathbf{z}, \mathbf{z}') = \int_{S^d_{1/R,2}} (\theta^\top \mathbf{z})^\alpha_+ (\theta^\top \mathbf{z}')^\alpha_+ \tau(d\theta)$. Analysis of these kernels is done in a similar way as translation invariant kernels, of which an example is given in section 6.1, through Fourier Analysis. The functions are assumed to be on the sphere $S^d_{1,2} \subset \mathbb{R}^{d+1}$. For example for $d = 1$, $S^1_{1,2}$ is the unit circle in $\mathbb{R}^2$. The unit circle is isomorphic to the segment $[0, 2\pi] \subset \mathbb{R}$ and functions thus can be decomposed into their Fourier Series (7.36). For $d > 1$ Spherical Harmonics are used for the analysis. A bound $C$ is put on the $\gamma_{2,\text{homo}}$-norm of the homogeneous two-layer neural networks. Through Fourier analysis of these kernels, this approximation can be bounded by $||g - f||_{L_\infty(S^d_{1,2})} \leq \Gamma(d, \alpha) L \left(C/L\right)^{1/(\alpha+(d-1)/2)} \log\left(C/L\right)$. Here $\Gamma(\alpha, d)$ is a constant that is only dependent on $\alpha$ and $d$. For the derivation of this bound the extra assumption is necessary that $g$ is even if $\alpha$ is odd and vice versa [Bac14].
As described in subsection 3.3, since $\mathcal{F}_{2,\text{homo}} \subset \mathcal{F}_{1,\text{homo}}$ this approximation result can be transferred when approximating by functions $f : S^d_{1,2} \to \mathbb{R}$ of the form $f(\mathbf{z}) = \int_{S^d_{1/R,2}} (\theta^\top \mathbf{z})^\alpha_+ p(\theta) \tau(d\theta)$ with $\gamma_{1,\text{homo}}(f) \leq C$. We get the same bound $||g - f||_{L_\infty(S^d_{1,2})} \leq \Gamma(d, \alpha) L \left(C/L\right)^{1/(\alpha+(d-1)/2)} \log\left(C/L\right)$, only $C$ is now the bound on the $\gamma_{1,\text{homo}}$-norm instead of the $\gamma_{2,\text{homo}}$-norm [Bac14]. This bound is sub-optimal when approximating by these functions $f \in \mathcal{F}_{1,\text{homo}}$. The above bounds for functions on the sphere $S^d_{1,2}$ can be extended to functions in the entire space $\mathbb{R}^{d+1}$.

## 6.4 Approximation of functions of projections

In this section we consider approximating functions with a known underlying structure, namely functions that depend on the projections on a 1-dimensional subspace. We proof that neural networks show adaptivity to these underlying structures. First the the function of projections are properly defined. Subsequently the key notion on which this adaptivity is based, is explained and proven. Lastly is shown how this key notion can be used to improve the bounds in the previous sections 6.2 and 6.3 when the function to be approximated has an underlying structure. These results were

mentioned implicitly in [Bar93] and [Bac14] but proven here explicitly.
We consider a function $h : \mathbb{R}^{d+1} \to \mathbb{R}$ dependent on a projection defined as:

$$h(\mathbf{z}) \coloneqq g(\mathbf{w}^\top \mathbf{z}) \text{ with } \mathbf{w}, \mathbf{z} \in \mathbb{R}^{d+1}, \tag{12}$$

with $g : \mathbb{R} \to \mathbb{R}$ a univariate function of variable $\mathbf{w}^\top \mathbf{z} = z \in \mathbb{R}$. The improvement of the bounds mentioned above, for these functions of projections, are based on the key notion that the $\gamma_{\mathbf{B}}$-norm of the functions $h$ is not larger than the $\gamma_{\mathbf{B}}$-norm of the function $g$ and hence independent of the input dimension $d$.
Again, in [Bar93] this was implicitly mentioned for the first time, considering a function $h \in L^1(\mathbb{R}^{d+1}) \cap L^2(\mathbb{R}^{d+1})$ and a function $g \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ with respectively $\gamma_{\mathbf{B}_{c.e.}}(\nabla h)$ and $\gamma_{\mathbf{B}_{c.e.}}(\nabla g)$ finite.

**Theorem 6.6.** *[BV07] A measurable function $g$ on $X_2$ is integrable w.r.t. the image measure $m_*\nu$ if and only if the composition $g \circ m$ is integrable w.r.t. the measure $\nu$:*

$$\int_{X_1} (k \circ m)(z)\ \nu(dz) = \int_{X_2} k(\mathbf{z})\ (\nu \circ m^{-1})(d\mathbf{z}),$$

*where the total variation $|\nu|(X_1)$ is non increasing and thus larger or equal than the total variation of the image measure $|\nu \circ m^{-1}|(X_2)$*

**Lemma 6.7.** *For a function of a projection of the form (12), with functions $h \in L^1(\mathbb{R}^{d+1}) \cap L^2(\mathbb{R}^{d+1})$ and a function $g \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, $\gamma_{\mathbf{B}_{c.e.}}(g) \geq \gamma_{\mathbf{B}_{c.e.}}(h)$ and $\gamma_{\mathbf{B}_{c.e.}}(\nabla g) \geq \gamma_{\mathbf{B}_{c.e.}}(\nabla h)$*

*Proof.* For both $h(\mathbf{z}) = \int_{\mathbb{R}^{d+1}} e^{i\theta^\top \mathbf{z}} \mu(d\theta)$ and $g(z) = \int_{\mathbb{R}} e^{i\xi z} \nu(d\xi)$, respectively the measures $\mu$ and $\nu$ are unique due to the Levy Continuity Theorem (6.3). The norm $\gamma_{\mathbf{B}_{c.e.}}(\nabla h)$ is given by $\{|i\theta\mu|(\mathbb{R}^{d+1}) \big| \nabla h(\mathbf{z}) = \int_{\mathbb{R}^{d+1}} e^{i\theta^\top \mathbf{z}} i\theta \mu(d\theta) \forall \mathbf{z} \in \mathbb{R}^{d+1}\} = \int_{\mathbb{R}^{d+1}} ||\theta||\ |\mu|(d\theta)$ and $\gamma_{\mathbf{B}_{c.e.}}(\nabla g) = \{|i\xi\nu|(\mathbb{R}) \big| \nabla(g)(z) = \int_{\mathbb{R}} e^{i\xi z} i\xi \nu(d\xi) \forall z \in \mathbb{R}\} = \int_{\mathbb{R}} |\xi| |\nu|(d\xi)$. Applying theorem 6.6 to rewrite $h$ and $\nabla h$ in the setting by [Bar93] we have:

$$\begin{aligned}
h(\mathbf{z}) \coloneqq g(\mathbf{w}^\top \mathbf{z}) &= \int_{\mathbb{R}} e^{i\xi \mathbf{w}^\top \mathbf{z}} \nu(d\xi) = \int_{\mathbb{R}^{d+1}} e^{i\theta^\top \mathbf{z}} (\nu \circ (\mathbf{w}^\top \mathbf{z})^{-1})(d\theta) = \int_{\mathbb{R}^{d+1}} e^{i\theta^\top \mathbf{z}} \mu(d\theta) \\
\nabla h(\mathbf{z}) \coloneqq \nabla g(\mathbf{w}^\top \mathbf{z}) &= \frac{\partial \mathbf{w}^\top \mathbf{z}}{\partial \mathbf{z}} \frac{\partial g(\mathbf{w}^\top \mathbf{z})}{\partial \mathbf{w}^\top \mathbf{z}} = \mathbf{w} \int_{\mathbb{R}} e^{i\xi \mathbf{w}^\top \mathbf{z}} i\xi \nu(d\xi) = \int_{\mathbb{R}} e^{i\xi \mathbf{w}^\top \mathbf{z}} i\xi \mathbf{w} \nu(d\xi) \\
&= \int_{\mathbb{R}^{d+1}} e^{i\theta^\top \mathbf{z}} i\theta (\nu \circ (\mathbf{w}^\top \mathbf{z})^{-1})(d\theta) = \int_{\mathbb{R}^{d+1}} e^{i\theta^\top \mathbf{z}} i\theta \mu(d\theta),
\end{aligned}$$

where in both cases we have the change of variable $\xi \mathbf{w} = \theta$. The image measure $\nu \circ (\mathbf{w}^\top \mathbf{z})^{-1} = \mu$ is obtained by transferring $\nu$ from the measurable space $\mathbb{R}$ to another measurable space $\mathbb{R}^{d+1}$, mapping $\xi$ to $\xi \mathbf{w} = \theta$. From this we can conclude that both $|\nu|(\mathbb{R}) \geq |\mu|(\mathbb{R}^{d+1})$ and $|i\xi\nu|(\mathbb{R}) \geq |i\theta\mu|(\mathbb{R}^{d+1})$. Since $\nu$ and $\mu$ are necessarily unique, $\gamma_{\mathbf{B}_{c.e.}}(g) \geq \gamma_{\mathbf{B}_{c.e.}}(h)$ and $\gamma_{\mathbf{B}_{c.e.}}(\nabla g) \geq \gamma_{\mathbf{B}_{c.e.}}(\nabla h)$. ■

Since $\gamma_{\mathbf{B}_{c.e.}}(g)$ is independent of dimension $d$, we have an upper-bound on $\gamma_{\mathbf{B}_{c.e.}}(h)$ that is independent on $d$. Since the approximation bounds are a function of the norm $\gamma_{\mathbf{B}_{c.e.}}(h)$, the so called 'Curse of dimensionality' is broken when approximating functions of projections with sigmoidal activated two-layer neural networks.
For a more general setting as in [Bac14], with the only assumption on the functions $h : \mathbb{R}^{d+1} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ that they are in $\mathbf{B}$ and thus the measure characterizing a function is not necessarily unique.

**Lemma 6.8.** *For a function of a projection of the form (12), with $h : \mathbb{R}^{d+1} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ in $\mathbf{B}$, $\gamma_{\mathbf{B}}(h)$ is not larger than $\gamma_{\mathbf{B}}(g)$.*

*Proof.* Since $h$ is in $\mathbf{B}$ there exist a function from the family of functions $\varphi : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \mathbb{R}$ such that $\varphi(\mathbf{z}, \cdot) \in C_0(\mathbb{R}^{d+1}, \mathbb{R})\ \forall \mathbf{z} \in \mathbb{R}^{d+1}$ and at least one measure $\mu$ such that $h(\mathbf{z}) = \int_{\mathbb{R}^{d+1}} \varphi(\mathbf{z}, \theta) \mu(d\theta)$.

Here $\varphi(\mathbf{z},\theta) = \rho(\theta^\top \mathbf{z})$ with a certain function $\rho : \mathbb{R} \to \mathbb{R}$ and thus $h(\mathbf{z}) = \int_{\mathbb{R}^{d+1}} \rho(\theta^\top \mathbf{z})\mu(d\theta)$. Similarly for $g(z) = \int_{\mathbb{R}} \varphi(z,\xi)\nu(d\xi)$ we have at least one measure $\nu$ and a function from the family of functions $\varphi : \mathbb{R} \to \mathbb{R}$ such that $\varphi(z,\cdot) \in C_0(\mathbb{R},\mathbb{R})\ \forall z \in \mathbb{R}$. This $\varphi(z,\xi) = \rho(\xi z)$ with the same function $\rho : \mathbb{R} \to \mathbb{R}$ as used in the construction of $h$ and thus $g(z) = \int_{\mathbb{R}} \rho(\xi z)\nu(d\xi)$. The $\gamma_{\mathbf{B}}$-norm of $h$ and $g$ are given by $\inf_{\mu\in\mathcal{M}(\mathbb{R}^{d+1})}\{|\mu|(\mathbb{R}^{d+1}) \mid h(\mathbf{z}) = \int_{\mathbb{R}^{d+1}} \rho(\theta^\top \mathbf{z})\mu(d\theta)\forall \mathbf{z} \in \mathbb{R}^{d+1}\}$ and $\inf_{\nu\in\mathcal{M}(\mathbb{R})}\{|\nu|(\mathbb{R}) \mid g(z) = \int_{\mathbb{R}} \rho(\xi z)\nu(d\xi)\forall z \in \mathbb{R}\}$ respectively. We take $\varepsilon > 0$ arbitrarely small, by the infimum definition of $\gamma_{\mathbf{B}}$ there exists some measure $\nu$ over $\mathbb{R}$ such that $|\nu|(\mathbb{R}) \leq \gamma_{\mathbf{B}}(g) + \varepsilon$. We again can apply theorem 6.6 with the change of variables $\xi\mathbf{w} = \theta$:

$$h(\mathbf{z}) \coloneqq g(\mathbf{w}^\top \mathbf{z}) = \int_{\mathbb{R}} \rho(\xi \mathbf{w}^\top \mathbf{z})\nu(d\xi) = \int_{\mathbb{R}^{d+1}} \rho(\theta^\top \mathbf{z})(\nu \circ (\mathbf{w}^\top \mathbf{z})^{-1})(d\theta) = \int_{\mathbb{R}^{d+1}} \rho(\theta^\top \mathbf{z})\mu(d\theta).$$

So $|\nu|(\mathbb{R}) \geq |\mu|(\mathbb{R}^{d+1})$. From this we can conclude $\gamma_{\mathbf{B}}(h) \leq |\mu|(\mathbb{R}^{d+1}) \leq |\nu|(\mathbb{R}) \leq \gamma_{\mathbf{B}}(g) + \varepsilon$, so $\gamma_{\mathbf{B}}(h)$ is not larger than $\gamma_{\mathbf{B}}(g)$. ■

Thus we can derive an upper bound for $\gamma_{\mathbf{B}}(h)$ that is independent on the dimension $d$. In a similar way we can derive $\gamma_1(h) \leq \gamma_1(g)$ by only considering a smaller space of measures $\mathcal{M}_\tau(\Theta)$ instead of $\mathcal{M}(\Theta)$, with the norm $||\mu||_{\mathcal{M}_\tau(\Theta)} = ||p||_{L_1(\Theta,\tau)}$. For the RKHS $\mathcal{F}_2$ no such result can be proved, so $\gamma_2(h) \not\leq \gamma_2(g)$. From this follows $\gamma_{1,\sigma}(h) \leq \gamma_{1,\sigma}(g)$ for each fixed basis, characterized by activation function $\sigma$ defined as in subsection 4.1, within the family of functions $\varphi$.
The bound for approximating $L$-Lipschitz continuous functions can be improved when it depends on some underlying structure. A function of projections $h(\mathbf{z}) = g(\mathbf{w}^\top \mathbf{z})$, with $g : \mathbb{R} \to \mathbb{R}$ $L$-Lipschitz, $||\mathbf{w}||_2 < \eta$ and $||\mathbf{z}||_2 \leq \sqrt{2}R$. The function $g$ is thus contained in $[-\sqrt{2}R\eta, \sqrt{2}R\eta] \subset \mathbb{R}$ and on this interval has a greater or equal $\gamma_{1,\text{homo}}$-norm than the function $h : S^d_{1,2} \to \mathbb{R}$. Since by definition $\gamma_{1,\text{homo}}(g) \leq \gamma_{2,\text{homo}}(g)$ we can improve the bound for approximating $L$-Lipschitz continuous functions with functions with a bounded $\gamma_{1,\text{homo}}$-norm.

**Theorem 6.9.** *[Bac14] Every function of a projection $h(\mathbf{z}) = g(\mathbf{w}^\top \mathbf{z}) : S^d_{1,2} \to \mathbb{R}$ with $g : \mathbb{R} \to \mathbb{R}$ an $L$-Lipschitz continuous function and $||\mathbf{w}||_2 \leq \eta$, there exists a function $f : S^d_{1,2} \to \mathbb{R}$ of the form $f(\mathbf{z}) = \int_{S^d_{1/R,2}} (\theta^\top \mathbf{z})^\alpha_+ p(\theta)\tau(d\theta)$ with $\gamma_{1,homo}(f) \leq CR\eta$, such that:*

$$||h - f||_{L_\infty(S^d_{1,2})} \leq \Gamma(\alpha)L\left(\frac{CR\eta}{L}\right)^{\frac{1}{\alpha}} \log\left(\frac{CR\eta}{L}\right),$$

*where $S^d_{1,2}, S^d_{1/R,2} \subset \mathbb{R}^{d+1}$ and $\Gamma(\alpha)$ is a constant only dependent on $\alpha$. For the derivation of this bound the extra assumption is necessary that $h$ is even if $\alpha$ is odd and vice versa.*

Just as in section 6.3 the bound in theorem 6.9 for functions on the sphere $S^d_{1,2}$ can be extended to functions in the entire space $\mathbb{R}^{d+1}$. Again we like to emphasize this bound is initially proven for functions with a bounded $\gamma_{2,\text{homo}}$-norm and extended to functions with a bounded $\gamma_{2,\text{homo}}$-norm because $\mathcal{F}_{2,\text{homo}} \subset \mathcal{F}_{1,\text{homo}}$, and hence $\gamma_{1,\text{homo}}(\cdot) \leq \gamma_{2,\text{homo}}(\cdot)$. To sum up the scheme used to derive the bound in theorem 6.9: $\gamma_{1,\text{homo}}(h) \leq \gamma_{1,\text{homo}}(g) \leq \gamma_{2,\text{homo}}(g) = \mathcal{O}(1)$.

**Approximation of function of multiple projections** The results for function of a single projections can be easily extended to functions of multiple projections, functions $H : \mathbb{R}^{d+1} \to \mathbb{R}$ defined as:

$$H(\mathbf{z}) \coloneqq G(\mathbf{W}^\top \mathbf{z}) \text{ with } \mathbf{W} \in \mathbb{R}^{(d+1)\times s}, \mathbf{z} \in \mathbb{R}^{d+1}, s < d,$$

with $G : \mathbb{R}^s \to \mathbb{R}$ a $s$-variate function of variable $\mathbf{W}^\top \mathbf{z} = \tilde{z} \in \mathbb{R}^s$. The same reasoning applies as in lemma's 6.7 and 6.8 that, through a change of variables, $\gamma_{\mathbf{B}}(H) \leq \gamma_{\mathbf{B}}(G)$ and $\gamma_1(H) \leq \gamma_1(G)$. So when approximating a function of multiple projections $H : S^d_{1,2} \to \mathbb{R}$ the norm $\gamma_{\mathbf{B}}(H)$ becomes instead of dependent on $d$, dependent on $s < d$. Equivalently for a function of multiple projections $H(\mathbf{z}) = G(\mathbf{W}^\top \mathbf{z})$ with $G : \mathbb{R}^s \to \mathbb{R}$, the largest singular value of $\mathbf{W}$ bounded by $\eta$ and $||\mathbf{z}||_2 \leq \sqrt{2}R$ the bound can be improved. The function $G : \mathbb{R}^s \to \mathbb{R}$ is thus contained in $[-\sqrt{2}R\eta, \sqrt{2}R\eta]^s \subset \mathbb{R}^s$ and has a larger or equal $\gamma_{1,\text{homo}}$-norm as the function $H : S^d_{1,2} \to \mathbb{R}$.

**Theorem 6.10.** *[Bac14] For every function of a projection $H(\mathbf{z}) = G(\mathbf{W}^\top \mathbf{z}) : S^d_{1,2} \to \mathbb{R}$ with $G : \mathbb{R}^s \to \mathbb{R}$ an $L$-Lipschitz continuous function, and the largest singular value of $\mathbf{W}$ is bounded by $\eta$, there exists a function $f : S^d_{1,2} \to \mathbb{R}$ of the form $f(\mathbf{z}) = \int_{S^d_{1/R,2}} (\theta^\top \mathbf{z})^\alpha_+ p(\theta)\tau(d\theta)$ with $\gamma_{1,homo}(f) \leq CR\eta$, such that:*

$$||H - f||_{L_\infty(S^d_{1,2})} \leq \Gamma(s,\alpha) L \left(\frac{CR\eta}{L}\right)^{\frac{1}{\alpha + \frac{s-1}{2}}} \log\left(\frac{CR\eta}{L}\right),$$

*where $S^d_{1,2}, S^d_{1/R,2} \subset \mathbb{R}^{d+1}$ and $\Gamma(s,\alpha)$ is a constant only dependent on $s < d$ and $\alpha$. For the derivation of this bound the extra assumption is necessary that $H$ is even if $\alpha$ is odd and vice versa.*

Also this bound can be extended from functions on the sphere $S^d_{1,2}$ to functions on the entire space $\mathbb{R}^{d+1}$.

**Approximating compositional Multi-Index functions** This idea of adaptivity to underlying structures can be extended to multilayer neural networks. This is elaborated in [Sch17] where compositional functions of the form $H(\mathbf{z}) = G_L \circ G_{L-1} \circ ... \circ G_1 \circ G_0(\mathbf{z})$ are approximated by $L$-layer neural networks. The $L$-layer neural network shows approximation benefits up to a compositional function of level $L$. For example this would mean a three-layer neural network shows adaptivity when approximating a compositional multi-index function:

$$H(\mathbf{z}) = G\left(W_G^\top \mathbf{y}\right) \text{ where } \mathbf{y} = \begin{bmatrix} g_1(W_1^\top \mathbf{z}) \\ \vdots \\ g_d(W_d^\top \mathbf{z}) \end{bmatrix} \text{ with } W_1, ...W_d, W_G \in \mathbb{R}^{(d+1)\times s} \text{ and } g_1, ..., g_d, G : \mathbb{R}^s \to \mathbb{R}.$$

The above is not proven yet, if the approximation analysis for neural networks represented as functions parameterized by measures, can be extended to multi-layer neural network is an open question.

# 7 Conclusion

We have discussed a representation of two-layer neural networks in terms of a basis function, the function of a single neuron in the hidden layer, parametrized by a measure. This representation makes the function space of neural networks, of which there is little understanding of the mathematical structure, linear and fit for the construction of different norms and inner products. By fixing a reference probability measure with full support, and only considering measures who are absolutely continuous with respect to this fixed measure, we can distinguish two main spaces of interest. Those are a big Banach space and a subspace which is a Hilbert space. The Banach space is proven to be a Reproducing Kernel Banach Space and the Hilbert space a Reproducing Kernel Hilbert Space, for both a representer theorem and a reproducing kernel is derived. The representer theorems give insight in the inductive bias of two-layer neural networks, by stating the form of the optimal solution when there is a finite number $n$ of training data points. From the representer theorem for the RKBS we can conclude that **even if one considers absolutely continuous measures, corresponding to a potentially infinite wide neural network, the optimal measure is an atomic measure supported on at most $n$ points in the parameter space**. Do note that the identity, the parameters of these $n$ point in the parameter space, is unknown. But this does indicate that the optimal function of neural network has at most $n$ neurons. This result contradicts the empirical findings of the use of overparameterization, which suggest that adding more and more parameters keeps improving the generalization performance. A possible explanation for this behaviour is that this single global minimum neural network function with at most $n$ neurons, is really hard to find exactly. Overparameterizing could possibly make the loss landscape more suitable for local search methods such as gradient descent. This is a hypothesis that would need further research.
**Dependent on the norm put on the function space of neural networks, the approximation properties change significantly**. Considering the RKHS-norm on the space of neural networks, allows us to analyse the neural networks as kernels. They can be optimized with kernel

methods, which gives a clear computational advantage. Finding the optimal solution is equivalent to finding the optimal coefficients in the linear combination of kernel evaluation at the training points. But the Hilbert space is shown to be too small to capture a lot of functions, it does not capture the full behaviour and potential of the space of neural networks. When considering the RKBS-norm on the space of neural networks, the resulting Banach space is able to capture way more functions but loses the computational advantage. In particular when approximating sparse functions, functions that depend on a projection on a lower dimensional space, the Banach Space outperforms the Hilbert space. The Banach space shows adaptivity to such underlying structures and approximation bounds independent of dimension $d$ can be derived, breaking the curse of dimensionality. The Hilbert space does not have this adaptivity. **The difference in structure and approximation performance between the Hilbert and Banach spaces, shows neural networks are indeed significantly different from kernels**.
In this thesis the main focus is on approximation and representation results. Especially the direction of optimization would also be interesting to research for this representation of two-layer neural network as a basis function parametrized by a measure, because we can study convex optimization schemes. Other possible future directions could be to extend these results to more complicated neural network architectures, such as adding more layers or convolutions.

# References


[FW56] Marguerite Frank and Philip Wolfe. "An algorithm for quadratic programming". In: *Naval Research Logistics Quarterly* 3.1-2 (1956), pp. 95–110. URL: https://EconPapers.repec.org/RePEc:wly:navlog:v:3:y:1956:i:1-2:p:95-110.

[DS58] N. Dunford and J.T. Schwartz. *Linear Operators: General theory*. Linear Operators. Interscience Publishers, 1958, pp. 285–305. URL: https://books.google.nl/books?id=DuJQAAAAMAAJ.

[Par67] K.R. Parthasarathy. "II - PROBABILITY MEASURES IN A METRIC SPACE". In: *Probability Measures on Metric Spaces*. Ed. by K.R. Parthasarathy. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press, 1967, pp. 26–55. URL: http://www.sciencedirect.com/science/article/pii/B9781483200224500065.

[Tre67] François Treves. Pure and Applied Mathematics, Vol. 25. Academic Press, 1967. ISBN: 9781483223629.

[Maz85] Vladimir G Maz'ja. Springer Series in Soviet Mathematics. Berlin–Heidelberg–New York: Springer-Verlag, 1985. ISBN: 0-387-13589-8.

[Bou87] Nicolas Bourbaki. *Topological vector spaces*. Elements of mathematics. Springer New York, 1987. ISBN: 978-3-540-13627-9.

[Rud87] Walter Rudin. *Real and Complex Analysis, 3rd Ed.* USA: McGraw-Hill, Inc., 1987. ISBN: 0070542341.

[Wil91] David Williams. *Probability with Martingales*. Cambridge mathematical textbooks. Cambridge University Press, 1991, pp. I–XV, 1–251. ISBN: 978-0-521-40605-5.

[H92] Whitney H. *Analytic Extensions of Differentiable Functions Defined in Closed Sets*. Birkhäuser Boston, 1992. ISBN: 978-1-4612-7740-8.

[Bar93] Andrew R. Barron. "Universal Approximation Bounds for Superpositions of a Sigmoidal Function". In: *930 IEEE TRANSACTIONS ON INFORMAI10N THEORY* 39, NO.3 (May 1993).

[Con94] J.B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 1994. ISBN: 9780387972459. URL: https://books.google.nl/books?id=ix4P1e6AkeIC.

[Bil95] Patrick Billingsley. *Probability and measure*. Wiley series in probability and mathematical statistics. New York:Wiley, 1995. ISBN: 0-471-00710-2.

[Mit97] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[Roc97] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, 1997.

[Pin99] Allan Pinkus. "Approximation theory of the MLP model in neural networks". In: *ACTA NUMERICA* 8 (1999), pp. 143–195.

[KS01] Kurkova and Sanguineti. "Bounds on Rates of Variable-Basis and Neural-Network Approximation". In: (Sept. 2001).

[Sch01] Smola A.J. Schölkopf B. Herbrich R. *A Generalized Representer Theorem*. Springer, Berlin, Heidelberg, 2001. ISBN: 978-3-540-42343-0.

[Coh03] Albert Cohen. *Numerical Analysis of Wavelet Methods*. Studies in Mathematics and Its Applications. 2003.

[Mha04] H. N. Mhaskar. "On the tractability of multivariate integration and approximation by neural networks". In: (2004).

[Lan05] S. Lang. *Algebra*. Graduate Texts in Mathematics. Springer New York, 2005. ISBN: 9780387953854. URL: https://books.google.nl/books?id=Fge-BwqhqIYC.

[BV07] Bogachev and Vladimir. "Measure Theory, Berlin: Springer Verlag, ISBN 9783540345138". In: (2007), Sections 3.6–3.7.

[CT08] de Vito Carmeli and Umanita Toigo. "Vector valued reproducing kernel Hilbert spaces and universality". In: (2008). URL: https://arxiv.org/pdf/0807.1659.pdf.

[RR08] Ali Rahimi and Benjamin Recht. "Random Features for Large-Scale Kernel Machines". In: (2008). Ed. by J. C. Platt et al., pp. 1177–1184. URL: http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines.pdf.

[Tao10] T. Tao. *An Epsilon of Room, I: Real Analysis*. An Epsilon of Room. American Mathematical Society, 2010. ISBN: 9780821852781. URL: https://books.google.nl/books?id=royDAwAAQBAJ.

[KSH12a] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: (2012). Ed. by F. Pereira et al., pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[KSH12b] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Lake Tahoe, Nevada: Curran Associates Inc., 2012, pp. 1097–1105.

[GMH13] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks". In: (2013), pp. 6645–6649. URL: http://dx.doi.org/10.1109/icassp.2013.6638947.

[Tao13] T. Tao. *An Introduction to Measure Theory*. Graduate studies in mathematics. American Mathematical Society, 2013. ISBN: 9781470409227. URL: https://books.google.nl/books?id=SPGJjwEACAAJ.

[Bac14] Francis R. Bach. "Breaking the Curse of Dimensionality with Convex Neural Networks". In: *CoRR* abs/1412.8690 (2014). arXiv: 1412.8690. URL: http://arxiv.org/abs/1412.8690.

[Bac15] Francis R. Bach. "On the Equivalence between Quadrature Rules and Random Features". In: *CoRR* abs/1502.06800 (2015). arXiv: 1502.06800. URL: http://arxiv.org/abs/1502.06800.

[Pol15] David Pollard. *Total variation distance between measures*. Asymptopia. 2015.

[CSV16] Patrick L. Combettes, Saverio Salzo, and Silvia Villa. *Regularized Learning Schemes in Feature Banach Spaces*. 2016. arXiv: 1410.6847 [math.ST].

[MP16] Hrushikesh Mhaskar and Tomaso A. Poggio. "Deep vs. shallow networks : An approximation theory perspective". In: *CoRR* abs/1608.03287 (2016). URL: http://arxiv.org/abs/1608.03287.

[Sil+16] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529 (2016), pp. 484–503. URL: http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html.

[Zha+16] Chiyuan Zhang et al. "Understanding deep learning requires rethinking generalization". In: (2016). Published in ICLR 2017. URL: http://arxiv.org/abs/1611.03530.

[Sch17] Johannes Schmidt-Hieber. *Nonparametric regression using deep neural networks with ReLU activation function*. 2017. arXiv: 1708.06633 [math.ST].

[BMM18] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. "To Understand Deep Learning We Need to Understand Kernel Learning". In: *Proceedings of the 35th International Conference on Machine Learning*. Proceedings of Machine Learning Research. 2018, pp. 541–549. URL: http://proceedings.mlr.press/v80/belkin18a.html.

[Bel+18] Mikhail Belkin et al. *Reconciling modern machine learning practice and the bias-variance trade-off*. 2018. arXiv: 1812.11118 [stat.ML].

[Gun+18] Suriya Gunasekar et al. *Characterizing Implicit Bias in Terms of Optimization Geometry*. 2018. arXiv: 1802.08246 [stat.ML].

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *CoRR* abs/1806.07572 (2018). arXiv: 1806.07572. URL: http://arxiv.org/abs/1806.07572.

[Sch18] Kevin Schlegel. *When is there a Representer Theorem? Nondifferentiable Regularisers and Banach spaces*. 2018. arXiv: 1804.09605 [cs.LG].

[Son18] Phan-Minh Nguyen Song Mei Andrea Montanari. "A Mean Field View of the Landscape of Two-Layers Neural Networks". In: (Aug. 2018).

[AL19] Zeyuan Allen-Zhu and Yuanzhi Li. "What Can ResNet Learn Efficiently, Going Beyond Kernels?" In: *CoRR* abs/1905.10337 (2019). arXiv: 1905.10337. URL: http://arxiv.org/abs/1905.10337.

[DL19] Xialiang Dou and Tengyuan Liang. "Training Neural Networks as Learning Data-adaptive Kernels: Provable Representation and Approximation Benefits". In: (Jan. 2019).

[LZZ19] Rongrong Lin, Haizhang Zhang, and Jun Zhang. "On Reproducing Kernel Banach Spaces: Generic Definitions and Unified Framework of Constructions". In: *CoRR* abs/1901.01002 (2019). arXiv: 1901.01002. URL: http://arxiv.org/abs/1901.01002.

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA: PMLR, 25–28 Jun 2019, pp. 2388–2464. URL: http://proceedings.mlr.press/v99/mei19a.html.

[Uns19] Michael Unser. *A unifying representer theorem for inverse problems and machine learning*. 2019. arXiv: 1903.00687 [math.OC].

[YS19] Gilad Yehudai and Ohad Shamir. "On the Power and Limitations of Random Features for Understanding Neural Networks". In: *CoRR* abs/1904.00687 (2019). URL: http://arxiv.org/abs/1904.00687.

[LZB20] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. "Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning". In: (2020). arXiv: 2003.00307 [cs.LG].

[PN20] Rahul Parhi and Robert D. Nowak. *Neural Networks, Ridge Splines, and TV Regularization in the Radon Domain*. 2020. arXiv: 2006.05626 [stat.ML].

[Tou+20] Hugo Touvron et al. *Fixing the train-test resolution discrepancy: FixEfficientNet*. 2020. arXiv: 2003.08237 [cs.CV].

# Appendix A: Mathematical Facts

**Definition 7.1.** *[RR08] Consider training data* $(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n) \in \mathcal{X} \times \mathcal{Y}$. *Let* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a kernel and* $\mathcal{H}$ *the corresponding Reproducing Kernel Hilbert Space (3.4) with inner product* $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, *Reproducing Kernel* $K$ *and RKHS-norm* $|| \cdot ||_{\mathcal{H}}$. *A Kernel Method is a discrimination rule of the form, for* $\lambda > 0$*:*

$$\widehat{f} = \underset{f \in \mathcal{H}}{\arg\min} \left\{ L\left(f(\mathbf{x}_1), \mathbf{y}_1, ..., f(\mathbf{x}_n), \mathbf{y}_n\right) + M(||f||_{\mathcal{H}}) \right\},$$

*where* $L$ *only depends on the function* $f$ *through* $f(\mathbf{x}_1), ..., f(\mathbf{x}_n)$ *and labels* $\mathbf{y}_i, ..., \mathbf{y}_n$, *and* $M$ *is non-decreasing*

**Definition 7.2.** *[Con94] A Locally Compact Second Countable Space* $X$ *satisfies:*

- $X$ *is a Hausdorff (7.3) topological space*
- *For all* $x \in X$ *there is an open set* $V$ *and a compact set* $K$ *such that* $x \in V \subset K$
- *There is a countable family* $\mathcal{V} = (V_i)_{i \in \mathbb{N}}$ *of open subsets of* $X$ *such that every open subset of* $X$ *can be written as the union of elements of some subfamily of* $\mathcal{V}$

**Definition 7.3.** *[Con94] In a Hausdorff space any two distinct points have disjoint neighbourhoods.*

**Definition 7.4.** *[Con94] The Borel* $\sigma$*-algebra* $\Sigma(X)$ *of a space* $X$ *is the smallest* $\sigma$*- algebra containing all open sets*

**Definition 7.5.** *[Tao13] The tuple* $(X, \Sigma(X))$ *is a Measurable Space, if* $X$ *a set and* $\Sigma(X)$ *a* $\sigma$*-algebra over* $X$.

**Definition 7.6.** *[Tao13] Let* $(X, \Sigma(X))$ *be a Measurable Space, than* $E \in \Sigma(X)$ *is a Measurable Set.*

**Definition 7.7.** *[Tao13] A Measure on a Measurable Space* $(X, \Sigma(X))$ *is a function* $\mu : \Sigma(X) \to [0, \infty]$, *that satisfies:*

- *Null empty set:* $\mu(\emptyset) = 0$, *where* $\emptyset$ *denotes an empty set*
- *Countable additivity:* $\mu\left(\bigcup_{k-1}^{\infty} E_k\right) = \sum_{i=1}^{\infty} \mu(E_k)$, *holds for all countable collections of pairwise disjoint sets in* $\Sigma(X)$*:* $\{E_i\}_{i=1}^{\infty}$

*I.e. a measure is function that assigns a real number to all Measurable Sets in a Measurable Space.*

**Definition 7.8.** *[Tao13] The triple* $(X, \Sigma(X), \mu)$, *is a Measure Space if* $X$ *is a set,* $\Sigma(X)$ *a* $\sigma$*-algebra on* $X$ *and* $\mu : \Sigma(X) \to \mathbb{R}^+$ *is a measure on* $(X, \Sigma(X))$. *An example of a measure space is the Euclidean space* $\mathbb{R}^d$, *the collection* $\Sigma(X) = \mathcal{L}[\mathbb{R}^d]$ *of all Lebesgue measurable subsets of* $\mathbb{R}^d$ *and* $\mu(E)$ *the d-dimensional Lebesgue measure.*

**Definition 7.9.** *[Tao13] A function* $f : X \to \mathbb{R}$, *with free variable* $x$, *is Measurable in a Measure Space* $(X, \Sigma(X), \mu)$ *when it is integrable with respect to the measure, i.e.* $\int_X f(x)\mu(dx)$ *is finite.*

**Definition 7.10.** *[Con94] A complex-valued regular Borel measures* $\mu$ *on a measurable space* $(X, \Sigma(X))$ *is a map* $\mu : \Sigma(X) \to \mathbb{C}$ *such that:*

- $\mu(\emptyset) = 0$
- $\mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$ *for all countable families* $(E_i)_{i \in \mathbb{N}}$ *of disjoint sets belong to* $\Sigma(X)$ *the series converges unconditionally*

- *Regularity condition:* $\mu(K) < \infty$ *for all compact subsets* $K$ *of* $X$

**Definition 7.11.** *[Con94] If* $\mu$ *is a measure on* $(X, \Sigma(X))$ *and* $E \in \Sigma(X)$*, the variation norm of the measure is defined as:*

$$|\mu|(E) = \sup\left\{\sum_{i=1}^{m} |\mu(E_i)| \middle| \{E_i\}_{i=1}^{m} \text{ is a measurable partition of } E\right\}.$$

**Definition 7.12.** *[Tao13] A Dirac Measure* $\delta_a$ *of a set* $E$ *is defined as* $\delta_a(E) = \mathbb{1}_E(a)$*. Thus the measure* $\delta_a$ *of a set* $E$ *is* 1 *if it contains* $a$ *and* 0 *otherwise.*

**Definition 7.13.** *[Bil95] A Probability measure* $\tau$ *is a real-valued measure on a Measurable Space* $(\Omega, \mathcal{F})$ *that forms a Probability space* $(\Omega, \mathcal{F}, \tau)$*. Here* $\Omega$ *is the set of all possible outcomes,* $\mathcal{F}$ *all possible events (set of outcomes). A probability measure* $\tau$ *returns a value between 0 and 1 and the values assigned to entire set add up to 1.*

**Definition 7.14.** *[Bou87] The Support of a non-negative measure* $\mu : \Sigma(X) \to \mathbb{R}^+$ *on a measurable space* $(X, \mathcal{B}(X))$ *is the subset of* $\Sigma(X)$ *for which holds:* $supp(\mu) := \left\{\overline{E \in \Sigma(X) : \mu(E) > 0}\right\}$

**Definition 7.15.** *[Bil95] A measure* $\mu$ *is Discrete with respect to a measure* $\nu$*, both defined on the measurable space* $(X, \Sigma(X))$*, if there is a at most countable subset* $S$ *of* $X$*. For this* $S$ *all subsets are measurable,* $\nu(S) = 0$ *and* $\mu(X \setminus S) = 0$*. I.e. the measure* $\mu$ *is concentrated on an at most countable set.*

**Definition 7.16.** *[Bil95] A measure* $\mu$ *on a measurable space* $(X, \mathcal{B}(X))$ *is Atomic if for a subset* $A \subset X$*, the measure assigns a value* $\mu(A) > 0$*. And for any measurable set* $B \supset A$ *with* $\mu(B) < \mu(A)$*,* $\mu(B)$ *is automatically equal to 0.*

**Definition 7.17.** *[Bil95] A measure* $\mu$ *is Absolutely Continuous with respect to a measure* $\tau$*, both defined on the measurable space* $(X, \Sigma(X))$*, if for every measurable set* $E \in \Sigma(X)$*,* $\tau(E) = 0$ *implies* $\mu(0)$*. In words: the measure* $\mu$ *is dominated by the measure* $\tau$*.*

**Definition 7.18.** *[Bil95] When there are two measures,* $\mu$ *and* $\tau$*, that are both defined on the measurable space* $(X, \Sigma(X))$*, and* $\mu$ *is absolutely continuous with respect to* $\tau$*, there exists a measurable functions* $p : X \to \mathbb{R}^+$*, such that for all measurable sets* $E \in \Sigma(X)$

$$\mu(E) = \int_E p(x)\tau(dx).$$

*The function* $p(x) = \mu(dx)/\tau(dx)$ *is the Density (Radon-Nikodym derivative) of a measure* $\mu$ *with respect to the measure* $\tau$

**Definition 7.19.** *[Bou87] A space of functions* $\mathcal{F}$ *from any fixed set* $X$ *to a field* $F$ *(e.g.* $\mathbb{R}^d \to \mathbb{R}$*) is a Linear Function Space (also Vector Function Space) if the following properties hold for any* $f, g \in \mathcal{F}$ *and any* $a \in F$*:*

- *Addition:* $(f + g) \in \mathcal{F}$
- *Scalar multiplication:* $a \cdot f \in \mathcal{F}$

**Definition 7.20.** *[Bou87] A Topological Linear Function Space is a Linear Function Space (7.19) a with suitable topology (a structure that allows the notion of some elements being close to or far away from each other). Suitable meaning that the addition and scalar multiplication are both a continuous map (continuous function between two topological spaces). The field* $F$ *thus needs to be a topological field,* $\mathbb{R}$ *is an example of this.*

**Definition 7.21.** *[Lan05] A Complete Topological Linear Function Space is a Topological Linear Function Space (7.20) in which any Cauchy sequence (sequence whose elements become arbitrarily close to each over as the sequence advances) has a limit, i.e. all limits are contained within the space.*

**Definition 7.22.** *[Tre67] A Banach Space $X$ is a Complete Topological Linear Function Space (7.21) with a norm $||\cdot||_X$ as topology, i.e. a complete normed linear space.*

**Definition 7.23.** *[Tre67] A Hilbert Space is a Complete Topological Linear Function Space (7.21) with the norm induced by an inner product $\langle f, g\rangle$. All Hilbert Spaces are thus by definition also a Banach space with a norm defined as $||f|| = \sqrt{\langle f, f\rangle}$*

**Definition 7.24.** *[Con94] Function have a Compact Support on a topological space $X$ if its closed support is a compact subset of $X$. Compact meaning it is closed (all limit point are within the space) and bounded (of finite size).*

**Definition 7.25.** *[Tao10] The Topological Dual Space of a topological linear function space $V$ (7.19) consists of all continuous linear forms on $V$, together with structure of addition and scalar multiplication. The topological dual is a subspace of the dual space, which consists out of all linear linear forms on $v$.*

**Definition 7.26.** *[Con94] A bilinear form is a function $B : V \times W \to K$ that is linear in each argument separately, such that for all $v_1, v_2 \in V$, all $w_1, w_2 \in W$ and all $C \in K$ holds:*

- $B(v_1 + v_2, w) = B(v_1, w) + B(v_2, w)$
- $B(C \cdot v, w) = C \cdot B(v, w)$
- $B(v, w_1 + w_2) = B(v, w_1) + B(v, w_2)$
- $B(v, C \cdot w) = C \cdot B(v, w)$

**Definition 7.27.** *[Tao10] The Canonical Pairing of a topological vector space $X$ and its topological dual space $X^*$ is the bilinear form between an element $\mathbf{x} \in X$ and $\mathbf{x}' \in X^*$ denoted as:*

$$\langle \mathbf{x}, \mathbf{x}' \rangle = \mathbf{x}'(\mathbf{x}).$$

**Definition 7.28.** *[Con94] A Reflexive Banach Space is a Banach Space $\mathcal{B}$ with dual space $\mathcal{B}^*$ and bidual space $(\mathcal{B}^*)^*$ such that:*

- *The evaluation map $J : \mathcal{B} \to (\mathcal{B}^*)^*$ is surjective*
- *The evaluation map $J : \mathcal{B} \to (\mathcal{B}^*)^*$ is an isometric isomorphism*
- *The evaluation map $J : \mathcal{B} \to (\mathcal{B}^*)^*$ is an isomorphism*

**Definition 7.29.** *[Con94] Consider vector space $X$ and $Y$ over a field $\mathbb{K}$ and $b : X \times Y \to \mathbb{K}$ a bilinear form. Denote $\forall \mathbf{x} \in X$, $b(\mathbf{x}, \cdot) : Y \to \mathbb{K}$ as the linear functional $\mathbf{y} \mapsto b(\mathbf{x}, \mathbf{y})$ and $\forall \mathbf{y} \in Y$, $b(\cdot, \mathbf{y}) : X \to \mathbb{K}$ as the linear functional $\mathbf{x} \mapsto b(\mathbf{x}, \mathbf{y})$. The Weak Topology on $X$ induced by $Y$ is the coarsest topology, the topology with the fewest open sets, for which all maps $b(\cdot, \mathbf{y}) : X \to \mathbb{K}$ remain continuous. The Weak Topology on $Y$ induced by $X$ is the coarsest topology, the topology with the fewest open sets, for which all maps $b(\mathbf{x}, \cdot) : Y \to \mathbb{K}$ remain continuous.*
*If $Y$ is the topological dual of $X$, $X^*$, the weak topology on $X$ is the weak topology on $X$ with respect to the (7.27) $\langle \cdot, \cdot \rangle : X \times X^* \to \mathbb{K}$.*
*The weak topology is characterized by weak convergence, i.e. for every $\mathbf{x} \in X$ there exists a sequence $\mathbf{x}_n \in X$ such that $\langle \mathbf{x}_n, \mathbf{x}' \rangle \to \langle \mathbf{x}, \mathbf{x}' \rangle$ as $n \to \infty$ for all $\mathbf{x}' \in X^*$.*

**Definition 7.30.** *[Con94] If $X^*$ is the topological dual of a topological vector space $X$, the Weak* topology on $X^*$ is the weak topology (7.29) on $X^*$ with respect to the (7.27) $\langle \cdot, \cdot \rangle : X \times X^* \to \mathbb{K}$.*
*The weak* topology is characterized by weak* convergence, i.e. for every $\mathbf{x}' \in X^*$ there exists a sequence $\mathbf{x}'_n \in X^*$ such that $\langle \mathbf{x}, \mathbf{x}'_n \rangle \to \langle \mathbf{x}, \mathbf{x}' \rangle$ as $n \to \infty$ for all $\mathbf{x} \in X$.*

**Definition 7.31.** *[Con94] If $\phi : V \to W$ is a map over a domain $V$, then the Image or range of $V$ under $\phi$ is defined as all values in $W$ $\phi$ can assume as its input varies over $V$:*

$$Im(\phi) = \phi(V) = \{f(v) | v \in V\}.$$

**Definition 7.32.** *[Con94] A mapping $\phi : V \to W$ is Injective if for all $a, b \in V$, $\phi(a) = phi(b)$ implies $a = b$. Equivalently $a \neq b$ $\phi(a) \neq \phi(b)$*

**Definition 7.33.** *[Con94] The Kernel or null space of a linear mapping, is the set of vectors in the domain of the mapping which are mapped to the zero vector. For a linear map $\phi : V \to W$ the kernel is given by*

$$Ker(\phi) = \{v \in V | \phi(v) = \mathbf{0}\},$$

*where $\mathbf{0}$ denotes the zero vector.*

**Definition 7.34.** *[Con94] A pairing of sets $V$ and $W$ Bijection if:*

1. *Each element $v \in V$ is paired with at least one element $w \in W$*
2. *No element $v \in V$ is paired with more than one element $w \in W$*
3. *Each element $w \in W$ is paired with at least one element $v \in V$*
4. *No element $w \in W$ is paired with more than one element $v \in V$*

**Definition 7.35.** *[Con94] Consider a vector space $V$ over a field $K$ and a subspace $W \subset V$. An equivalence relation $\sim$ can be defined stating $x \sim y$ if $x - y \in W$. The equivalence class of $x$ is denoted as $[x] = x + W$. The Quotient $V/W$ is the set of all equivalence classes over $V$ by $\sim$.*

**Definition 7.36.** *[Rud87] The Fourier Series for a function $f : \mathbb{R} \to \mathbb{C}$, is defined as:*

$$f_N(x) = \sum_{n=-N}^{N} c_n \cdot e^{inx},$$
$$with\ c_n = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-inx} dx.$$

**Definition 7.37.** *[Rud87] The Fourier Transform as a function $f : \mathbb{R} \to \mathbb{C}$, is defined as:*

$$\widehat{f}(\omega) = \int_{\mathbb{R}} f(x) e^{-i\omega x} dx.$$

*And the Fourier Inversion theorem states that the function $f$ can be retrieved from the fourier transform $\widehat{f}$:*

$$f(x) = \int_{\mathbb{R}} \widehat{f}(\omega) e^{i\omega x} d\omega.$$

**Definition 7.38.** *[LZZ19] A Continuous bilinear form between two normed vector spaces $V_1, V_2$ is a function $\langle \cdot, \cdot \rangle_{V_1, V_2}$ from $V_1 \times V_2$ to $\mathbb{C}$ that is linear to both arguments and for which a constant $C \in \mathbb{R}^+$ exists such that:*

$$| \langle \cdot, \cdot \rangle_{V_1 \times V_2} | \leq C ||f||_{V_1} ||g||_{V_2} \ \forall f \in V_1, g \in V_2.$$

**Definition 7.39.** *[LZZ19] The linear span of a set $A \subseteq \mathcal{W}_1$ is dense in $\mathcal{W}_1$ with respect to the bilinear form $\langle \cdot, \cdot \rangle_{\mathcal{W}_1 \times \mathcal{W}_2}$ if for any $v \in \mathcal{W}_2$:*

$$\langle a, v \rangle_{\mathcal{W}_1 \times \mathcal{W}_2} = 0 \ \forall a \in A,$$

*implies $v = 0$.*

**Definition 7.40.** *[Rud87] Jensen Inequality states the following: If $(X, \Sigma(X), \mu)$ is a probability space, $f$ a convex function on $\mathbb{R}$ and $g$ is measurable on $(X, \Sigma(X), \mu)$, the following holds:*

$$f\left( \int_X g(x) \mu(dx) \right) \leq \int_X (f \circ g)(x) \mu(dx).$$

**Definition 7.41.** *[Maz85] If $f$ is a function in $L^1([a,b])$, than $g \in L^1([a,b])$ is a Weak Derivative of $f$ if:*

$$\int_a^b f(x)h'(x)dx = -\int_a^b g(x)h(x)dx,$$

*for all infinitely differentiable function $h$ for which $h(a) = h(b) = 0$.*

**Definition 7.42.** *[RR08] A Translation Invariant Kernel is a kernel for which holds:*

$$K(x,x') = K(x-x') = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{K}(\omega)e^{i\omega(x-x')},$$

*the Fourier domain is natural for translation invariant kernels.*

# Appendix B: Proof RKHS $\mathcal{F}_2$ with RKHS-norm $\gamma_2$

**Theorem 7.43.** *The set $\mathcal{F}_2 = \{f \in \mathbb{C}^{\mathcal{X}} : \gamma_2(f) < \infty\}$ where $\gamma_2^2(f)$ is the infimum $\int_\Theta |p(\theta)|^2\tau(d\theta)$ over all decomposition of $f(\mathbf{x}) = \int_\Theta \varphi(\mathbf{x},\theta)p(\theta)\tau(d\theta)$, is an RKHS with corresponding RKHS-norm $\gamma_2$ and kernel $K_{\mathcal{F}_2}(\mathbf{x},\mathbf{x}') = \int_\Theta \varphi(\mathbf{x},\theta)\varphi(\mathbf{x}',\theta)\tau(d\theta)$ for any compact space $\Theta$.*

*Proof.* We introduce the linear mapping $\psi : L_2(\Theta,\tau) \to \mathcal{F}_2$ by $\psi(p)(\mathbf{x}) = \int_\Theta \varphi(\mathbf{x},\theta)p(\theta)\tau(d\theta)$ with kernel $\mathrm{Ker}(\psi)$. From this we obtain a bijection $U$ from the orthogonal complement $\mathrm{Ker}(\psi)^\perp$ to $\mathcal{F}_2$. We define $\mathcal{F}_2$-dotproduct as $\langle f,g\rangle_{\mathcal{F}_2} = \int_\Theta (U^{-1}f)(\theta)(U^{-1}g)(\theta)\tau(d\theta)$. First we prove this defines an RKHS with corresponding kernel $K_{\mathcal{F}_2}(\mathbf{x},\mathbf{x}')$:

$$\begin{aligned}
\forall \mathbf{x}' \in \mathcal{X} \text{ we have} &: K_{\mathcal{F}_2}(\cdot,\mathbf{x}') \in \mathcal{F}_2 \\
&: p = U^{-1}K_{\mathcal{F}_2}(\cdot,\mathbf{x}') \in \mathrm{Ker}(\psi)^\perp \\
&: q : \theta \to \varphi(\mathbf{x}',\theta) \\
&: p - q \in \mathrm{Ker}(\psi).
\end{aligned}$$

$$\begin{aligned}
\text{Thus } \langle f, K_{\mathcal{F}_2}(\cdot,\mathbf{x}')\rangle_{\mathcal{F}_2} &= \int_\Theta (U^{-1}f)(\theta)(U^{-1}K_{\mathcal{F}_2}(\cdot,\mathbf{x}'))(\theta)\tau(d\theta) \\
&= \int_\Theta (U^{-1}f)(\theta)p(\theta)\tau(d\theta) \\
&= \int_\Theta (U^{-1}f)(\theta)q(\theta)\tau(d\theta) \\
&= \int_\Theta (U^{-1}f)(\theta)\varphi(\theta,\mathbf{x}')\tau(d\theta) \\
&= \psi(U^{-1}f)(\mathbf{x}') \\
&= f(\mathbf{x}').
\end{aligned}$$

Secondly we prove the RKHS-norm we have defined is $\gamma_2$.

$$\begin{aligned}
\forall f \in \mathcal{F}_2 \text{ s.t } f = \psi(p) \text{ for } p \in L_2(\Theta,\tau) \text{ we have:} \\
p = U^{-1}f + q \text{ where } q \in \mathrm{Ker}(\psi).
\end{aligned}$$

$$\begin{aligned}
\text{Thus } \int_\Theta |p(\theta)|^2\tau(d\theta) &= ||p||^2_{L_2(\Theta,\tau)} \\
&= ||U^{-1}f||^2_{L_2(\Theta,\tau)} + ||q||^2_{L_2(\Theta,\tau)} \\
&= ||f||^2_{\mathcal{F}_2} + ||q||^2_{L_2((\Theta,\tau))}.
\end{aligned}$$

$$\begin{aligned}
\text{This implies } \int_\Theta |p(\theta)|^2\tau(d\theta) &\geq ||f||^2_{\mathcal{F}_2} \text{ with equality if and only if } q = 0 \\
\text{This shows } \gamma_2^2(f) &= ||f||^2_{\mathcal{F}_2}\ .
\end{aligned}$$

■

# Appendix C: Two related Regularization methods

Consider training data $(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n)$ and we want to minimize the Empirical Risk $\widehat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{y}_i, f(\mathbf{x}_i))$ for functions from an infinite dimensional space $\mathcal{F}$ by a penalty term $\phi : \mathcal{F} \to \mathbb{R}$, e.g. for $L_2 \subset \mathcal{F}$ the penalty $M(f) = ||f||_2^2$. We define:

$$f_\lambda^* = \arg\min_{f \in \mathcal{F}} \left\{ \widehat{L}(f) + \lambda M(f) \right\}$$

$$f_C^* = \arg\min_{f \in \mathcal{F}: M(f) \leq C} \left\{ \widehat{L}(f) \right\}.$$

**Theorem 7.44.** *Family of solutions derived by regularization with penalty parameter $\lambda > 0$ is included in the family of solutions derived by regularization through constraining within the ball with parameter $C > 0$, i.e. $\{f_\lambda^* \mid \lambda > 0\} \subseteq \{f_C^* \mid C > 0\}$. This means for every value of $\lambda > 0$ there exists a value of $C > 0$ such that $f_\lambda^* = f_C^*$.*

*Proof.* By contradiction: By definition $f_\lambda^* = \arg\min_{f \in \mathcal{F}} \left\{ \widehat{L}(f) + \lambda M(f) \right\}$. Take $C = M(f_\lambda^*)$. We assume there exists a function $f$ such that $M(f) \leq C = M(f_\lambda^*)$ and we know that always holds $\widehat{L}(f) < \widehat{L}(f_\lambda^*)$. Than we would have $\widehat{L}(f) + (f) < \widehat{L}(f_\lambda^*) + \lambda M(f_\lambda^*)$ which is a contradiction since $\widehat{L}(f_\lambda^*) + \lambda M(f_\lambda^*)$ is by definition the minimum. ■