Smooth Operators

Steffen Grünewälder STEFFEN@CS.UCL.AC.UK
Arthur Gretton[†] ARTHUR.GRETTON@GMAIL.COM
John Shawe-Taylor JST@CS.UCL.AC.UK
Computer Science and [†]Gatsby Unit, CSML, University College London, UK

Abstract

We develop a generic approach to form smooth versions of basic mathematical operations like multiplication, composition, change of measure, and conditional expectation, among others. Operations which result in functions outside the reproducing kernel Hilbert space (such as the product of two RKHS functions) are approximated via a natural cost function, such that the solution is guaranteed to be in the targeted RKHS. This approximation problem is reduced to a regression problem using an adjoint trick, and solved in a vector-valued RKHS, consisting of continuous, linear, smooth operators which map from an input, real-valued RKHS to the desired target RKHS. Important constraints, such as an almost everywhere positive density, can be enforced or approximated naturally in this framework, using convex constraints on the operators. Finally, smooth operators can be composed to accomplish more complex machine learning tasks, such as the sum rule and kernelized approximate Bayesian inference, where state-of-the-art convergence rates are obtained.

1. Motivation

One of the important ideas that make functional analysis a powerful tool in all branches of mathematics is that basic mathematical operations, like multiplication or composition, may be represented and studied with linear operators. Multiplication fg, for example, is for a fixed f a linear operation in g, and under suitable restrictions, this operation can be described with the help of a bounded linear operator $\mathbf{M}_f$, i.e. $\mathbf{M}_f g = fg$.

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

The study of such basic operations in reproducing kernel Hilbert spaces (RKHSs, Aronszajn (1950); Berlinet & Thomas-Agnan (2004)) suffers from a crucial difficulty: these spaces are not closed under many such operations. For example, if we consider an RKHS $\mathcal{H}_X$ and two functions $f, g \in \mathcal{H}_X$ then in most cases fg will not lie in $\mathcal{H}_X$. This simple fact has far reaching consequences, both for theoretical and practical problems, as one cannot simply apply basic mathematical operations on functions in the RKHS and expect to obtain an RKHS function. In many practical problems, for example, the reproducing property is of major importance in keeping computation costs at bay, and to avoid dealing explicitly with high dimensional feature spaces. To each RKHS $\mathcal{H}_X$ there corresponds an associated reproducing kernel $k(x, y)$, and the reproducing property states that $f(x) = \langle f, k(x, \cdot)\rangle_k$ for any function f from $\mathcal{H}_X$. Since the product of two RKHS functions is likely not in $\mathcal{H}_X$, however, the reproducing property will not hold for this product.

Our main contribution is a way to address these difficulties by approximating linear operators such as $\mathbf{M}_f$ with operators $\mathbf{F}_f : \mathcal{H}_X \to \mathcal{H}_X$ that map back into the RKHS $\mathcal{H}_X$. We will refer to such operators as *smooth operators*. By smooth we mean in a broad sense RKHS functions with low norm. The intuition is that an RKHS-norm is a measure of smoothness very similar to a Sobolev-norm, which measures (weak) derivatives of functions and calculates the norm based on how large these derivatives are. The operator $\mathbf{F}_f$ preserves smoothness in this sense, but we also model $\mathbf{F}_f$ itself as an element of a more complex RKHS.

This more complex RKHS is one of the key tools in the paper. It is based on a vector-valued kernel function $\boldsymbol{\Xi}(f, g)$, where $f, g \in \mathcal{H}_X$. The importance of this kernel is that the corresponding RKHS $\mathcal{H}_{\boldsymbol{\Xi}}$ consists only of bounded linear operators mapping from $\mathcal{H}_X$ to a second RKHS $\mathcal{H}_Y$ (in the case of a product of functions, we have the special case $\mathcal{H}_Y = \mathcal{H}_X$). This vector-valued kernel is in the simplest case a subset of

the Hilbert-Schmidt operators. We will make use of well established vector-valued RKHS tools to approximate and estimate operators like $\mathbf{M}_f$.

It turns out that for the intuitive risk functions in many settings, an adjoint trick is useful to make estimation tractable. Typically, we have an expression of the form $(\mathbf{F}h)(x)$, where $h \in \mathcal{H}_Y$, and we want to separate h from $\mathbf{F}$ (recall that our goal is to estimate $\mathbf{F}$, which is assessed by its action on some test function h evaluated at x). The trick is simple: as $\mathbf{F}$ is a bounded linear operator, there exists an adjoint operator $\mathbf{F}^*$ with which we can transform the term

$$(\mathbf{F}h)(x) = \langle \mathbf{F}h, k(x,\cdot)\rangle_k = \langle h, \mathbf{F}^* k(x,\cdot)\rangle_l,$$

with l being the kernel of $\mathcal{H}_Y$; thus h is separated from $\mathbf{F}$. We prove there exists a natural adjoint kernel $\mathbf{\Xi}^*$ for $\mathbf{\Xi}$ such that $\mathbf{F}^* \in \mathcal{H}_{\mathbf{\Xi}^*}$ iff $\mathbf{F} \in \mathcal{H}_{\mathbf{\Xi}}$. This is important as we gain explicit control over the adjoint and the link between $\mathbf{F}$ and $\mathbf{F}^*$.

We can view this move to the adjoint operator as transforming our learning problem from one of estimating an operator $\mathbf{F}$ to that of estimating a mapping into an RKHS, $x \mapsto \mathbf{F}^* k(x,.)$, which can be viewed as a regression problem. We are thus able to obtain $\mathbf{F}$ through standard regression techniques. Since this is couched in the general RKHS framework, it can be applied to a very general class of mappings and applications. Our results show that these estimation problems are tractable both algorithmically and statistically.

Besides the problem of learning smooth approximations of non-smooth functions, an important application of smooth operators is in integration theory. Basic integrals of RKHS functions are studied with the help of *mean embeddings* (Berlinet and Thomas-Agnan, 2004; Smola, Gretton, Song, and Schölkopf, 2007; Sriperumbudur, Gretton, Fukumizu, Lanckriet, and Schölkopf, 2010). These mean embeddings are representer $m_X \in \mathcal{H}_X$ of an integral or expectation, in that the expectation over an RKHS function $f \in \mathcal{H}_X$ can be efficiently calculated as $\mathbb{E}f = \langle m_X, f\rangle_k$. Integration theory itself is a field rich in sophisticated methods to transform integrals for all sorts of practical problems. We focus here on two such transformations: the change of measure rule, and conditional expectations. We show these can be approached within the operator framework, and produce sample based estimates for these transformations which do not leave the underlying RKHSs. The covariate shift problem (Huang, Smola, Gretton, Borgwardt, and Schölkopf, 2007; Gretton, Smola, Huang, Schmittfull, Borgwardt, and Schölkopf, 2009; Yu and Szepesvari, 2012) is closely related to the change of measure transformation, and our conditional expectation approach follows up on the work of Song, Huang, Smola, and Fukumizu (2009); Grünewälder, Lever, Baldassarre, Patterson, Gretton, and Pontil (2012a).

The Radon-Nikodým theorem often allows us to reduce a change of measure transformation to a multiplication: an integral of a function f over a changed measure reduces to an integral of the product f with a Radon-Nikodým derivative r over the original measure. This problem is close to that of learning a multiplication operator $\mathbf{M}_r$, however a Radon-Nikodým derivative is almost everywhere positive. Constraints of this form occur often and are difficult to enforce. If we consider the space L^2 with inner product $\langle f, g\rangle_{L^2} = \int fg$, and a multiplication operator $\mathbf{M}_r$ with $r \in L^2$, then r is a.e. positive when the multiplication operator $\mathbf{M}_r$ is positive; that is, if $\langle \mathbf{M}_r f, f\rangle_{L^2} = \int rf^2 \geq 0$ for all square integrable f. The important point is that positivity of $\mathbf{M}_r$ can be enforced by a convex constraint, illustrating the broader principle that difficult constraints can in certain cases be replaced or approximated with convex constraints on the operators.

Finally, we consider the problem of combining basic operations to perform more complex operations. Key applications of conditional expectations and changes of measure include the sum rule for marginalising out a random variable in a multivariate distribution (Song, Huang, Smola, and Fukumizu, 2009), and kernel-based approximations to Bayes' rule for inference without parametric models (Fukumizu, Song, and Gretton, 2011; Song, Fukumizu, and Gretton, 2013). We show that these problems can be addressed naturally with smooth operators. In particular, the development of estimators is considerably simplified: we derive natural estimators for both rules in a few lines, by first transforming the relevant integrals and then approximating these transformations with estimated operators. This is a significant shortening of the derivation of an estimator when performing approximate Bayesian inference, albeit at the expense of a non-vanishing bias.

We give a brief overview of the sum rule approach. The task is to estimate the expected value of a function h wrt. a measure $\mathbb{Q}_Y$ that is unobserved. We observe $\mathbb{Q}_X$, a second measure $\mathbb{P}_{X\times Y}$, and we know that the conditional measures are equal, i.e. $\mathbb{P}_{Y|x} = \mathbb{Q}_{Y|x}$. It is easy to obtain the quantity $\mathbb{E}_{\mathbb{Q}_Y} h$ from these observed measures, via the integral transformations

$$\mathbb{E}_{\mathbb{Q}_Y} h = \mathbb{E}_{\mathbb{Q}_X} \mathbb{E}_{\mathbb{Q}_{Y|x}} h = \mathbb{E}_{\mathbb{Q}_X} \mathbb{E}_{\mathbb{P}_{Y|x}} h.$$

We can approximate the two operations on the right, i.e. the expectations $\mathbb{E}_{\mathbb{Q}_X}$ and $\mathbb{E}_{\mathbb{P}_{Y|x}}$, with operators. The advantage of the approach is that the two operators can be composed together, since the approximation of $\mathbb{E}_{\mathbb{P}_{Y|x}}$ maps back into the relevant RKHS.

Our approach to composition of operators has another advantage: the error of the composite operation is bounded by the errors of the basic operations that are combined. We demonstrate this on the sum rule and on the kernel Bayes' rule, by bounding the risk of the estimators via the risk of the conditional expectations, means, and approximation errors, which are easily estimated. We show in the case of the sum rule that these bounds can yield state-of-the-art convergence rates.

The problems that can be addressed with our approach have direct practical application. Besides covariate shift and Bayesian inference as discussed above, additional applications include spectral methods for inference in hidden Markov models, and reinforcement learning (Song et al., 2010; Grünewälder et al., 2012b; Nishiyama et al., 2012).

We like to think that the main text of this paper is readable with a basic knowledge of functional analysis and scalar valued RKHS theory. Obviously, we also use techniques from the vector-valued RKHS literature, however this is kept to a minimum in the main text, and the reader can go a long way with the concrete form of the kernel $\mathbf{\Xi}$ from eq. 3, and treat terms of the form $\|\mathbf{F}\|_{\mathbf{\Xi}}$ by analogy with the scalar case $\|f\|_k$. In the supplement, a basic understanding of vector-valued RKHSs is needed. Excellent introductions to this topic are Micchelli and Pontil (2005); Carmeli, De Vito, and Toigo (2006).

2. Smooth Operators

We begin by introducing a natural risk function and a generic way of minimising it to motivate the approach. We then introduce the operator valued kernel and its adjoint. For the purposes of illustration, we apply this basic approach to the multiplication, composition and quotient (Suppl. A.3) operations.

2.1. A Natural Risk Function

Assume we have a linear operator $\mathbf{G}$, acting on functions h from an RKHS $\mathcal{H}_Y$ with kernel $l(y, y')$ and mapping to some function space $\mathcal{F}$, which we want to approximate with an operator $\mathbf{F} \in \mathcal{H}_{\mathbf{\Xi}}$ mapping from $\mathcal{H}_Y$ to $\mathcal{H}_X$. We first need to define in which sense we want to approximate $\mathbf{G}$. A natural choice is to consider the actions of $\mathbf{G}$ and $\mathbf{F}$ on elements h, and minimise the difference between the two, i.e. to minimise the error $((\mathbf{F}h)(x) - (\mathbf{G}h)(x))^2$. There are two free variables here, x and h. An intuitive choice is now to average the error over x wrt. a suitable measure and to take the supremum over $\|h\|_l \leq 1$ to be robust against the worst case h. The corresponding risk function, which we call the natural risk, is

$$\sup_{\|h\|_l \leq 1} \mathbb{E}_X((\mathbf{F}h)(x) - (\mathbf{G}h)(x))^2.$$

2.2. A Generic Approach

This natural risk has the disadvantage that h can be a rather complicated object, and optimising over all possible h is difficult. We can transform the problem into a simpler problem, however. As we will see, there often exists an operator $\mathbf{X}$ acting directly on data x and mapping to $\mathcal{H}_Y$ such that

$$(\mathbf{G}h)(x) = \langle h, \mathbf{X}(x)\rangle_l. \tag{1}$$

(we will provide examples shortly). Furthermore, as $\mathbf{F}$ is in $\mathcal{H}_{\mathbf{\Xi}}$, we can use the adjoint trick to transform $(\mathbf{F}h)(x) = \langle h, \mathbf{F}^* k(x, \cdot)\rangle_l$. Applying both transformations to the natural risk gives us

$$\begin{aligned}&\sup_{\|h\|_l \leq 1} \mathbb{E}_X((\mathbf{F}h)(x) - (\mathbf{G}h)(x))^2 \\ &= \sup_{\|h\|_l \leq 1} \mathbb{E}_X \langle h, \mathbf{F}^* k(x, \cdot) - \mathbf{X}(x)\rangle_l^2.\end{aligned}$$

We still have h in the equation, but it is separated from $\mathbf{F}^*$. Applying Cauchy-Schwarz removes h altogether,

$$\mathbb{E}_X \|\mathbf{F}^* k(x, \cdot) - \mathbf{X}(x)\|_l^2.$$

This is an upper bound for the natural risk which contains no supremum, but only observable quantities that depend on the data x. The objective is still difficult to optimise, as we may not easily be able to compute the expectation $\mathbb{E}_X$. We fall back on a sampling approach, and replace $\mathbb{E}_X$ with a finite sample estimate. We further add a regulariser that penalizes the complexity of $\mathbf{F}^*$, to guarantee solutions that are robust in sparsely sampled regions. This gives us a vector-valued regression problem,

$$\sum_{i=1}^n \|\mathbf{F}^* k(x_i, \cdot) - \mathbf{X}(x_i)\|_l^2 + \lambda \|\mathbf{F}^*\|_{\mathbf{\Xi}^*}^2,$$

where $\lambda \in [0, \infty[$ is the regularisation parameter and $\{x_i\}_{i=1}^n$ a sample from the underlying probability measure. The minimiser of this problem is known to be

$$\mathbf{F}^* f = \sum_{i,j=1}^n f(x_i) \mathbf{W}_{ij} \mathbf{X}(x_j), \tag{2}$$

with $\mathbf{W} = (\mathbf{K} + \lambda \mathbf{I})^{-1}$ and $\mathbf{K}$ the kernel matrix, in case that the kernels $\mathbf{\Xi}$ from (3) below are used with $\mathbf{A}$ and $\mathbf{B}$ being the identities (Micchelli & Pontil, 2005).

We have a one-to-one relation between operators in $\mathcal{H}_{\mathbf{\Xi}}$ and their adjoints by Theorem 2.3 below, so we

extract $\mathbf{F}$ together with $\mathbf{F}^*$. In summary, the recipe to approximate the operator $\mathbf{G}$ is extremely simple: find a transformation $\mathbf{X}$ and use the adjoint of the corresponding estimator in (2). There remains an important question, however: How tight is the upper bound? While in general this bound is not tight, the minima of the upper bound and the natural risk are often related (see Supplement A.1).

2.3. An RKHS of Bounded Linear Operators

We now develop the necessary mathematical tools for the smooth operator approach. The first step is to define a vector-valued kernel $\mathbf{\Xi}$, such that the corresponding RKHS $\mathcal{H}_{\mathbf{\Xi}}$ consists of linear bounded operators between $\mathcal{H}_X$ and $\mathcal{H}_Y$. A suitable choice is

$$\mathbf{\Xi}(f, g) := \langle f, \mathbf{A}g\rangle_k \mathbf{B}, \tag{3}$$

where $\mathbf{A} \in L(\mathcal{H}_X), \mathbf{B} \in L(\mathcal{H}_Y)$ are positive, self-adjoint operators. The most important case is where $\mathbf{A}$ and $\mathbf{B}$ are the identities.

As in the case of scalar kernels, there exist point evaluators that are closely related to the kernel. These are $\mathbf{\Xi}_f[h]$, where $\mathbf{\Xi}_f : \mathcal{H}_Y \to \mathcal{H}_{\mathbf{\Xi}}$ with $\langle \mathbf{F}, \mathbf{\Xi}_f[h]\rangle_{\mathbf{\Xi}} = \langle \mathbf{F}f, h\rangle_l$ (see Micchelli & Pontil (2005)[Sec. 2]). These point evaluators have a natural interpretation as a tensor product in case that $\mathbf{A}$ and $\mathbf{B}$ are the identities; that is, $\mathbf{\Xi}_f[h] = h \otimes f$. We have in this case that $\langle h, \mathbf{\Xi}(f, g)u\rangle_l = \langle \mathbf{\Xi}_f[h], \mathbf{\Xi}_g[u]\rangle_{\mathbf{\Xi}} = \langle h \otimes f, u \otimes g\rangle_{\mathrm{HS}}$.

The theorems we prove hold for the general form in eq. 3, as long as all the scalar kernels used are bounded, e.g. $\sup_{x\in X} k(x, x) < \infty$. In the applications we restrict ourselves for ease of exposition to the case that $\mathbf{A}$ and $\mathbf{B}$ are the identities. Finally, we often need to integrate scalar valued RKHS functions, and we assume that these integrals are well defined (Supp. F).

In Carmeli et al. (2006)[Prop.1] a criterion is given which, if fulfilled, guarantees that a vector-valued RKHS exists with $\mathbf{\Xi}$ as its reproducing kernel. It is easy to verify this criterion applies, and that $\mathbf{\Xi}$ has an associated RKHS $\mathcal{H}_{\mathbf{\Xi}}$ (see Supp. A.2). The importance of this space is that it consists of bounded linear operators. A standard tensor product argument shows that $\mathcal{H}_{\mathbf{\Xi}}$ is a subset of the Hilbert-Schmidt operators in case that $\mathbf{A}$ and $\mathbf{B}$ are the identities.

Corollary 2.1. *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are the identities then* $\mathcal{H}_{\mathbf{\Xi}} \subset$ HS *and the inner products are equal.*

In the general case we still have:

Theorem 2.1 (Proof in supplement, p. 11)**.** *Each* $\mathbf{F} \in \mathcal{H}_{\mathbf{\Xi}}$ *is a bounded linear operator from* $\mathcal{H}_X$ *to* $\mathcal{H}_Y$*.*

Another useful fact about this RKHS is that all $\mathbf{F}$ are uniquely defined by the values $\mathbf{F}k(x, \cdot)$.

Theorem 2.2 (Proof in supp., p. 11)**.** *If for* $\mathbf{F}, \mathbf{G} \in \mathcal{H}_{\mathbf{\Xi}}$ *and all* $x \in X$ *it holds that* $\mathbf{F}k(x, \cdot) = \mathbf{G}k(x, \cdot)$ *then* $\mathbf{F} = \mathbf{G}$*. Furthermore, if* $k(x, \cdot)$ *is continuous in* x *then it is sufficient that* $\mathbf{F}k(x, \cdot) = \mathbf{G}k(x, \cdot)$ *on a dense subset of* X*.*

2.4. Adjoint Kernels and Operators

We now define an adjoint kernel $\mathbf{\Xi}^*(h, u) = \langle h, \mathbf{B}u\rangle_l \mathbf{A}$ for $\mathbf{\Xi}$. Here $l(y, y')$ denotes the kernel corresponding to $\mathcal{H}_Y$, and $\langle \cdot, \cdot\rangle_l$ is the $\mathcal{H}_Y$ inner product. With the same argument as for $\mathbf{\Xi}$ we show $\mathbf{\Xi}^*$ is a kernel with an associated RKHS $\mathcal{H}_{\mathbf{\Xi}^*}$ such that each element of $\mathcal{H}_{\mathbf{\Xi}^*}$ is a bounded linear operator from $\mathcal{H}_Y$ to $\mathcal{H}_X$. The following theorem is important for the adjoint trick.

Theorem 2.3 (Proof in supp., p. 12)**.** *For every* $\mathbf{F} \in \mathcal{H}_{\mathbf{\Xi}}$ *there exists an adjoint* $\mathbf{F}^*$ *in* $\mathcal{H}_{\mathbf{\Xi}^*}$ *such that for all* $f \in \mathcal{H}_X$ *and* $h \in \mathcal{H}_Y$

$$\langle \mathbf{F}f, h\rangle_l = \langle f, \mathbf{F}^*h\rangle_k.$$

In particular, we have for $\mathbf{F}f = \sum_{i=1}^n \mathbf{\Xi}_{f_i}[h_i](f) = \sum_{i=1}^n \langle f, \mathbf{A}f_i\rangle_k \mathbf{B}h_i$ *that the adjoint is*

$$(\mathrm{T}\mathbf{F})h = \mathbf{F}^*h = \sum_{i=1}^n \mathbf{\Xi}^*_{h_i}[f_i](h) = \sum_{i=1}^n \langle h, \mathbf{B}h_i\rangle_l \mathbf{A}f_i.$$

The operator $\mathrm{T}\mathbf{F} = \mathbf{F}^*$ *is an isometric isomorphism from* $\mathcal{H}_{\mathbf{\Xi}}$ *to* $\mathcal{H}_{\mathbf{\Xi}^*}$ *(*$\mathcal{H}_{\mathbf{\Xi}} \cong \mathcal{H}_{\mathbf{\Xi}^*}$ *and* $\|\mathbf{F}\|_{\mathbf{\Xi}} = \|\mathbf{F}^*\|_{\mathbf{\Xi}^*}$*).*

2.5. Constraints

As in the introductory example, it is usually known that the operation we estimate fulfills certain properties, like being symmetric in the sense that

$$\langle \mathbf{F}f, g\rangle_k = \langle f, \mathbf{F}g\rangle_k,$$

and one might want to have an estimate that shares this property of self-adjointness with $\mathbf{F}$.

In the case of operators acting on L^2, certain properties can be enforced by imposing convex constraints. We mentioned already the *a.e. positive Radon-Nikodým derivative* in the introduction, which can be enforced by a positivity constraint on the operator. *Symmetry* of an operation can be enforced by a linear constraint on the corresponding operator, to make the operator self-adjoint. Enforcing a *multiplication operator* is very similar to this case, as every bounded multiplication operator is self-adjoint and every self-adjoint operator is a multiplication operator in a suitable coordinate system, due to the spectral theorem. Self-adjointness might therefor be used as an easy to optimise proxy constraint. Other examples are *expectation operators*, which can be difficult to learn due to the required normalisation. Convex constraints can

be used to guarantee that the inferred operator represents an integral, however. This is similar to the positivity constraint discussed before: we have $\mathbf{F}f \geq 0$ for all positive continuous f iff there exists a (Radon-)measure μ such that $\mathbf{F}f = \int f d\mu$ under suitable conditions. This is the Riesz representation theorem for linear functionals (Fremlin, 2003)[436J].

The same constraints can be applied in the RKHS setting, although a real-valued RKHS is usually a proper subset of L^2 and this can weaken the implications. Quantifying this effect is a major piece of work on its own. Here, we illustrate on an example the relation between self-adjointness and linear constraints:

Theorem 2.4 (Proof in supp., p. 13)**.** *The set of self-adjoint operators in $\mathcal{H}_{\mathbf{\Xi}}$ is a closed linear subspace.*

2.6. Smooth Multiplication Operators

We demonstrate our approach on the example from the introduction by approximating the multiplication operator $\mathbf{G}g = fg$ with a smooth operator $\mathbf{M}_f : \mathcal{H}_X \to \mathcal{H}_X$, where $g \in \mathcal{H}_X$ and f is an arbitrary function. As noted in the introduction, fg is not in the RKHS even for $f \in \mathcal{H}_X$: in this case, the product $fg = \langle f \otimes g, \Psi(x)\rangle_{\mathrm{HS}}$ is a linear operation in the tensor feature space $\Psi(x) := k(x,\cdot) \otimes k(x,\cdot)$ with the standard Hilbert-Schmidt inner product, which corresponds to the RKHS with the squared kernel (Steinwart & Christmann, 2008, Theorem 7.25).

We apply the generic approach from Section 2.2, where in eq. 1 we use the mapping $\mathbf{X}(x) := f(x)k(x,\cdot)$, which is in $\mathcal{H}_X$ for a given x as required. An approximation $\mathbf{M}_f$ of $\mathbf{G}$ can now be gained from eq. 2 by moving from the adjoint $\mathbf{M}_f^*$ in eq. 2 to $\mathbf{M}_f$,

$$\mathbf{M}_f g = \sum_{i,j=1}^{n} f(x_j) g(x_j) \mathbf{W}_{ij} k(x_i,\cdot).$$

This is an intuitive solution: f and g are multiplied on our sample points x_j and this product is interpolated with the help of $k(x_i,\cdot)$. Indeed, it is the solution of the scalar-valued ridge regression,

$$\min_{q\in\mathcal{H}_X} \sum_{i=1}^{n} \left(f(x_i)g(x_i) - q(x_i)\right)^2 + \lambda \|q\|_k^2.$$

Returning to our setting from the introduction: if we wish to take the inner product of this approximation with a new function $h \in \mathcal{H}_X$, we get

$$\text{“}\langle fg, h\rangle_k\text{”} \approx \langle \mathbf{M}_f g, h\rangle_k = \sum_{i,j=1}^{n} f(x_j) g(x_j) \mathbf{W}_{ij} h(x_i).$$

It would further be useful to constrain the estimate either to be a multiplication operator or to be self-adjoint. In this case no closed form solution is available, and a numerical optimisation is needed.

2.7. Smooth Composition Operators

Assume we have given a function $\phi : X \to Y$, a function $h \in \mathcal{H}_Y$, and we want a smooth approximation of $\mathbf{G}h = h \circ \phi$ with Φh, where $\Phi \in \mathcal{H}_{\mathbf{\Xi}}$ maps from $\mathcal{H}_Y$ to $\mathcal{H}_X$. We again use the relation of eq. 1, where this time $\mathbf{X}(x) := l(\phi(x),\cdot)$, which is in $\mathcal{H}_Y$ for a given x. We then get the approximation

$$\Phi h = \sum_{i,j=1}^{n} h(\phi(x_j)) \mathbf{W}_{ij} k(x_i,\cdot).$$

3. RKHS Integration Theory: Basic Transformations

We discuss the change of measure rule and conditional expectations. The supplementary material contains a discussion of products and the Fubini theorem.

3.1. Covariate Shift: Ch. of Meas. on X

A standard integral transformation is the change of measure: given a measure $\mathbb{P}$ and a measure $\mathbb{Q}$ that is absolute continuous wrt. $\mathbb{P}$ ($\mathbb{Q} \ll \mathbb{P}$) there exists a Radon-Nikodým derivative r such that $\mathbb{E}_{\mathbb{Q}} f = \mathbb{E}_{\mathbb{P}} f \times r$. As in the multiplication case we have in general no guarantee that $f \times r$ is in $\mathcal{H}_X$, and it is useful to have an approximation $\mathbf{R}f$ that maps to $\mathcal{H}_X$. Furthermore, we do not know r, and we need to work with data. A potential risk function is $\sup_{\|f\|_k \leq 1} (\mathbb{E}_{\mathbb{Q}} f - \mathbb{E}_{\mathbb{P}} \mathbf{R} f)^2$, and a first optimisation approach would be to replace expectations with empirical expectations and minimize wrt. $\mathbf{R}$,

$$\sup_{\|f\|_k \leq 1} \Big(\sum_{j=1}^{m} \langle f, k(y_j,\cdot)\rangle_k - \sum_{i=1}^{n} \langle \mathbf{R}f, k(x_i,\cdot)\rangle_k \Big)^2$$
$$\leq \| \sum_{j=1}^{m} k(y_j,\cdot) - \mathbf{R}^* \sum_{i=1}^{n} k(x_i,\cdot) \|_k^2, \tag{4}$$

where $\{y_j\}_{j=1}^m$ is a sample from $\mathbb{Q}$ and $\{x_i\}_{i=1}^n$ from $\mathbb{P}$. The following $\mathbf{R}^*$ makes both errors zero,

$$\mathbf{R}^* = \frac{1}{\|m_{\mathbb{P}}\|^2} \langle m_{\mathbb{P}}, \cdot \rangle_k m_{\mathbb{Q}}, \quad \mathbf{R}^* m_{\mathbb{P}} = m_{\mathbb{Q}},$$

where $m_{\mathbb{P}} = \sum_{i=1}^n k(x_i,\cdot)$ and $m_{\mathbb{Q}} = \sum_{i=1}^m k(x'_i,\cdot)$. This is the minimum norm solution which fits both sides exactly (Micchelli & Pontil, 2005)[Th. 3.1].

The approach differs from our generic approach since we have no expectation in the risk function over which

the error is averaged. Instead, we have an interpolation problem. This interpolation transforms $\mathbb{P}$ completely to $\mathbb{Q}$, which can be interpreted as overfitting. There are at least two points where we can improve matters. First, $\mathbf{R}$ does not necessarily represent a multiplication, and constraints can be used to enforce this, or to enforce self-adjointness of $\mathbf{R}$, which is easier. Second, we do not verify the absolute continuity condition. If the measures are not absolutely continuous then it is not possible to transform one measure into the other by a multiplication operator. We further discuss absolute continuity in Suppl. C.1.1.

A heuristic to solve the constrained problem is to estimate a Radon-Nikodým derivative r from data and then, in a second step, to approximate the multiplication with an operator $\mathbf{R}$ to guarantee that $\mathbf{R}f \in \mathcal{H}_X$. There are several possible ways to estimate such a function. In Huang et al. (2007); Gretton et al. (2009); Yu & Szepesvari (2012) a quadratic program is given to estimate a weight vector β with non-negative entries, such that the following cost function is minimised, $\| \sum_{j=1}^m k(y_j, \cdot) - \sum_{i=1}^n \beta_i k(x_i, \cdot)\|_k$. This is eq. 4 with β instead of $\mathbf{R}^*$.

We can interpolate these β_i's with a non-negative function r if the x_i are disjoint. Applying the unconstrained multiplication estimate from Sec. 2.6 to $r \times f$ gives us the change-of-measure operator

$$\mathbf{R}f = \sum_{i,j=1}^n \beta_i f(x_i) \mathbf{W}_{ij} k(x_j, \cdot).$$

3.2. Conditional Expectation

Kernel-based approximations to conditional expectations have been widely studied, and their links with vector-valued regression are established (Song et al., 2009; Grünewälder et al., 2012a). The conditional expectation estimate introduced in these works can be represented by a vector-valued function $\mu : X \to \mathcal{H}_Y$. The approximation is $\mathbb{E}[h|x] \approx \langle h, \mu(x)\rangle_l$. Now, in line with our earlier reasoning, we can define a smooth operator $\mathbf{E}$ to represent the operation. To define such an operator, it is useful to treat the conditional expectation as an operator on h, i.e. $(h \mapsto \mathbb{E}[h|x])$.

By using our natural cost function and applying Jensen's inequality, we gain an upper bound that is very similar to the one in the generic case,

$$\begin{aligned}\mathcal{E}_c[\mathbf{E}] &:= \sup_{\|h\|_l \leq 1} \mathbb{E}_X(\mathbb{E}[h|x] - \mathbf{E}[h](x))^2 \\ &\leq \sup_{\|h\|_l \leq 1} \mathbb{E}_{X\times Y}(\langle h, l(y,\cdot)\rangle_l - \langle h, \mathbf{E}^* k(x,\cdot)\rangle_l)^2 \\ &\leq \mathbb{E}_{X\times Y} \|l(y,\cdot) - \mathbf{E}^*[k(x,\cdot)]\|_l^2.\end{aligned}$$

This differs from our approach of Section 2.2 in that $\mathbf{X}(x)$ is no longer deterministic, but takes the values $l(y, \cdot)$ according to the product distribution. With the usual (regularised) empirical version we get the estimate

$$\mathbf{E}h = \sum_{i,j=1}^n h(y_j) \mathbf{W}_{ij} k(x_i, \cdot), \tag{5}$$

where $\mathbf{W}$ is defined in eq. 2. The expression is very similar to the solution μ in (Grünewälder et al., 2012a), since $\mu(x) = \mathbf{E}^* k(x, \cdot)$ (see Supp. C.3).

4. Composite Transformations

4.1. Sum Rule – Change of Measure on Y

We next consider a smooth approximation to the sum rule, as introduced by Song et al. (2009)[eq. 6]; see also Fukumizu et al. (2012, Theorem 3.2). We have two measures $\mathbb{P}$ and $\mathbb{Q}$ on the product space $X \times Y$. We assume that for each x we have conditional measures $\mathbb{P}_{Y|x} = \mathbb{Q}_{Y|x}$. The task is to estimate the marginal distribution of $\mathbb{Q}$ on Y, i.e. $\mathbb{Q}_Y$, based on samples $\{(x_i, y_i)\}_{i=1}^n$ from $\mathbb{P}_{X\times Y}$ and $\{z_i\}_{i=1}^m$ from $\mathbb{Q}_X$.

In our setting the task is formulated naturally in a weak sense, i.e. we want to infer an RKHS element m_Y such that $\mathcal{E}_m[m_Y] := \sup_{\|h\|_l \leq 1} (\mathbb{E}_{\mathbb{Q}_Y} h - \langle m_Y, h\rangle_l)^2$ is small. We can reformulate the expectation to reduce it to quantities we observe. Formally, we have

$$\mathbb{E}_{\mathbb{Q}_Y} h = \mathbb{E}_{\mathbb{Q}_{X\times Y}} h = \mathbb{E}_{\mathbb{Q}_X} \mathbb{E}_{\mathbb{Q}}[h|x] = \mathbb{E}_{\mathbb{Q}_X} \mathbb{E}_{\mathbb{P}}[h|x]. \tag{6}$$

The problem of performing these transformations when we have only samples can now be addressed naturally in the operator framework. Using the samples from $\mathbb{P}_{X\times Y}$ we can infer a conditional expectation estimate $\mathbf{E}[h](x) \approx \mathbb{E}[h|x]$ via Sec. 3.2, and using samples $\{z_i\}_{i=1}^m$ from $\mathbb{Q}_X$, we can infer an $m_X = m^{-1} \sum_{i=1}^m k(z_i, \cdot)$ representing $\mathbb{Q}_X$. We can now form compositions of the approximate conditional expectation operation $\mathbf{E}$ and the approximate expectation operation $\langle m_X, \cdot\rangle_k$ as $\mathbf{E}$ maps into $\mathcal{H}_X$: $\langle m_X, \mathbf{E}h\rangle_k = \langle \mathbf{E}^* m_X, h\rangle_l$. A natural estimate m_Y is hence $\mathbf{E}^* m_X$. With the expectation estimate from eq. 5 and $\mathbf{W}$ from eq. 2 we have

$$m_Y = \mathbf{E}^* m_X = \sum_{i,j=1}^n m_X(x_i) \mathbf{W}_{ij} l(y_j, \cdot),$$

which is the estimate of Song et al. (2009).

4.1.1. Estimation Error

Assuming we have control over the approximation error $\mathcal{E}_c[\mathbf{E}]$ of $\mathbf{E}$ and $\mathcal{E}_m[m_X]$ of m_X, and we want to get

error approximations for m_Y, i.e. upper bounds on $\mathcal{E}_m[m_Y]$. The next theorem provides these. The proof uses the transformation in eq. 6 and the link of the involved quantities to the estimates $\mathbf{E}$ and m_X. The kernel function is $\Xi(h, h') := \langle h, \mathbf{A}h'\rangle_l \mathbf{B}$.

Theorem 4.1 (Proof in supp., p. 16)**.** *We assume that the integrability assumptions from Supp. F hold, $\mathbb{Q}_X \ll \mathbb{P}_X$, and the corresponding Radon-Nikodým derivative r is a.e. upper bounded by b. Defining $c = \|\mathbf{A}^{1/2}\|_{op}^2 \|\mathbf{B}\|_{op}$, we have that*

$$\mathcal{E}_m[m_Y] \leq b\mathcal{E}_c[\mathbf{E}] + c\|\mathbf{E}\|_{\Xi}^2 \mathcal{E}_m[m_X].$$

The error is controlled by scaled versions of the errors of $\mathbf{E}$ and m_X, which is as we would hope. The convergence rate of $\mathcal{E}_m[m_Y]$ in terms of sample size is controlled by the slower rate of $\mathcal{E}_c[\mathbf{E}]$ and $\mathcal{E}_m[m_X]$ when $\|\mathbf{E}\|_{\Xi}^2$ stays bounded.

4.2. Bayes' Rule – Ch. of Meas. on $X|y$

Closely related to the approximate sum rule is an approximate Bayesian inference setting, as described by Fukumizu et al. (2011); Song et al. (2013). As in the case of the sum rule, we have two measures $\mathbb{P}$ and $\mathbb{Q}$ on the product space $X \times Y$, samples $\{(x_i, y_i)\}_{i=1}^n$ from $\mathbb{P}_{X\times Y}$, samples $\{z_i\}_{i=1}^m$ from $\mathbb{Q}_X$, and we assume $\mathbb{P}_{Y|x} = \mathbb{Q}_{Y|x}$. The difference compared with the sum rule is that we are not interested in the marginal $\mathbb{Q}_Y$, but in $\mathbb{Q}_{X|y}$.

It is intuitive to consider this problem in a weak sense: that is, instead of estimating the full distribution, we want to learn a version of the conditional expectation acting on functions f, i.e., to minimise

$$\mathcal{E}_c[\mathbf{G}] = \sup_{\|f\|_k \leq 1} \mathbb{E}_{\mathbb{Q}_Y}(\mathbb{E}_{\mathbb{Q}_X}[f|y] - \mathbf{G}[f](y))^2.$$

Unlike the problem of estimating conditional expectations, however, we observe only $\mathbb{P}$ on the product space $X \times Y$, and not the $\mathbb{Q}$ for which we want the conditional expectation. In this setting multiple operations must be combined, and the operator approach shows its strength in terms of keeping the manipulations simple.

We begin by linking the problem of estimating $\mathbb{E}[f|y]$ with $\mathbf{G}$ to the easier problem of estimating $\mathbb{E}[h|x]$ with $\mathbf{E}$. The latter problem is easier since $\mathbb{Q}_{Y|x} = \mathbb{P}_{Y|x}$ and we can use the usual approach to estimate the conditional expectation with samples from $\mathbb{P}$. As with the sum rule, the quality of this estimate as an estimate of $\mathbb{Q}_{Y|x}$ depends on the Radon-Nikodým derivative of the marginal measures, as the estimate is optimised wrt. $\mathbb{E}_{\mathbb{P}_X}$ and not $\mathbb{E}_{\mathbb{Q}_X}$.

We can use integral transformations to link the conditional expectations. One of the challenges is the introduction of an integral over $\mathbb{Q}_Y$ such that we can move from $\mathbb{E}[f|y]$ to a product integral, and from the product integral to the conditional expectation $\mathbb{E}[h|x]$. One way to do this is to approximate a δ-peak at y with a function $\bar{\delta}_y$. This function should be concentrated around y, and should be normalised to 1 wrt. $\mathbb{Q}_Y$ to approximate the point evaluator at y. In this case we can approximate $\mathbb{E}[f|y]$ with

$$\begin{aligned}\mathbb{E}_{Y'} \frac{\bar{\delta}_y(y')}{\mathbb{E}_{Y'}\bar{\delta}_y(y')} \mathbb{E}[f|y'] &= \mathbb{E}_{X\times Y'} f \times \frac{\bar{\delta}_y(y')}{\mathbb{E}_{Y'}\bar{\delta}_y(y')} \\ &= \frac{1}{\mathbb{E}_{Y'}\bar{\delta}_y(y')} \mathbb{E}_X f \, \mathbb{E}_{Y'}[\bar{\delta}_y(y')|x].\end{aligned}$$

An RKHS kernel function $l(y, \cdot)$ can serve as a smoothed approximation to a point-evaluator. For example, a Gaussian kernel with a bandwidth parameter σ becomes concentrated around y for small σ. We thus choose $\bar{\delta}_y = l(y, \cdot)$, bearing in mind that this will introduce a non-vanishing bias. With this choice, and by approximating the last term with the estimate $\mathbf{E}$, we get

$$\begin{aligned}&\mathbb{E}_X f(x) \mathbb{E}[l(y,\cdot)|x] \approx \mathbb{E}_X f(x) \mathbf{E}[l(y,\cdot)](x) \\ &= \mathbb{E}_X \langle f, k(x,\cdot)\rangle_k \langle \mathbf{E} l(y,\cdot), k(x,\cdot)\rangle_k \\ &= \mathbb{E}_X \langle f, \langle \mathbf{E} l(y,\cdot), k(x,\cdot)\rangle_k k(x,\cdot)\rangle_k,\end{aligned}$$

The term $\mathbb{E}_Y l(y, \cdot)$ is approximated by the mean estimate $\langle m_Y, l(y,\cdot)\rangle_l$, computed via change of measure.

We next approximate the above with $\mathbf{G}[f](y)$ to estimate $\mathbb{E}[f|y]$. By defining a suitable distribution $\mathbb{R}_Y$ over Y to approximate $\mathbb{E}[f|y]$, and following the usual approach, we get

$$\begin{aligned}&\sup_{\|f\|_k \leq 1} \mathbb{E}_Y \Big(\langle \mathbf{G} f, l(y,\cdot)\rangle_l - \\ &\quad (\langle m_Y, l(y,\cdot)\rangle_l)^{-1} \mathbb{E}_X \langle f, \langle \mathbf{E} l(y,\cdot), k(x,\cdot)\rangle_k k(x,\cdot)\rangle_k \Big)^2 \\ &\leq \mathbb{E}_{X\times Y} \|\mathbf{G}^* l(y,\cdot) - u(x,y) k(x,\cdot)\|_k^2,\end{aligned} \tag{7}$$

where the product measure is over the independent probability measures $\mathbb{Q}_X$ and $\mathbb{R}_Y$ which we choose, and we are approximating the function

$$\begin{aligned}u(x,y) &= \langle \mathbf{E} l(y,\cdot), k(x,\cdot)\rangle_k \langle m_Y, l(y,\cdot)\rangle_l^{-1} \\ &\approx \left(\mathbb{E}_{Y|x} l(y,\cdot)\right) \left(\mathbb{E}_{Y|x}\mathbb{E}_{\mathbb{Q}_X} l(y,\cdot)\right)^{-1}.\end{aligned}$$

The above is an estimate (via $\mathbf{E}$) of a ratio of smoothed densities, the numerator being a smoothed conditional density. If the bandwidth parameter of the kernel on $\mathcal{H}_Y$ is fixed, then this smoothing remains a source of bias, and shows up as an approximation error in Th. 4.2 below. If we now use the empirical and λ-

regularised version of the upper bound, we get an estimate for $\mathbb{E}[f|y]$,

$$\mathbf{G}f = \sum_{i,j=1}^{n} f(x_j)\mathbf{E}\left[\frac{l(y_j,\cdot)}{\langle m_Y, l(y_j,\cdot)\rangle_l}\right](x_j)\mathbf{W}_{ij}l(y_i,\cdot),$$

with $\mathbf{W} = (\mathbf{L} + \lambda\mathbf{I})^{-1}$, $\mathbf{L}$ being the kernel matrix, $\{x_i\}_{i=1}^n$ being samples from $\mathbb{Q}_X$ and $\{y_i\}_{i=1}^n$ from $\mathbb{R}_Y$. Note that this expression is not the same as the kernel Bayes' rule of Fukumizu et al. (2012, Figure 1); an empirical comparison of the two approaches remains a topic for future work.

4.2.1. Estimation Error

The error of the estimator $\mathbf{G}$ can be bounded by the errors of the mean estimate m_X, the error of $\mathbf{E}$, an approximation error

$$\mathcal{E}_a[l] := \sup_{\|h\|_1\leq 1} \mathbb{E}_Y\left(h(y) - \mathbb{E}_{Y'}\frac{l(y,y')}{\mathbb{E}_{Y'}l(y,y')}h(y')\right)^2$$

where $y, y' \sim \mathbb{Q}_Y$, and the risk of $\mathbf{G}$ in the top line of eq. 7. We denote this risk with $\mathcal{E}_K[\mathbf{G}]$. The following theorem states the bound. The risks in the theorem are measured wrt. $\mathbb{Q}$ for all but the estimate $\mathbf{E}$ and the constant C, and can be found in the supplement.

Theorem 4.2 (Proof in supp., p. 17)**.** *We assume that the integrability assumptions from Supp. F hold, that $\mathbb{Q}_X \ll \mathbb{P}_X$, and that the corresponding Radon-Nikodým derivative is a.e. upper bounded by b. Furthermore, we assume that there exists a constant $q > 0$ such that $\mathbb{E}_{y'\sim\mathbb{P}_Y} l(y,y') \geq q$ for all $y \in Y$ and that the approximation error of m_Y is such that $|\mathbb{E}_{y'\sim\mathbb{P}_Y} l(y,y') - \langle m_Y, l(y,\cdot)\rangle_l| \leq |\mathbb{E}_{y'\sim\mathbb{P}_Y} l(y,y')|/2$. There exists a positive constant C such that*

$$\mathcal{E}_c[\mathbf{G}] \leq \mathcal{E}_K[\mathbf{G}] + C\left(\mathcal{E}_a[l] + \|\mathbf{E}\|^2_{\mathbf{\Xi}}\mathcal{E}_m[m_X] + \mathcal{E}_{c,\mathbb{P}}[\mathbf{E}]\right).$$

The assumption on m_Y guarantees that we are reasonably close to the true expectation. This is fulfilled with high probability after finitely many steps for the standard estimate. The assumption $\mathbb{E}_{y'\sim\mathbb{P}_Y} l(y,y') \geq q$ guarantees that we have a good approximate point evaluator at y'.

4.3. A Short Note on Convergence Rates

Convergence rates are obviously a big topic and we do not want to go into too much depth here. We therefore keep the necessary assumptions simple, and we derive rates only for the approximate sum rule, which we compare with the rates of (Fukumizu et al., 2012). We make a number of assumptions, which can be found in Sec. E.1. The main assumption is that $\mathcal{H}_X$ and $\mathcal{H}_Y$ are finite dimensional. The $\mathcal{H}_Y$ assumption is crucial, however the $\mathcal{H}_X$ assumption can be avoided with some extra effort. Another assumption concerns the probability measures over which the convergence occurs. We refer the reader here to Caponnetto & De Vito (2007) for details, and we take $\mathfrak{P}$ to be the class of priors from Def. 1 with $b = \infty$. There is an approximation error in the theorem which measures how well we can approximate the true conditional expectation (see Supp. E for the definition). Finally, we assume that we have a rate of $\alpha \in]0, 1]$ to estimate the mean of $\mathbb{Q}_X$.

Theorem 4.3 (Proof in Supp. E)**.** *Let $\mathbf{E}_*$ be a minimiser of the approximation error $\mathcal{E}_A$, and let the schedule for the regulariser for $\mathbf{E}_n$ be chosen according to Caponnetto & De Vito (2007)[Thm 1]. Under assumptions E.1 and if $\mathbb{Q}_X \ll \mathbb{P}_X$ with a bounded Radon-Nikodým derivative, we have that for every $\epsilon > 0$ there exist constants a, b, c, d such that*

$$\limsup_{n\to\infty}\sup_{\mathbb{P}\in\mathfrak{P}}(\mathbb{P}\otimes\mathbb{Q})^n\left[\mathcal{E}_m[m_Y^n] > \left(a\|\mathbf{E}_n\|^2_{\mathbf{\Xi}} n^{-\alpha} + \mathcal{E}_A[\mathbf{E}_*]\left(1 + \sqrt{b + c\|\mathbf{E}_n\|_{\Xi}}\right) + dn^{-\frac{1}{2}}\right)^2\right] < \epsilon.$$

The value $\|\mathbf{E}_n\|_{\Xi}$ is of obvious importance. $\mathbf{E}_n$ is the minimiser of the empirical regularised risk, and if this minimiser converges with high probability to the minimiser of the regularised risk, then one can infer from Caponnetto & De Vito (2007)[Prop. 3] that $\mathbf{E}_n$ will be bounded with high probability. This then guarantees a rate of convergence of $n^{-\alpha}$, which matches the state of art rates of Fukumizu et al. (2012)[Th. 6.1] which are between $n^{-2/3\alpha}$ and $n^{-\alpha}$, depending on the smoothness assumptions made.

5. Conclusion

We have presented an approach for estimating linear operators acting on an RKHS. Derivations of estimates are often generic, and operations can naturally be combined to form complex estimates. Risk bounds for these complex rules can be expressed straightforwardly in terms of risk bounds of the basic estimates used in building them. There are obviously many routes to explore from here. Most immediately, improved estimation techniques would be helpful, incorporating sparsity and other constraints. It would also be interesting to consider additional machine learning settings in this framework.

Acknowledgements The authors want to thank for the support of the EPSRC #EP/H017402/1 (CARDyAL) and the European Union #FP7-ICT-270327 (Complacs), as well as the reviewers for helpful suggestions.

References

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68 (3):337–404, 1950.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in Probability and Statistics*. Kluwer, 2004.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Carmeli, C., De Vito, E., and Toigo, A. Vector valued reproducing kernel Hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(4):377–408, 2006.

Fremlin, D.H. *Measure Theory - Volume 1: The Irreducible Minimum*. Torres Fremlin, 2000.

Fremlin, D.H. *Measure Theory - Volume 2: Broad Foundations*. Torres Fremlin, 2001.

Fremlin, D.H. *Measure Theory - Volume 4: Topological Measure Spaces*. Torres Fremlin, 2003.

Fukumizu, K., Song, L., and Gretton, A. Kernel bayes' rule. In *NIPS*, 2011.

Fukumizu, K., Song, L., and Gretton, A. Kernel bayes' rule: Bayesian inference with positive definite kernels. *ArXiv*, 1009.5736v4, 2012.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate shift and local learning by distribution matching. In *Dataset Shift in Machine Learning*. 2009.

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. Conditional mean embeddings as regressors. In *ICML*, 2012a.

Grünewälder, S., Lever, G., Baldassarre, L., Pontil, M., and Gretton, A. Modelling transition dynamics in MDPs with RKHS embeddings. In *ICML*, 2012b.

Huang, J., Smola, A. J., Gretton, A., Borgwardt, K., and Schölkopf, B. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.

Micchelli, C.A. and Pontil, M.A. On learning vector-valued functions. *Neural Computation*, 17(1), 2005.

Nishiyama, Y., Boularias, A., Gretton, A., and Fukumizu, K. Hilbert space embeddings of POMDPs. In *UAI*, 2012.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *ALT*, 2007.

Song, L., Huang, J., Smola, A.J., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *ICML*, 2009.

Song, L., Boots, B., Siddiqi, S. M., Gordon, G. J., and Smola, A. J. Hilbert space embeddings of hidden Markov models. In *ICML*, 2010.

Song, L., Fukumizu, K., and Gretton, A. Kernel embeddings of conditional distributions. *IEEE Signal Processing Magazine*, To Appear, 2013.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer, 2008.

Werner, D. *Funktionalanalysis*. Springer, 4th edition, 2002.

Yu, Y. and Szepesvari, C. Analysis of kernel mean matching under covariate shift. In *ICML*, 2012.