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.

SUPPLEMENTARY

A. Section 2: Smooth Operators – Supplementary Results

A.1. On the Relation between the Natural and Surrogate Risk

We follow up on the discussion in Section 2.2 about the surrogate and natural risk. The surrogate risk itself is not a quantity we really care about, it is only an upper bound that makes optimisation feasible. In general, we have that upper bounds similar to the one we derived in Section 2.2 are loose. For the conditional expectation estimates in Section 3.2, for example, the upper bound corresponds to something like the variance of the underlying distribution at points x and might be arbitrarily high for all estimates, while the natural risk can be decreased to zero with a reasonable estimator. Yet, the situation is not as grim as it seems. The reason for this is that the positions of the minimisers are often closely related, i.e. a minimum of the surrogate risk is in certain cases also a minimum of the natural risk. More generally, the minima often do not overlap exactly, but due to some continuity properties they are not located too far apart and we suffer only a minor penalty compared to the true minimiser by using the surrogate minimiser.

Why this is the case is easy to see for the setting in Section 3.2. If we are in the lucky situation that $\mathbf{X}(x)$ can be represented by a $\mathbf{G}^* k(x, \cdot)$ then this $\mathbf{G}^* k(x, \cdot)$ is the minimiser for both the upper bound and the natural risk function. Furthermore, the bound becomes tight as the surrogate risk can be minimised to zero. If we can not represent $\mathbf{X}(x)$ exactly then the surrogate risk minimises the difference to $\mathbf{X}(x)$ and the natural risk is bounded by this approximation error.

Usually, we have a variation of the risk functions of Section 2.2 and relating the minimisers becomes more complicated. The problem of relating the risk functions is an important one and it is useful to have a rather general way to link these risk functions. One such approach is to use conditional expectations where we condition wrt. a σ-algebra Σ (Fremlin, 2001)[Chp. 233]. It is well known that such conditional expectations are in a suitable sense L^2 minimisers over all Σ-measurable functions (Fremlin, 2001)[244N]. Our setting is a bit more complicated then the standard L^2 setting, but, intuitively, if we can find a suitable Σ such that the conditional expectation wrt. Σ is a solution for both the natural and the surrogate risk and if the class of Σ-measurable functions overlaps with the functions we can represent with $\mathbf{G}^* k(x, \cdot)$ then we know that the minimisers are co-located. We use this argument in a form adapted to our setting for kernelized approximate Bayesian inference and the simple conditional expectation $\mathbb{E}[\,\cdot\,|x]$ to relate the risk functions.

A.2. Reproducing Kernel

We verify here that $\boldsymbol{\Xi}$ is a valid reproducing kernel. We use the criterion from Carmeli et al. (2006)[Prop.1] to verify this. The criterion resembles the positive-definiteness of a scalar valued kernel. The criterion is fulfilled, if $\boldsymbol{\Xi}(f, g) \in L(\mathcal{H}_Y)$ (which is fulfilled as $\boldsymbol{\Xi}(f, g) = c\mathbf{B}$, for a $c \in \mathbb{R}$) and for all $n \in \mathbb{N}$, $\{c_i\}_{i=1}^n$, $c_i \in \mathbb{R}$, $\{f_i\}_{i=1}^n$, $f_i \in \mathcal{H}_X$, and all $h \in \mathcal{H}_Y$ it holds that

$$\sum_{i=1}^{n}\sum_{i=j}^{n} c_i c_j \langle \boldsymbol{\Xi}(f_i, f_j)h, h\rangle_l = \langle \sum_{i=1}^{n} c_i f_i, \mathbf{A}\sum_{i=1}^{n} c_i f_i\rangle_k \langle \mathbf{B}h, h\rangle_l = \| \sum_{i=1}^{n} \mathbf{A}^{1/2} c_i f_i\|_k^2 \|\mathbf{B}^{1/2} h\|_l^2$$

is greater than zero. This is obviously fulfilled and $\boldsymbol{\Xi}$ has an associated RKHS $\mathcal{H}_{\boldsymbol{\Xi}}$.

A.3. Case Study III: Smooth Quotient Operators

Analogously to multiplication one can derive an operator for forming quotients, $f/g \approx \mathbf{Q}f$, where $f \in \mathcal{H}_X$ and $g(x) \neq 0$ for all x. In the unconstrained case we can find a suitable operator $\mathbf{X}$ by using eq. 1 with $\mathbf{X}(x) := \frac{k(x,\cdot)}{g(x)}$, which is in $\mathcal{H}_X$ for a given x, is a valid choice. The approximation is hence

$$\mathbf{Q}f = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{f(x_j)}{g(x_j)} \mathbf{W}_{ij} k(x_i, \cdot), \text{with } \mathbf{W} = (\mathbf{K} + \lambda \mathbf{I})^{-1}.$$

B. Section 2: Smooth Operators – Proofs

Theorem B.1. *Each $\mathbf{F} \in \mathcal{H}_{\boldsymbol{\Xi}}$ is a bounded linear operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$.*

Proof. **(a)** Each operator in $\mathcal{L} = \{\sum_{i=1}^n \boldsymbol{\Xi}(f_i, \cdot)h_i : n \in \mathbb{N}, f_i \in \mathcal{H}_X, h_i \in \mathcal{H}_Y\}$ linear as

$$\mathbf{F}[af + bg] = \sum_{i=1}^n \langle af + bg, \mathbf{A}f_i\rangle_k \mathbf{B}h_i = a\sum_{i=1}^n \langle f, \mathbf{A}f_i\rangle_k \mathbf{B}h_i + b\sum_{i=1}^n \langle g, \mathbf{A}f_i\rangle_k \mathbf{B}h_i = a\mathbf{F}f + b\mathbf{F}g.$$

(b) Also each operator in $\mathcal{H}_{\boldsymbol{\Xi}} = \text{clos}\,\mathcal{L}$ is linear – see, for example, the proof of Prop. 1 in Carmeli et al. (2006) for the equivalence of the closure of $\mathcal{L}$ and $\mathcal{H}_{\boldsymbol{\Xi}}$.

P Since $\mathcal{L}$ is dense we can find for each $\epsilon > 0$ and $\mathbf{F} \in \mathcal{H}_{\boldsymbol{\Xi}}$ an operator $\mathbf{F}_\delta \in \mathcal{L}$ such that $\|\mathbf{F} - \mathbf{F}_\delta\|_{\boldsymbol{\Xi}} < \delta$. We have for an arbitrary $g \in \mathcal{H}_X$ that

$$\begin{aligned}\|\mathbf{F}g - \mathbf{F}_\delta g\|_l &= \|(\mathbf{F} - \mathbf{F}_\delta)g\|_l \leq \|\mathbf{F} - \mathbf{F}_\delta\|_{\boldsymbol{\Xi}} \|\boldsymbol{\Xi}(g, g)\|_{\text{op}}^{1/2} = \|\mathbf{F} - \mathbf{F}_\delta\|_{\boldsymbol{\Xi}} \|\mathbf{A}^{1/2} g\|_k \|\mathbf{B}\|_{\text{op}}^{1/2} \\ &\leq \|\mathbf{F} - \mathbf{F}_\delta\|_{\boldsymbol{\Xi}} \|\mathbf{A}^{1/2}\|_{\text{op}} \|\mathbf{B}\|_{\text{op}}^{1/2} \|g\|_k,\end{aligned}$$

where we used Prop 2.1 (f) from Micchelli & Pontil (2005) for the first inequality and the positivity and self-adjointness of $\mathbf{A}$ and $\mathbf{B}$ to guarantee the existence of square-roots. As $\mathbf{A}, \mathbf{B}$ are bounded we can pick for a given g a δ such that $\|\mathbf{F}g - \mathbf{F}_\delta g\|_l < \epsilon$.

Now, we can also pick a δ such that $\|\mathbf{F}[af + bg] - \mathbf{F}_\delta[af + bg]\|_k$, $\|a\mathbf{F}f - a\mathbf{F}_\delta f\|_k$ and $\|b\mathbf{F}g - b\mathbf{F}_\delta g\|_k$ are simultaneously smaller than $\epsilon/3$.

Hence, for a given ϵ we have a $\mathbf{F}_\epsilon$ such that

$$\|\mathbf{F}[af + bg] - a\mathbf{F}f - b\mathbf{F}g\|_k \leq \|\mathbf{F}[af + bg] - \mathbf{F}_\delta[af + bg]\|_k + \|a\mathbf{F}_\delta f + b\mathbf{F}_\delta g - a\mathbf{F}f - b\mathbf{F}g\|_k \leq \epsilon.$$

Since this holds for every $\epsilon > 0$ we have that $\|\mathbf{F}[af + bg] - a\mathbf{F}f - b\mathbf{F}g\|_k = 0$ and $\mathbf{F}[af + bg] = a\mathbf{F}f + b\mathbf{F}g$, i.e. $\mathbf{F}$ is linear. **Q**

(c) Each $\mathbf{F}$ maps into $\mathcal{H}_Y$. This is implicitly in Th. 2.1 from (Micchelli & Pontil, 2005), but is also easy to derive: we want to show that $\mathbf{F}f \in \mathcal{H}_Y$. We know this holds for any $\mathbf{F}' \in \mathcal{L}$ and we can for any $\mathbf{F} \in \mathcal{H}_{\boldsymbol{\Xi}}$ find a sequence $\{\mathbf{F}_n\}_{n=1}^\infty$ in $\mathcal{L}$ that converges to $\mathbf{F}$, and is hence a Cauchy sequence. Now, as $\mathcal{H}_Y$ is complete it is sufficient for convergence to show that for a given $f \in \mathcal{H}_X$, $\mathbf{F}_n f$ is a Cauchy sequence. Similarly, like in (b), we have

$$\|\mathbf{F}_n f - \mathbf{F}_m f\|_l \leq \|\mathbf{F}_n - \mathbf{F}_m\|_{\boldsymbol{\Xi}} \|\boldsymbol{\Xi}(f, f)\|_{\text{op}}^{1/2}.$$

Since $\boldsymbol{\Xi}(f, f)$ is a bounded operator we have shown that $\{\mathbf{F}_n f\}_{n=1}^\infty$ is a Cauchy sequence in $\mathcal{H}_Y$ and has hence a limit $\tilde{\mathbf{F}}f$ in $\mathcal{H}_Y$. We have

$$\|\tilde{\mathbf{F}}f - \mathbf{F}f\|_l \leq \|\tilde{\mathbf{F}}f - \mathbf{F}_n f\|_l + \|\mathbf{F}_n - \mathbf{F}\|_{\boldsymbol{\Xi}} \|\boldsymbol{\Xi}(f, f)\|_{\text{op}}^{1/2}.$$

Since $\mathbf{F}_n f$ converges to $\tilde{\mathbf{F}}f$ in $\mathcal{H}_Y$ and $\mathbf{F}_n$ converges to $\mathbf{F}$ in $\mathcal{H}_{\boldsymbol{\Xi}}$ we have that $\tilde{\mathbf{F}}f = \mathbf{F}f \in \mathcal{H}_Y$.

(d) Finally, each $\mathbf{F}$ is bounded as an operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$ as

$$\|\mathbf{F}f\|_l \leq \|\mathbf{F}\|_{\boldsymbol{\Xi}} \|\boldsymbol{\Xi}(f, f)\|_{\text{op}}^{1/2} = \|\mathbf{F}\|_{\boldsymbol{\Xi}} \|\langle f, \mathbf{A}f\rangle_k \mathbf{B}\|_{\text{op}}^{1/2} \leq \|\mathbf{F}\|_{\boldsymbol{\Xi}} \|\mathbf{A}^{1/2}\|_{\text{op}} \|f\|_k \|\mathbf{B}\|_{\text{op}}^{1/2} \leq C\|f\|_k.$$

□

Theorem B.2. *If for $\mathbf{F}, \mathbf{G} \in \mathcal{H}_{\boldsymbol{\Xi}}$ 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.*

Proof. As $\mathbf{F}$ and $\mathbf{G}$ are continuous it follows that they are uniquely defined by their values on the dense subset $\mathcal{L}_X$ of $\mathcal{H}_X$. Now, let $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ be an arbitrary element in $\mathcal{L}_X$ then $\mathbf{F}f = \sum_{i=1}^n \alpha_i \mathbf{F}k(x_i, \cdot)$ and $\mathbf{F}f = \mathbf{G}f$ if $\mathbf{F}k(x, \cdot) = \mathbf{G}k(x, \cdot)$ for all $x \in X$. This proves the first statement.

Now, assume that we only know that both operators are equal on a dense set $\mathcal{D}$ of X. Take an arbitrary $x \in X$. There exists a sequence $\{x_j\}_{j=1}^\infty$ in $\mathcal{D}$ converging to x. We have that $\mathbf{F}k(x, \cdot) = \mathbf{F}\lim_{n\to\infty} k(x_n, \cdot) = \lim_{n\to\infty} \mathbf{F}k(x_n, \cdot) = \lim_{n\to\infty} \mathbf{G}k(x_n, \cdot) = \mathbf{G}k(x, \cdot)$ and both operators are equal on all $k(x, \cdot)$. □

Theorem B.3. *For every $\mathbf{F} \in \mathcal{H}_{\boldsymbol{\Xi}}$ there exists an adjoint $\mathbf{F}^*$ in $\mathcal{H}_{\boldsymbol{\Xi}^*}$ 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 $\mathbf{F}f = \sum_{i=1}^n \boldsymbol{\Xi}_{f_i}[h_i](f) = \sum_{i=1}^n \langle f, \mathbf{A}f_i\rangle_k \mathbf{B}h_i$ that the adjoint is

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

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

Proof. **(a)** We first derive the explicit expression of $\mathbf{F}^*$ for $\mathbf{F} \in \mathcal{L}$. This is nearly trivial, we have

$$\langle \mathbf{F}f, h\rangle_l = \sum_{i=1}^n \langle f, \mathbf{A}f_i\rangle_k \langle h_i, \mathbf{B}h\rangle_l = \sum_{i=1}^n \langle h_i, \mathbf{B}h\rangle_l \langle f, \mathbf{A}f_i\rangle_k = \langle f, \sum_{i=1}^n \langle h, \mathbf{B}h_i\rangle_l \mathbf{A}f_i\rangle_k = \langle f, \mathbf{F}^* h\rangle_k.$$

(b) Next, we verify some properties of $(\mathrm{T}\restriction \mathcal{L})$, where $(\mathrm{T}\restriction \mathcal{L})[\sum_{i=1}^n \boldsymbol{\Xi}_{f_i} h_i] = \sum_{i=1}^n \boldsymbol{\Xi}^*_{h_i} f_i$. **(i)** $(\mathrm{T}\restriction \mathcal{L})$ is linear as

$$(\mathrm{T}\restriction \mathcal{L})[a\mathbf{F} + b\mathbf{G}] = (\mathrm{T}\restriction \mathcal{L})[\sum_{i=1}^n \boldsymbol{\Xi}_{f_i} a h_i + \sum_{j=1}^m \boldsymbol{\Xi}_{g_j} b u_j] = \sum_{i=1}^n \boldsymbol{\Xi}^*_{ah_i} f_i + \sum_{j=1}^m \boldsymbol{\Xi}^*_{bu_j} g_j$$
$$= a\sum_{i=1}^n \langle \cdot, \mathbf{B}h_i\rangle_l \mathbf{A}f_i + b\sum_{j=1}^m \langle \cdot, \mathbf{B}u_j\rangle_l \mathbf{A}g_j = a(\mathrm{T}\restriction \mathcal{L})\mathbf{F} + b(\mathrm{T}\restriction \mathcal{L})\mathbf{G}.$$

where we used that $\boldsymbol{\Xi}_u$ is a linear operator. **(ii)** $(\mathrm{T}\restriction \mathcal{L})$ is norm preserving, as

$$\|(\mathrm{T}\restriction \mathcal{L})\mathbf{F}\|^2_{\boldsymbol{\Xi}^*} = \sum_{i=1}^n \sum_{j=1}^n \langle \boldsymbol{\Xi}^*_{h_i}[f_i], \boldsymbol{\Xi}^*_{h_j}[f_j]\rangle_{\boldsymbol{\Xi}^*} = \sum_{i=1}^n \sum_{j=1}^n \langle f_i, \boldsymbol{\Xi}^*(h_i, h_j) f_j\rangle_k$$
$$= \sum_{i=1}^n \sum_{j=1}^n \langle h_i, \mathbf{B}h_j\rangle_l \langle f_i, \mathbf{A}f_j\rangle_k = \sum_{i=1}^n \sum_{j=1}^n \langle h_i, \boldsymbol{\Xi}(f_i, f_j) h_j\rangle_l = \|\mathbf{F}\|^2_{\boldsymbol{\Xi}}.$$

Furthermore, $(\mathrm{T}\restriction \mathcal{L})$ is continuous as $\|(\mathrm{T}\restriction \mathcal{L})\|_{\mathrm{op}} = \sup_{\mathbf{F}\in\mathcal{L}, \|\mathbf{F}\|_{\boldsymbol{\Xi}}=1} \|(\mathrm{T}\restriction \mathcal{L})\mathbf{F}\|_{\boldsymbol{\Xi}^*} = \sup_{\mathbf{F}\in\mathcal{L}, \|\mathbf{F}\|_{\boldsymbol{\Xi}}=1} \|\mathbf{F}\|_{\boldsymbol{\Xi}} = 1$. **(iii)** $(\mathrm{T}\restriction \mathcal{L})$ is bijective. Take an arbitrary $\mathbf{G} \in \mathcal{L}^*$ then $\mathbf{G} = \sum_{i=1}^n \boldsymbol{\Xi}^*_{h_i}[f_i]$ for suitable choices of n, f_i, h_i. We have that $(\mathrm{T}\restriction \mathcal{L})[\sum_{i=1}^n \boldsymbol{\Xi}_{f_i}[h_i]] = \mathbf{G}$ and, hence, $(\mathrm{T}\restriction \mathcal{L})$ is surjective. $(\mathrm{T}\restriction \mathcal{L})$ is also injective, take $\mathbf{F}, \mathbf{F}'$ such that $(\mathrm{T}\restriction \mathcal{L})\mathbf{F} = (\mathrm{T}\restriction \mathcal{L})\mathbf{F}'$ then, we have

$$\|\mathbf{F} - \mathbf{F}'\|_{\boldsymbol{\Xi}} = \|(\mathrm{T}\restriction \mathcal{L})[\mathbf{F} - \mathbf{F}']\|_{\boldsymbol{\Xi}^*} = \|(\mathrm{T}\restriction \mathcal{L})\mathbf{F} - (\mathrm{T}\restriction \mathcal{L})\mathbf{F}'\|_{\boldsymbol{\Xi}^*} = 0,$$

as $(\mathrm{T}\restriction \mathcal{L})$ is norm preserving and we conclude $\mathbf{F} = \mathbf{F}'$.

(c) As $\mathcal{L}$ is dense and $(\mathrm{T}\restriction \mathcal{L})$ a bounded linear operator from $\mathcal{L}$ to $\mathcal{H}_{\boldsymbol{\Xi}^*}$ there exists a unique continuous extension $\mathrm{T} : \mathcal{H}_{\boldsymbol{\Xi}} \mapsto \mathcal{H}_{\boldsymbol{\Xi}^*}$ of $(\mathrm{T}\restriction \mathcal{L})$ (Werner, 2002)[Satz II.1.5]. Furthermore, $\|\mathrm{T}\|_{op} = \|(\mathrm{T}\restriction \mathcal{L})\|_{op} = 1$.

We verify again a couple of properties. **(i)** T is injective. **P** Assume that for $\mathbf{F}, \mathbf{G} \in \mathcal{H}_{\boldsymbol{\Xi}}$ it holds that $\mathbf{F} \neq \mathbf{G}$ and $\mathrm{T}\mathbf{F} = \mathrm{T}\mathbf{G}$. As $\mathbf{F} \neq \mathbf{G}$ we have that $\|\mathbf{F} - \mathbf{G}\|_{\boldsymbol{\Xi}^*} > \epsilon$ for an $\epsilon > 0$. Now, as T is continuous it is also uniformly continuous and there exists for arbitrary $\eta > 0$ a $\delta > 0$ such that for all $\mathbf{H}$ it holds that $\|\mathrm{T}\mathbf{H} - \mathrm{T}\mathbf{H}_\eta\|_{\boldsymbol{\Xi}^*} < \eta$, whenever $\|\mathbf{H} - \mathbf{H}_\eta\|_{\boldsymbol{\Xi}} < \delta$. In the following we use $\eta = \epsilon/6$ and we denote the associated δ with $\delta_{\epsilon/6}$.

For $\mathbf{F}', \mathbf{G}' \in \mathcal{L}$ we have that

$$\|\mathrm{T}\mathbf{F}' - \mathrm{T}\mathbf{G}'\|_{\boldsymbol{\Xi}^*} = \|\mathbf{F}' - \mathbf{G}'\|_{\boldsymbol{\Xi}} = \|\mathbf{F}' - \mathbf{F} + \mathbf{F} - \mathbf{G} + \mathbf{G} - \mathbf{G}'\|_{\boldsymbol{\Xi}} \geq |\|\mathbf{F} - \mathbf{G}\|_{\boldsymbol{\Xi}} - \|\mathbf{F}' - \mathbf{F} + \mathbf{G} - \mathbf{G}'\|_{\boldsymbol{\Xi}}|.$$

As $\mathcal{L}$ is dense, we can pick $\mathbf{F}'$ and $\mathbf{G}'$ such that $\|\mathbf{F}' - \mathbf{F}\|_{\boldsymbol{\Xi}}, \|\mathbf{G} - \mathbf{G}'\|_{\boldsymbol{\Xi}} < \min\{\epsilon/6, \delta_{\epsilon/6}\}$. Hence, $\|\mathbf{F}' - \mathbf{F} + \mathbf{G} - \mathbf{G}'\|_{\boldsymbol{\Xi}} \leq \|\mathbf{F}' - \mathbf{F}\|_{\boldsymbol{\Xi}} + \|\mathbf{G} - \mathbf{G}'\|_{\boldsymbol{\Xi}} < \epsilon/3$ and, consequently, that $\|\mathrm{T}\mathbf{F}' - \mathrm{T}\mathbf{G}'\|_{\boldsymbol{\Xi}^*} > 2/3\epsilon$.

Furthermore, $\|\mathrm{T}\mathbf{F} - \mathrm{T}\mathbf{F}'\|_{\mathbf{\Xi}^*}, \|\mathrm{T}\mathbf{G}' - \mathrm{T}\mathbf{G}\|_{\mathbf{\Xi}^*} < \epsilon/6$ and we have

$$\|\mathrm{T}\mathbf{F} - \mathrm{T}\mathbf{G}\|_{\mathbf{\Xi}^*} \geq |\|\mathrm{T}\mathbf{F} - \mathrm{T}\mathbf{F}'\|_{\mathbf{\Xi}^*} - \|\mathrm{T}\mathbf{F}' - \mathrm{T}\mathbf{G}'\|_{\mathbf{\Xi}^*} + \|\mathrm{T}\mathbf{G}' - \mathrm{T}\mathbf{G}\|_{\mathbf{\Xi}^*}| > 1/3 > 0$$

and $\mathbf{F} \neq \mathbf{G}$. **Q**

(ii) T is surjective, and hence bijective. **P** Consider an arbitrary $\mathbf{G} \in \mathcal{H}_{\mathbf{\Xi}^*}$ and chose a sequence $\{\mathbf{G}_n\}_{n=1}^{\infty}$ in $\mathcal{L}^*$ converging to $\mathbf{G}$. Now, we have exactly one $\mathbf{F}_n \in \mathcal{L}$ such that $\mathrm{T}\mathbf{F}_n = \mathbf{G}_n$. As $\{\mathbf{G}_n\}_{n=1}^{\infty}$ is a Cauchy-sequence, it follows that $\{\mathbf{F}_n\}_{n=1}^{\infty}$ is also a Cauchy-sequence:

$$\|\mathbf{F}_n - \mathbf{F}_m\|_{\mathbf{\Xi}} = \|\mathrm{T}\mathbf{F}_n - \mathrm{T}\mathbf{F}_m\|_{\mathbf{\Xi}^*} = \|\mathbf{G}_n - \mathbf{G}_m\|_{\mathbf{\Xi}^*}$$

and because of the completeness of $\mathcal{H}_{\mathbf{\Xi}}$ the sequence $\{\mathbf{F}_n\}_{n=1}^{\infty}$ has a limit $\mathbf{F}$.

Because of the continuity of T it follows that

$$\mathbf{G} = \lim_{n\to\infty} \mathrm{T}\mathbf{F}_n = \mathrm{T}\mathbf{F}$$

and T is surjective. **Q**

(iii) T has a continuous inverse T^{-1}. That follows from an application of the open mapping theorem, e.g. Kor.IV.3.4 in Werner (2002).

(iv) T is norm preserving. For an arbitrary $\mathbf{F} \in \mathcal{H}_{\mathbf{\Xi}}$ pick a sequence $\{\mathbf{F}_n\}_{n=1}^{\infty}$ in $\mathcal{L}$ that converges to it. Then

$$\|\mathrm{T}\mathbf{F}\|_{\mathbf{\Xi}^*} = \|\lim_{n\to\infty} \mathrm{T}\mathbf{F}_n\|_{\mathbf{\Xi}^*} = \lim_{n\to\infty} \|\mathrm{T}\mathbf{F}_n\|_{\mathbf{\Xi}^*} = \lim_{n\to\infty} \|\mathbf{F}_n\|_{\mathbf{\Xi}} = \|\mathbf{F}\|_{\mathbf{\Xi}}$$

as T is continuous and preserves the norm for elements in $\mathcal{L}$.

(v) That T maps to the adjoint can be seen in a similar way. For an arbitrary $\mathbf{F} \in \mathcal{H}_{\mathbf{\Xi}}$ pick a sequence $\{\mathbf{F}_n\}_{n=1}^{\infty}$ in $\mathcal{L}$ that converges to it. Then

$$\langle \mathbf{F}f, h\rangle_k = \lim_{n\to\infty} \langle \mathbf{F}_n f, h\rangle_k = \lim_{n\to\infty} \langle f, \mathrm{T}\mathbf{F}_n h\rangle_l = \langle f, \mathrm{T}\mathbf{F}h\rangle_l$$

as T is continuous and maps to the adjoint for elements in $\mathcal{L}$. □

Theorem B.4. *The set of self-adjoint operators in $\mathcal{H}_{\mathbf{\Xi}}$ is a closed linear subspace.*

Proof. The set is a linear subspace as for two self-adjoint operators $\mathbf{F}, \mathbf{G}$, scalar a, b and arbitrary $f, g \in \mathcal{H}_X$ it holds that

$$\langle (a\mathbf{F} + b\mathbf{G})f, g\rangle_k = a\langle \mathbf{F}f, g\rangle_k + b\langle \mathbf{G}f, g\rangle_k = \langle f, (a\mathbf{F} + b\mathbf{G})g\rangle_k.$$

The subspace is closed. To see this let $\mathbf{F}$ be a limit of a sequence $\{\mathbf{F}_n\}_{n=1}^{\infty}$ of self-adjoint operators. For a given $f, g \in \mathcal{H}_X$ we have that

$$|\langle \mathbf{F}f, g\rangle_k - \langle \mathbf{F}_n f, g\rangle_k| \leq \|(\mathbf{F} - \mathbf{F}_n)f\|_k \|g\|_k \leq \|\mathbf{F} - \mathbf{F}_n\|_{\mathbf{\Xi}} \|\mathbf{\Xi}(f, f)\|_{op}^{1/2} \|g\|_k$$

and, as the operator $\mathbf{\Xi}(f, f)$ is bounded there exists for any upper bound $\epsilon > 0$ of the right side a N such that for all $n \geq N$ we have that $|\langle \mathbf{F}f, g\rangle_k - \langle \mathbf{F}_n f, g\rangle_k| < \epsilon$.

Using this we have that for arbitrary f, g

$$\langle \mathbf{F}f, g\rangle_k = \lim_{n\to\infty} \langle \mathbf{F}_n f, g\rangle_k = \lim_{n\to\infty} \langle f, \mathbf{F}_n g\rangle_k = \langle f, \mathbf{F}g\rangle_k$$

and $\mathbf{F}$ is also self-adjoint. □

C. Section 3: RKHS Integration Theory: Basic Transformations – Supplementary Results

C.1. Change of Measure

C.1.1. Absolute Continuity

We discuss now a way to test for a lack of absolute continuity and how to split the problem into the part of $\mathbb{Q}$ that is singular wrt. $\mathbb{P}$ and the absolute continuous part.

If $\mathbb{Q} \not\ll \mathbb{P}$ then there is a set on which $\mathbb{P}$ is zero while $\mathbb{Q}$ is not and there exists a strictly positive measurable function f – for example, the characteristic function for that set – for which $\mathbb{E}_{\mathbb{Q}} f > 0$, while $\mathbb{E}_{\mathbb{P}} f = 0$. Now, we have only control over RKHS functions and not arbitrary measurable functions, but we might consider the point-evaluators $k(x, \cdot)$ as a form of δ-function at x and test for $\mathbb{E}_{\mathbb{Q}} k(x, \cdot) > 0$, while $\mathbb{E}_{\mathbb{P}} k(x, \cdot) = 0$. If we consider the empirical version $\hat{m}_{\mathbb{Q}} = \sum_{i=1}^n k(y_i, \cdot)$ then $\hat{\mathbb{E}}_{\mathbb{Q}} k(x, \cdot) = \langle \hat{m}_{\mathbb{Q}}, k(x, \cdot) \rangle_k = \sum_{i=1}^n k(y_i, x) > 0$ implies $k(x, \cdot) \not\perp \{k(y_i, \cdot)\}_{i=1}^n$. So we might restrict our test for abs. continuity to the elements $\{k(y_i, \cdot)\}_{i=1}^n$ of which $\hat{m}_{\mathbb{Q}}$ is formed. If there is a $k(y_i, \cdot)$ which is perpendicular to every $k(x_j, \cdot)$, where $\hat{m}_{\mathbb{P}} = \sum_{j=1}^m k(x_j, \cdot)$ then we have a strong indicator that the empirical measures are not absolute continuous.

There are two effects here which might lead us to a wrong conclusion: (1) $k(y_i, \cdot)$ might take positive and negative values which cancel exactly when averaged over the empirical version of $\mathbb{P}$; (2) $\hat{\mathbb{E}}_{\mathbb{Q}} k(y_i, \cdot)$ might be 0 despite $k(y_i, \cdot)$ being an element of the sum defining $\hat{\mathbb{E}}_{\mathbb{Q}}$. So if the $k(y_i, \cdot)$ is a strictly positive function and $\hat{\mathbb{E}}_{\mathbb{Q}} k(y_i, \cdot) \neq 0$ then we know that for the empirical versions $\hat{\mathbb{Q}} \not\ll \hat{\mathbb{P}}$ holds.

We can split the sample into two parts, the $k(y_i, \cdot)$'s which we just discussed. These reflect the singular part of $\hat{\mathbb{Q}}$ wrt. to $\hat{\mathbb{P}}$. We can use the remaining samples to define $\hat{\mathbb{Q}}_a$, i.e. the absolute continuous part and estimate $\mathbf{R}$ for $\hat{\mathbb{Q}}_a$ and $\mathbb{P}$. One important point is that we do not have guarantees that $\hat{\mathbb{Q}}_a$ is in a measure theoretic sense absolute continuous as we test only with kernel functions if we can break absolute continuity and not with arbitrary measurable functions, i.e. the above statement is only a necessary condition for absolute continuity and not a sufficient one.

An interesting question is whether this can be turned into a proper test by increasing either the size of the RKHS, for example, by using a universal RKHS, or by making use of a bandwidth parameter which will decrease to 0 in the sample size.

C.2. Product Integral – Fubini

Integrals or expectations over product spaces $X \times Y$ are common in many applications. There are two settings that appear to be of broader interest: The case where we associate with X the RKHS $\mathcal{H}_X$ and with Y the RKHS $\mathcal{H}_Y$. Now, for $f \in \mathcal{H}_X, h \in \mathcal{H}_Y$ we like to take expectations over $f \times h$ with respect to a measure $\mathbb{P}_{X \times Y}$ on the product space. This case can be addressed with the help of the product RKHS $\mathcal{H}_X \otimes \mathcal{H}_Y$ that is introduced in Aronszajn (1950)[Sec. 8]. The RKHS $\mathcal{H}_X \otimes \mathcal{H}_Y$ has the reproducing kernel

$$p(x_1, y_1, x_2, y_2) = k(x_1, x_2) l(y_1, y_2). \tag{8}$$

We denote the RKHS with $\mathcal{H}_{X \times Y} := \mathcal{H}_X \otimes \mathcal{H}_Y$.

We have that $f \times h \in \mathcal{H}_{X \times Y}$ and expectations can be calculated in the usual way by replacing m_X with a suitable $m_{X \times Y} \in \mathcal{H}_{X \times Y}$, i.e. if $\mathcal{H}_{X \times Y} \subset L^2(X \times Y, \mathbb{P}_{X \times Y})$ and the corresponding expectation operator is bounded on $\mathcal{H}_{X \times Y}$ then the Riesz theorem guarantees us that such an element exists with which

$$\mathbb{E}_{X \times Y} f \times h = \langle m_{X \times Y}, f \times h \rangle_{X \times Y}.$$

It is often useful to reduce the product integral to two integrals with the help of the Fubini theorem. That is that, under suitable assumptions, $\mathbb{E}_{X \times Y} g(x, y) = \mathbb{E}_X \mathbb{E}_Y g(x, y)$.

For expectations over $g \in \mathcal{H}_X \otimes \mathcal{H}_Y$ we can do something similar. In case that $g(x, y) = f(x)h(y)$ for suitable $f \in \mathcal{H}_X, h \in \mathcal{H}_Y$, f, g, h are integrable and we have suitable representer $m_{X \times Y}, m_X, m_Y$ then the Fubini theorem

guarantees us that

$$\langle m_{X\times Y}, g\rangle_{X\times Y} = \mathbb{E}_{X\times Y} g(x,y) = \mathbb{E}_{X\times Y} f(x)h(y)$$
$$= \mathbb{E}_X f \mathbb{E}_Y h = \langle m_X, f\rangle_k \langle m_Y, h\rangle_l.$$

Note, that not every $g \in \mathcal{H}_X \otimes \mathcal{H}_Y$ needs to be of this particular form as $\mathcal{H}_X \otimes \mathcal{H}_Y$ is the completion of the direct product between $\mathcal{H}_X$ and $\mathcal{H}_Y$.

The second case of interest is when you have a kernel on the product space $X \times Y$ that does not arise from kernels on X and Y, i.e. the kernel $p(x_1, y_1, x_2, y_2)$ has not the form from eq. 8. This approach is also useful to deal with the limit points in $\mathcal{H}_X \otimes \mathcal{H}_Y$.

Expectations over elements g from the corresponding RKHS $\mathcal{H}_{X\times Y}$ can be taken like in the first case. The more interesting problem is to have a form of the Fubini theorem to turn the product integral into two separate integrals that can be efficiently evaluated using the RKHS framework. To do so we can use a kernel k on the space X to define an RKHS $\mathcal{H}_X$ and we try to approximate the inner integral, i.e. to find an operator $\mathbf{E} : \mathcal{H}_{X\times Y} \to \mathcal{H}_X$ such that

$$(\mathbf{E}g)(x) \approx \mathbb{E}_Y g(x,y).$$

The free variables are here x and g. Taking the supremum over the unit ball in $\mathcal{H}_{X\times Y}$ and the average over X wrt. $\mathbb{P}_X$ we get

$$\sup_{\|g\|_{X\times Y}\leq 1} \mathbb{E}_X(\mathbb{E}_Y g(x,y) - (\mathbf{E}g)(x))^2$$
$$\sup_{\|g\|_{X\times Y}\leq 1} \mathbb{E}_X(\mathbb{E}_Y \langle g, p(x,y,\cdot,\cdot) - \mathbf{E}^* k(x,\cdot)\rangle_{X\times Y})^2$$
$$\leq \mathbb{E}_{X\times Y} \|p(x,y,\cdot,\cdot) - \mathbf{E}^* k(x,\cdot)\|_{X\times Y}.$$

Using the usual regularised empirical version and $\mathbf{W}$ from eq. 2 we get the estimate

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

C.3. Conditional Expectation

The adjoint of the estimate we derived for the conditional expectation in eq. 5 is

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

with $\mathbf{W}$ defined in eq. 2. If we use $f = k(x,\cdot)$ we get

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

which is exactly the estimate $\mu(x)$ from Grünewälder et al. (2012a)[p. 4] with the vector-valued kernel $\Gamma(x,x') = k(x,x')\mathbf{I}$. Furthermore, we have that $\mathbf{E}[h] = \langle h, \mu(\cdot)\rangle_l$ and because $\mathbf{E}$ maps to $\mathcal{H}_X$ we know that $\langle h, \mu(\cdot)\rangle_l \in \mathcal{H}_X$.

This is also straight forward from a direct evaluation of $\mu(x)$ as

$$\langle h, \mu(x)\rangle_l = \sum_{i,j=1}^n k(x_i, x)\mathbf{W}_{ij} h(y_j) = \sum_{i=1}^n \beta_i k(x_i, x),$$

with $\beta_i = \sum_{j=1}^n \mathbf{W}_{ij} h(y_j)$.

More generally, one might consider the set $\mathcal{L} = \{\sum_{i=1}^n k(x_i,\cdot)h_i : n \in \mathbb{N}, x_i \in X, h_i \in \mathcal{H}_Y\}$ which is dense in the vector-valued RKHS $\mathcal{H}_\Gamma$. Because, elements $\mu \in \mathcal{L}$ are finite sums we have that

$$\langle h, \mu(x)\rangle_l = \sum_{i=1}^n k(x_i,\cdot)\langle h, h_i\rangle_l = \sum_{i=1}^n \alpha_i k(x_i,\cdot) \in \mathcal{H}_X$$

where $\alpha_i = \langle h, h_i \rangle_l$.

The more difficult question is if for the limit of functions in $\mathcal{L}$, i.e. functions $\mu \in \mathcal{H}_\Gamma$, it holds that $\langle h, \mu(\cdot)\rangle_l \in \mathcal{H}_Y$ for every $h \in \mathcal{H}_Y$. One might try to show for a Cauchy-sequence $\{\mu_n\}_{n=1}^\infty$ converging to μ in $\mathcal{H}_\Gamma$ that $\{\langle h, \mu_n(\cdot)\rangle_l\}_{n=1}^\infty$ is a Cauchy-sequence in $\mathcal{H}_X$, i.e. that

$$\|\langle h, \mu_n(\cdot)\rangle_l - \langle h, \mu_m(\cdot)\rangle_l\|_k = \|\langle h, \mu_n(\cdot) - \mu_m(\cdot)\rangle_l\|_k$$

is below a given ϵ after some finite number N. It is not directly obvious how to approach this. One might consider the Cauchy-Schwarz inequality, which tells us that $|\langle h, \mu_n(\cdot) - \mu_m(\cdot)\rangle_l| \leq \|h\|_l \|\mu_n(\cdot) - \mu_m(\cdot)\|_l$. Then one might show that $\|\mu_n(\cdot) - \mu_m(\cdot)\|_l$ is in $\mathcal{H}_X$ and try to prove that the norm of the upper bound is higher than the norm of the original sequence – this is not directly obvious as norms can measure different properties. In the operator approach these problems do not arise as by construction it is guaranteed that $\mathbf{E}h \in \mathcal{H}_X$ independent of $\mathbf{E}$ being a finite sum or a limit point in $\mathcal{H}_{\mathbf{\Xi}}$.

D. Section 4: RKHS Integration Theory: Composite Transformations – Proofs

D.1. Sum Rule – Change of Measure on Y

Theorem D.1. *We assume that the integrability assumptions from suppl. F hold, that $\mathbb{Q}_X \ll \mathbb{P}_X$ and the corresponding Radon-Nikodým derivative r is a.e. upper bounded by b we have with $c = \|\mathbf{A}^{1/2}\|_{op}^2 \|\mathbf{B}\|_{op}$ that*

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

Proof. Under our assumptions $\mathbb{E}_{\mathbb{Q}_X}$ has a representer $m_{\mathbb{Q}_X} \in \mathcal{H}_X$ due to the Riesz-theorem as each $f \in \mathcal{H}_X$ is integrable and $\mathbb{E}_{\mathbb{Q}_X}$ is bounded as $\mathbb{E}_{\mathbb{Q}_X} f \leq \|f\|_k \sqrt{k(x,x)}$. Using the transformation in eq. 6, we get

$$\begin{aligned}
\mathcal{E}_m[m_Y] &= \sup_{\|h\|_l \leq 1} (\mathbb{E}_{\mathbb{Q}_Y} h - \langle m_Y, h\rangle_l)^2 = \sup_{\|h\|_l \leq 1} (\mathbb{E}_{\mathbb{Q}_X} \mathbb{E}_{\mathbb{P}}[h|x] - \langle m_X, \mathbf{E}h\rangle_k)^2 \\
&= \sup_{\|h\|_l \leq 1} (\mathbb{E}_{\mathbb{Q}_X} \mathbb{E}_{\mathbb{P}}[h|x] - \mathbb{E}_{\mathbb{Q}_X} \mathbf{E}[h] + \mathbb{E}_{\mathbb{Q}_X} \mathbf{E}[h] - \langle m_X, \mathbf{E}h\rangle_k)^2 \\
&\leq \sup_{\|h\|_l \leq 1} \mathbb{E}_{\mathbb{Q}_X} (\mathbb{E}_{\mathbb{P}}[h|x] - \mathbf{E}[h])^2 + \sup_{\|h\|_l \leq 1} (\langle m_{\mathbb{Q}_X} - m_X, \mathbf{E}[h]\rangle_k)^2.
\end{aligned}$$

The first term can be transformed in case $\mathbb{Q}$ is absolute continuous wrt. $\mathbb{P}$. Assuming the corresponding Radon-Nikodým derivative $r(x)$ is a.e. upper bounded by b, we get:

$$\sup_{\|h\|_l \leq 1} \mathbb{E}_{\mathbb{Q}_X} (\mathbb{E}_{\mathbb{P}}[h|x] - \mathbf{E}[h])^2 = \sup_{\|h\|_l \leq 1} \mathbb{E}_{\mathbb{P}_X} r(x) (\mathbb{E}_{\mathbb{P}}[h|x] - \mathbf{E}[h])^2 \leq b \sup_{\|h\|_l \leq 1} \mathbb{E}_{\mathbb{P}_X} (\mathbb{E}_{\mathbb{P}}[h|x] - \mathbf{E}[h])^2 \leq b\mathcal{E}_c[\mathbf{E}]. \quad (9)$$

Using Micchelli & Pontil (2005)[Prop. 2.1 (f)], the second term can be bounded by

$$\begin{aligned}
&\sup_{\|h\|_l \leq 1} (\langle m_{\mathbb{Q}_X} - m_X, \mathbf{E}[h]\rangle_k)^2 \leq \sup_{\|h\|_l \leq 1} \langle m_{\mathbb{Q}_X} - m_X, \frac{\mathbf{E}[h]}{\|\mathbf{E}h\|_k} \rangle_k^2 \|\mathbf{E}h\|_k^2 \\
&\leq \sup_{\|f\|_k \leq 1} \langle m_{\mathbb{Q}_X} - m_X, f\rangle_k^2 \|\mathbf{E}\|_{\mathbf{\Xi}}^2 \sup_{\|h\|_l \leq 1} \|\mathbf{\Xi}(h,h)\|_{op} \leq \mathcal{E}_m[m_X] \|\mathbf{E}\|_{\mathbf{\Xi}}^2 \|\mathbf{A}^{1/2}\|_{op}^2 \sup_{\|h\|_l \leq 1} \|h\|_l^2 \|\mathbf{B}\|_{op} \\
&= c\|\mathbf{E}\|_{\mathbf{\Xi}}^2 \mathcal{E}_m[m_X],
\end{aligned}$$

with $c = \|\mathbf{A}^{1/2}\|_{op}^2 \|\mathbf{B}\|_{op}$.

In total, we get the upper bound

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

□

D.2. Kernel Bayes' Rule – Change of Measure on $X|y$

In the following we assume that $d = \sup_{x\in X} k(x,x) < \infty$, $c = \sup_{y\in Y} l(y,y) < \infty$. For the theorem we use subscripts at the risk functions to denote the measure with which they are evaluated, i.e. $\mathcal{E}_{c,\mathbb{Q}}$ for the conditional expectation risk evaluated wrt. $\mathbb{Q}$. The kernel function is here $\mathbf{\Xi}(h,h') := \langle h, \mathbf{A}h'\rangle_l \mathbf{B}$.

Theorem D.2. *We assume that the integrability assumptions from suppl. 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 $|\mathbb{E}_{y'\sim\mathbb{P}_Y} l(y,y') - \langle m_Y, l(y,\cdot)\rangle_l| \leq |\mathbb{E}_{y'\sim\mathbb{P}_Y} l(y,y')|/2$. We have that*

$$\mathcal{E}_{c,\mathbb{Q}}[\mathbf{G}] \leq d^2\mathcal{E}_{a,\mathbb{Q}}[l] + \mathcal{E}_{K,\mathbb{Q}}[\mathbf{G}] + \frac{4cd}{q^2}\left(\frac{c^2}{q^2}\left(b\mathcal{E}_{c,\mathbb{P}}[\mathbf{E}] + \|\mathbf{A}^{1/2}\|_{op}^2\|\mathbf{B}\|_{op}\|\mathbf{E}\|_{\mathbf{\Xi}}^2\mathcal{E}_{m,\mathbb{Q}}[m_X]\right) + b\mathcal{E}_{c,\mathbb{P}}[\mathbf{E}]\right),$$

in other words there exists a positive constant C such that

$$\mathcal{E}_{c,\mathbb{Q}}[\mathbf{G}] \leq \mathcal{E}_{K,\mathbb{Q}}[\mathbf{G}] + C\left(\mathcal{E}_{a,\mathbb{Q}}[l] + \|\mathbf{E}\|_{\mathbf{\Xi}}^2\mathcal{E}_{m,\mathbb{Q}}[m_X] + \mathcal{E}_{c,\mathbb{P}}[\mathbf{E}]\right).$$

Proof. In the following, we use the short form $\mathbb{E}_{Y'}$ for $\mathbb{E}_{y'\sim\mathbb{P}_Y}$.

(a) We follow the chain of arguments from Section 4.2. We use here the measure $\mathbb{Q}$. A change of measure is needed at the end to bound the error of $\mathbf{E}$. We have that

$$\begin{aligned}
\mathcal{E}_{c,\mathbb{Q}}[\mathbf{G}] &= \sup_{\|f\|_k\leq 1} \mathbb{E}_Y(\mathbb{E}[f|y] - \mathbf{G}[f](y))^2 \\
&\leq \sup_{\|f\|_k\leq 1} \mathbb{E}_Y\left(\mathbb{E}[f|y] - \mathbb{E}_{Y'}\frac{l(y,y')}{\mathbb{E}_{Y'}l(y,y')}\mathbb{E}[f|y']\right)^2 + \sup_{\|f\|_k\leq 1} \mathbb{E}_Y\left(\mathbb{E}_{Y'}\frac{l(y,y')}{\mathbb{E}_{Y'}l(y,y')}\mathbb{E}[f|y'] - \mathbf{G}[f](y)\right)^2 \\
&= d^2\mathcal{E}_{a,\mathbb{Q}}[l] + \sup_{\|f\|_k\leq 1} \mathbb{E}_Y\left(\frac{1}{\mathbb{E}_{Y'}l(y,y')}\mathbb{E}_X f\mathbb{E}_{Y'}[l(y,y')|x] - \mathbf{G}[f](y)\right)^2 \\
&= d^2\mathcal{E}_{a,\mathbb{Q}}[l] + \sup_{\|f\|_k\leq 1} \mathbb{E}_Y\left(\frac{1}{\mathbb{E}_{Y'}l(y,y')}\mathbb{E}_X f\mathbb{E}_{Y'}[l(y,y')|x] - \frac{1}{\langle m_Y, l(y,\cdot)\rangle_l}\mathbb{E}_X f\mathbf{E}[l(y,\cdot)](x)\right)^2 \\
&\qquad + \sup_{\|f\|_k\leq 1} \mathbb{E}_Y\left(\frac{1}{\langle m_Y, l(y,\cdot)\rangle_l}\mathbb{E}_X f\mathbf{E}[l(y,\cdot)](x) - \mathbf{G}[f](y)\right)^2 \\
&= d^2\mathcal{E}_{a,\mathbb{Q}}[l] + \sup_{\|f\|_k\leq 1} \mathbb{E}_Y\left(\frac{1}{\mathbb{E}_{Y'}l(y,y')}\mathbb{E}_X f\mathbb{E}_{Y'}[l(y,y')|x] - \frac{1}{\langle m_Y, l(y,\cdot)\rangle_l}\mathbb{E}_X f\mathbf{E}[l(y,\cdot)](x)\right)^2 + \mathcal{E}_{K,\mathbb{Q}}[\mathbf{G}].
\end{aligned}$$

We address the approximation error in (b), we verify in (c) that the integral transformation in the third line is valid and we bound the error of the middle term of the last line in (d),(e) and (f). Finally, the error for $\mathbf{E}$ can be bound in terms of the error $\mathcal{E}_{c,\mathbb{P}}[\mathbf{E}]$ wrt. the measure $\mathbb{P}$ from which we can sample. This is the part where the change of measure is used and the bound on the Radon-Nikodým derivative is needed. We derived the necessary bound already in eq. 9: $\mathcal{E}_{c,\mathbb{Q}}[\mathbf{E}] \leq b\mathcal{E}_{c,\mathbb{P}}[\mathbf{E}]$.

(b) We have that

$$\sup_{\|f\|_k\leq 1} \mathbb{E}_Y\left(\mathbb{E}[f|y] - \mathbb{E}_{Y'}\frac{l(y,y')}{\mathbb{E}_{Y'}l(y,y')}\mathbb{E}[f|y']\right)^2 \leq d^2 \sup_{\|h\|_{L^1(\mathbb{Q}_Y)}\leq 1} \mathbb{E}_Y\left(h(y) - \mathbb{E}_{Y'}\frac{l(y,y')}{\mathbb{E}_{Y'}l(y,y')}h(y')\right)^2 = d^2\mathcal{E}_{a,\mathbb{Q}}[l].$$

To see this we first observe that the conditional expectation $\mathbb{E}[f|y]$ is integrable wrt. $\mathbb{Q}_Y$ and

$$\|\mathbb{E}[f|y]\|_{L^1(\mathbb{Q}_Y)} \leq \mathbb{E}_{\mathbb{Q}_Y}\mathbb{E}[|f||y] \leq \sup_{x\in X}|f(x)| \leq \sup_{x\in X}\|f\|_k\sqrt{k(x,x)} \leq d\|f\|_k.$$

Hence, if we take the supremum over all $\mathbb{Q}_Y$ integrable functions h with norm $\|h\|_{L^1(\mathbb{Q}_Y)} \leq d$, we also include

every $\mathbb{E}[f|y]$ with $\|f\|_k = 1$. Finally, we can pull the scaling outside through

$$\sup_{\|h\|_{L^1(\mathbb{Q}_Y)}\leq d} \mathbb{E}_Y\left(h(y) - \mathbb{E}_{Y'}\frac{l(y,y')}{\mathbb{E}_{Y'}l(y,y')}h(y')\right)^2 = \sup_{\|\frac{h}{d}\|_{L^1(\mathbb{Q}_Y)}\leq 1} \mathbb{E}_Y\left(h(y) - \mathbb{E}_{Y'}\frac{l(y,y')}{\mathbb{E}_{Y'}l(y,y')}h(y')\right)^2$$
$$= \sup_{\|h\|_{L^1(\mathbb{Q}_Y)}\leq 1} d^2\mathbb{E}_Y\left(h(y) - \mathbb{E}_{Y'}\frac{l(y,y')}{\mathbb{E}_{Y'}l(y,y')}h(y')\right)^2.$$

(c) The integral transformation $\mathbb{E}_{Y'}l(y,y')\mathbb{E}[f|y'] = \mathbb{E}_X f\mathbb{E}_{Y'}[l(y,y')|x]$ is easy to verify. We have that $l(y,\cdot) \in \mathcal{H}_Y$ and by assumption is $\mathbb{Q}_Y$-integrable. Similarly, $f \in \mathcal{H}_X$ is $\mathbb{Q}_X$-integrable and, using (Fremlin, 2001)[253D], we have that $l(y,\cdot)\otimes f$ is $\mathbb{Q}_{X\times Y}$-integrable. Now, $\mathbb{E}_{X\times Y'}l(y,\cdot)\otimes f = \mathbb{E}_{X\times Y'}\mathbb{E}[l(y,\cdot)\otimes f|x] = \mathbb{E}_X f\mathbb{E}[l(y,\cdot)|x]$. With the same argument we have $\mathbb{E}_{X\times Y'}l(y,\cdot)\otimes f = \mathbb{E}_{Y'}l(y,\cdot)\mathbb{E}[f|y']$.

(d) We have that

$$(\mathbb{E}_X f(x)\mathbb{E}[l(y,\cdot)|x] - \mathbb{E}_X f(x)\mathbf{E}[l(y,\cdot)](x))^2 \leq cd\|f\|_k^2\mathcal{E}_{c,\mathbb{Q}}[\mathbf{E}].$$

This is essentially due to the Jensen inequality and the fact that $f^2(x) = |\langle f, k(x,\cdot)\rangle_k|^2 \leq \|f\|_k^2 k(x,x) = d\|f\|_k^2$:

$$(\mathbb{E}_X f(x)\mathbb{E}[l(y,\cdot)|x] - \mathbb{E}_X f(x)\mathbf{E}[l(y,\cdot)](x))^2 \leq \mathbb{E}_X f^2(x)(\mathbb{E}[l(y,\cdot)|x] - \mathbf{E}[l(y,\cdot)](x))^2$$
$$\leq d\|f\|_k^2\mathbb{E}_X(\mathbb{E}[l(y,\cdot)|x] - \mathbf{E}[l(y,\cdot)](x))^2 \leq d\|f\|_k^2 l(y,y)\mathcal{E}_{c,\mathbb{Q}}[\mathbf{E}].$$

(e) Building up on (d) we get the bound

$$\left(\frac{1}{\mathbb{E}_{Y'}l(y,\cdot)}\mathbb{E}_X f(x)\mathbb{E}[l(y,\cdot)|x] - \frac{1}{\langle m_Y, l(y,\cdot)\rangle_l}\mathbb{E}_X f(x)\mathbf{E}[l(y,\cdot)](x)\right)^2$$
$$\leq \frac{4cd\|f\|_k^2}{|\mathbb{E}_Y l(y,\cdot)|^2}\left(\frac{c^2}{|\mathbb{E}_{Y'}l(y,\cdot)|^2}\mathcal{E}_{m,\mathbb{Q}}[m_Y] + b\mathcal{E}_{c,\mathbb{Q}}[\mathbf{E}]\right).$$

P We first address the quotients. Let us denote for this part $e := \mathbb{E}_{Y'}l(y,y')$ and $o = \langle m_Y, l(y,\cdot)\rangle_l$. We have that

$$\left|\frac{1}{e} - \frac{1}{o}\right| = \frac{|e-o|}{|eo|} = \frac{|e-o|}{|e||e-(e-o)|} \leq \frac{|e-o|}{|e|\,||e|-|e-o|\,|}$$

and $|e-o|^2 = |\mathbb{E}_{Y'}l(y,\cdot) - \langle m_Y, l(y,\cdot)\rangle_l|^2 = \|l(y,\cdot)\|_l^2|\mathbb{E}_{Y'}\frac{l(y,\cdot)}{\|l(y,\cdot)\|_l} - \langle m_Y, \frac{l(y,\cdot)}{\|l(y,\cdot)\|_l}\rangle_l|^2 \leq c\mathcal{E}_{m,\mathbb{Q}}[m_Y]$. Furthermore, using the assumption that $|e-o| \leq |e|/2$ we get that

$$\left|\frac{1}{e} - \frac{1}{o}\right|^2 \leq \frac{|e-o|^2}{|e|^2\,||e|-|e-o|\,|^2} \leq \frac{4|e-o|^2}{|e|^4} \leq \frac{4c\mathcal{E}_{m,\mathbb{Q}}[m_Y]}{|\mathbb{E}_{Y'}l(y,y')|^4}.$$

Next we combine this with (c). We use that $|o| = |o-e-(-e)| \geq |\,|o-e|-|e|\,| \geq \frac{|e|}{2}$ under our assumption and that $|l(y,y')| \leq c$. The bound is now

$$\left(\frac{1}{e}\mathbb{E}_X f(x)\mathbb{E}[l(y,\cdot)|x] - \frac{1}{o}\mathbb{E}_X f(x)\mathbf{E}[l(y,\cdot)](x)\right)^2$$
$$\leq (\mathbb{E}_X f(x)\mathbb{E}[l(y,\cdot)|x])^2\left(\frac{1}{e} - \frac{1}{o}\right)^2 + \frac{1}{o^2}\left(\mathbb{E}_X f(x)\mathbb{E}[l(y,\cdot)|x] - \mathbb{E}_X f(x)\mathbf{E}[l(y,\cdot)](x)\right)^2$$
$$\leq \|f\|_k^2 dc^2\left(\frac{1}{e} - \frac{1}{o}\right)^2 + \frac{4}{|e|^2}\left(\mathbb{E}_X f(x)\mathbb{E}[l(y,\cdot)|x] - \mathbb{E}_X f(x)\mathbf{E}[l(y,\cdot)](x)\right)^2$$
$$\leq \frac{4dc^3}{|\mathbb{E}_{Y'}l(y,y')|^4}\|f\|_k^2\mathcal{E}_{m,\mathbb{Q}}[m_Y] + \frac{4cd}{|\mathbb{E}_{Y'}l(y,y')|^2}\|f\|_k^2\mathcal{E}_{c,\mathbb{Q}}[\mathbf{E}]$$
$$\leq \frac{4dc^3}{|\mathbb{E}_{Y'}l(y,y')|^4}\|f\|_k^2\mathcal{E}_{m,\mathbb{Q}}[m_Y] + \frac{4bcd}{|\mathbb{E}_{Y'}l(y,y')|^2}\|f\|_k^2\mathcal{E}_{c,\mathbb{P}}[\mathbf{E}]. \quad \mathbf{Q}$$

(f) The final step is to use the sum rule theorem to bound the error $\mathcal{E}_{m,\mathbb{Q}}[m_y]$ and to take the supremum over f and integrate wrt. $\mathbb{E}_Y$. Under our assumption that $\mathbb{E}_{Y'} l(y,y') > q$ this turns into

$$\sup_{\|f\|_k \leq 1} \mathbb{E}_Y \left(\frac{1}{\mathbb{E}_{Y'} l(y,y')} \mathbb{E}_X f(x) \mathbb{E}[l(y,\cdot)|x] - \frac{1}{\langle m_Y, l(y,\cdot)\rangle_l} \mathbb{E}_X f(x) \mathbf{E}[l(y,\cdot)](x) \right)^2 \leq \frac{4dc^3}{q^4} \mathcal{E}_{m,\mathbb{Q}}[m_Y] + \frac{4bcd}{q^2} \mathcal{E}_{c,\mathbb{P}}[\mathbf{E}]$$
$$\leq \frac{4dc^3}{q^4} \left(b\mathcal{E}_c[\mathbf{E}] + \|\mathbf{A}^{1/2}\|_{op}^2 \|\mathbf{B}\|_{op} \|\mathbf{E}\|_{\Xi}^2 \mathcal{E}_{m,\mathbb{Q}}[m_X] \right) + \frac{4bcd}{q^2} \mathcal{E}_{c,\mathbb{P}}[\mathbf{E}].$$

□

E. Convergence Rates for the approximate sum rule

We use Theorem 4.1 and we bound the involved risk term for the conditional expectation in the following subsections. We follow here the approach in Grünewälder et al. (2012a). This approach is based on vector-valued convergence rates from Caponnetto & De Vito (2007). One of the restrictions of these rates is that they need a finite dimensional space at one point. The next section contains assumptions which are needed to be able to apply the convergence results from Caponnetto & De Vito (2007).

Before we proceed we need to discuss convergence rates for the standard mean estimate $\mathcal{E}_m[m_X]$. Convergence rates are known for the convergence of the mean element in the RKHS norm. This implies convergence of the estimate in our risk function $\mathcal{E}_m$, but is actually a lot stronger than what we need. The convergence rates are under suitable assumptions in the order of $O(n^{-\alpha})$ with $0 < \alpha \leq 1/2$ (see Fukumizu et al. (2011) and references therein). We have rates of the order n^{-1} for the conditional expectation estimates for our risk function, and one might hypothesize that these rates are also achievable for $\mathcal{E}_m$, as conditional expectation estimation is a more difficult task and as our risk function is weaker than the RKHS norm. We do not derive new rates for the mean estimates, but leave it as a parameter in the theorem with an $\alpha \in]0,1]$. In particular, we assume that for a given measure $\mathbb{Q}_X$ and for any $\epsilon > 0$ there exists a constant C such that

$$\limsup_{n\to\infty} \mathbb{Q}^n[\mathcal{E}_m[m_X^n] > Cn^{-\alpha}] < \epsilon, \tag{10}$$

holds for any iid sample $\{x_i\}_{i=1}^n$. We use here the notation m_X^n to denote the n-sample mean estimate and $\mathbb{Q}^n$ to denote the product measure for n copies of $\mathbb{Q}_X$.

E.1. Assumptions

We assume that the spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ are finite dimensional. The assumption for $\mathcal{H}_Y$ is implied by the approach in Caponnetto & De Vito (2007) as there the output space of the regression problem must be finite dimensional. For simplicity we also assume that $\mathcal{H}_X$ is finite dimensional, however, this assumption can be dropped with some extra effort.

We also assume that the kernel is measurable, that is that for arbitrary $h, h' \in \mathcal{H}_Y$ that the mapping: $(f,g) \mapsto \langle h, \Xi(f,g)h'\rangle_Y$ is measurable. Furthermore, we assume that $\|l(y,\cdot)\|_l^2$ is measurable wrt. y and in general that the integrability assumptions from F hold.

We need specific assumptions for the conditional expectation estimation problem. For this we assume that a minimiser of the regression problem exists, that is, that there exists a $\mathbf{E}_s \in \mathcal{H}_\Xi$ such that $\mathcal{E}_s[\mathbf{E}_s] = \inf_{\mathbf{E}\in\mathcal{H}_\Xi} \mathcal{E}_s[\mathbf{E}]$.

E.2. Rates for the Conditional Expectation

In this section we derive risk bounds and convergence rates for the natural risk function

$$\mathcal{E}_c[\mathbf{E}] = \sup_{\|h\|_l \leq 1} \mathbb{E}_X (\mathbb{E}[h|x] - \mathbf{E}[h](x))^2.$$

The approach we take is to derive convergence rates for the surrogate risk function

$$\mathcal{E}_s[\mathbf{E}] = \mathbb{E}_{X\times Y} \|l(y,\cdot) - \mathbf{E}^* k(x,\cdot)\|_l^2,$$

which can be done by a direct application of vector-valued regression rates and to link the two cost functions. We start by linking the two cost functions in the next.

E.2.1. Relating the Risk Functions

We reproduce now Theorem A.2 and A.3 from Grünewälder et al. (2012a) for our setting. The derivation is – modulo minor adaptations – like in Grünewälder et al. (2012a) and we include the proofs mainly for completeness. Also note that the approach is based on a conditional expectation argument as discussed is Supp. A.1.

Lemma E.1. *We assume that the integrability assumptions from suppl. F hold. If there exists* $\mathbf{E}_* \in \mathcal{H}_{\mathbf{\Xi}}$ *such that for any* $h \in \mathcal{H}_Y$*:* $\mathbb{E}[h|x] = \mathbf{E}_*[h](x)$ $\mathbb{P}_X$*-a.s., then for any* $\mathbf{E} \in \mathcal{H}_{\mathbf{\Xi}}$*:*

$$(i) \quad \mathbb{E}_{X\times Y}\mathbf{E}_*[l(y,\cdot)](x) = \mathbb{E}_X\|\mathbf{E}_*^* k(x,\cdot)\|_l^2,$$
$$(ii) \quad \mathbb{E}_{X\times Y}\mathbf{E}[l(y,\cdot)](x) = \mathbb{E}_X\langle \mathbf{E}_*^* k(x,\cdot), \mathbf{E}^* k(x,\cdot)\rangle_l.$$

Proof. (i) follows from (ii) by setting $\mathbf{E} := \mathbf{E}_*$. Using the assumption (ii) can be derived:

$$\mathbb{E}_X\langle \mathbf{E}_*^* k(x,\cdot), \mathbf{E}^* k(x,\cdot)\rangle_l = \mathbb{E}_X\langle k(x,\cdot), \mathbf{E}_*\mathbf{E}^* k(x,\cdot)\rangle_l = \mathbb{E}_X\mathbf{E}_*[\mathbf{E}^* k(x,\cdot)](x)$$
$$= \mathbb{E}_X\mathbb{E}_Y[\mathbf{E}^*[k(x,\cdot)](y)|x] = \mathbb{E}_{X\times Y}\langle l(y,\cdot), \mathbf{E}^* k(x,\cdot)\rangle_l = \mathbb{E}_{X\times Y}\mathbf{E}[l(y,\cdot)](x).$$

□

Theorem E.1. *If there exists a* $\mathbf{E}_* \in \mathcal{H}_{\mathbf{\Xi}}$ *such that for any* $h \in \mathcal{H}_Y$ *it holds that* $\mathbb{E}[h|x] = \mathbf{E}_*[h](x)$ $\mathbb{P}_X$*-a.s. then* $\mathbf{E}_*$ *is a solution of* $\operatorname{argmin}_{\mathbf{E}\in\mathcal{H}_{\mathbf{\Xi}}} \mathcal{E}_c[\mathbf{E}]$ *and* $\operatorname{argmin}_{\mathbf{E}\in\mathcal{H}_{\mathbf{\Xi}}} \mathcal{E}_s[\mathbf{E}]$*. Furthermore, any solution* $\mathbf{E}_\circ$ *of either of the two risk functions fulfills*

$$\mathbf{E}_* k(x,\cdot) = \mathbf{E}_\circ k(x,\cdot) \quad \mathbb{P}_X\text{-a.s.}$$

In particular, if k *is continuous and for any open set* $B \neq \emptyset$ *it holds that* $\mathbb{P}_X B > 0$ *then the minimisers of the two risk functions are equal to* $\mathbf{E}_*$*.*

Proof. We start by showing that the right side is minimised by $\mathbf{E}_*$ using the above lemma. Let $\mathbf{E}$ be any element in $\mathcal{H}_{\mathbf{\Xi}}$ then we have

$$\mathbb{E}_{X\times Y}\|l(y,\cdot) - \mathbf{E}^* k(x,\cdot)\|_l^2 - \mathbb{E}_{X\times Y}\|l(y,\cdot) - \mathbf{E}_*^* k(x,\cdot)\|_l^2$$
$$= \mathbb{E}_X\|\mathbf{E}^* k(x,\cdot)\|_l^2 - 2\mathbb{E}_{X\times Y}\mathbf{E}[l(y,\cdot)](x) + 2\mathbb{E}_{X\times Y}\mathbf{E}_*[l(y,\cdot)](x) - \mathbb{E}_X\|\mathbf{E}_*^* k(x,\cdot)\|_l^2$$
$$= \mathbb{E}_X\|\mathbf{E}^* k(x,\cdot)\|_l^2 - 2\mathbb{E}_X\langle \mathbf{E}_*^* k(x,\cdot), \mathbf{E}^* k(x,\cdot)\rangle_l + \mathbb{E}_X\|\mathbf{E}_*^* k(x,\cdot)\|_l^2 = \mathbb{E}_X\|\mathbf{E}^* k(x,\cdot) - \mathbf{E}_*^* k(x,\cdot)\|_l^2 \geq 0.$$

Hence, $\mathbf{E}_*$ is a minimiser of the surrogate risk functional. The minimiser is furthermore $\mathbb{P}_X$-a.s. unique: Assume there is a second minimiser $\mathbf{E}_\circ$ then above calculation shows that

$$0 = \mathbb{E}_{X\times Y}\|l(y,\cdot) - \mathbf{E}_\circ^* k(x,\cdot)\|_l^2 - \mathbb{E}_{X\times Y}\|l(y,\cdot) - \mathbf{E}_*^* k(x,\cdot)\|_l^2 = \mathbb{E}_X\|\mathbf{E}_*^* k(x,\cdot) - \mathbf{E}_\circ^* k(x,\cdot)\|_l^2.$$

Thus, $\|\mathbf{E}_*^* k(x,\cdot) - \mathbf{E}_\circ^* k(x,\cdot)\|_l = 0$ $\mathbb{P}_X$-a.s. (Fremlin, 2000)[122Rc], i.e. a measurable set M with $\mathbb{P}_X M = 1$ exists such that $\|\mathbf{E}_*^* k(x,\cdot) - \mathbf{E}_\circ^* k(x,\cdot)\|_l = 0$ holds for all $x \in M$. As $\|\cdot\|_l$ is a norm we have that $\mathbf{E}_*^* k(x,\cdot) = \mathbf{E}_\circ^* k(x,\cdot)$ $\mathbb{P}_X$-a.s. Now, let k be continuous, let $\mathbb{P}_X B > 0$ for any open set $B \neq \emptyset$ then and assume that there exists a point x such that $\mathbf{E}_*^* k(x,\cdot) \neq \mathbf{E}_\circ^* k(x,\cdot)$. Now, as $\mathbf{E}_*, \mathbf{E}_\circ$ and k are continuous there exists an open set B around x such that $\mathbf{E}_*^* k(x',\cdot) \neq \mathbf{E}_\circ^* k(x',\cdot)$ for all $x' \in B$ and, as $\mathbb{P}B > 0$, $\mathbf{E}_*^* k(x,\cdot) = \mathbf{E}_\circ^* k(x,\cdot)$ does not hold $\mathbb{P}_X$-a.s. with contradiction to the above. Now, Theorem 2.2 tells us that $\mathbf{E}_*^* = \mathbf{E}_\circ^*$ and because the adjoint identifies the operators uniquely we have $\mathbf{E}_* = \mathbf{E}_\circ$.

Now, for the minimisers of the natural risk function we first observe that for every $h \in \mathcal{H}_Y$, $\mathbb{E}_X(\mathbb{E}[h|x] - \mathbf{E}_*[h](x))^2 = 0$ by assumption and $\mathbf{E}_*$ is a minimiser. Uniqueness can be seen in the following way: Assume there is a second minimiser $\mathbf{E}_\circ$ then for all $h \in \mathcal{H}_Y$ we have

$$\mathbb{E}_X(\langle h, \mathbf{E}_\circ^* k(x,\cdot) - \mathbf{E}_*^* k(x,\cdot)\rangle_l)^2 \leq \mathbb{E}_X(\langle h, \mathbf{E}_\circ^* k(x,\cdot)\rangle_l - \mathbb{E}[h|x])^2 + \mathbb{E}_X(\mathbb{E}[h|x] - \langle h, \mathbf{E}_*^* k(x,\cdot)\rangle_l)^2 = 0.$$

Hence, $\langle h, \mathbf{E}_\circ^* k(x,\cdot) - \mathbf{E}_*^* k(x,\cdot)\rangle_l = 0$ $\mathbb{P}_X$-a.s. (Fremlin, 2000)[122Rc], i.e. a measurable set M with $\mathbb{P}_X M = 1$ exists such that $\langle h, \mathbf{E}_\circ^* k(x,\cdot) - \mathbf{E}_*^* k(x,\cdot)\rangle_l = 0$ holds for all $x \in M$. Assume that there exists a $x' \in M$ such that $\mathbf{E}_\circ^* k(x',\cdot) \neq \mathbf{E}_*^* k(x',\cdot)$ then pick $h := \mathbf{E}_\circ^* k(x',\cdot) - \mathbf{E}_*^* k(x',\cdot)$ and we have $0 = \langle h, \mathbf{E}_\circ^* k(x',\cdot) - \mathbf{E}_*^* k(x',\cdot)\rangle_l = \|\mathbf{E}_\circ^* k(x',\cdot) - \mathbf{E}_*^* k(x',\cdot)\|_l > 0$ as $\|\cdot\|_l$ is a norm. By contradiction we get , $\mathbf{E}_\circ^* k(x,\cdot) = \mathbf{E}_*^* k(x,\cdot)$ $\mathbb{P}_X$-a.s. With the same argument as in case one we can follow equivalence of the operators. □

The next theorem is the main theorem. It allows us to use convergence rates of the surrogate risk to infer convergence rates for the natural risk. Furthermore, it weakens the assumptions. The price we have to pay for this is an approximation error term.

Theorem E.2. *Let* $C = \|\mathbf{A}^{1/2}\|_{op}\|\mathbf{B}\|_{op}^{1/2}\sup_{x\in X}\sqrt{k(x,x)}$ *and assume that there exists an* $\eta > 0$ *and* $\mathbf{E}_* \in \mathcal{H}_{\mathbf{\Xi}}$ *such that* $\sup_{\|\mathbf{E}\|_{\mathbf{\Xi}}\leq 1}\mathbb{E}_X\left[\mathbb{E}[\mathbf{E}^*k(x,\cdot)|x] - \langle \mathbf{E}^*k(x,\cdot), \mathbf{E}_*^*k(x,\cdot)\rangle_l\right]^2 = \eta < \infty$. *Furthermore, let* $\mathbf{E}_s$ *be a minimiser of the surrogate risk and let* $\mathbf{E}_\circ$ *be an arbitrary element in* $\mathcal{H}_{\mathbf{\Xi}}$. *With* $\mathcal{E}_s[\mathbf{E}_\circ] \leq \mathcal{E}_s[\mathbf{E}_s] + \delta$ *we have*

$$(i)\quad \mathcal{E}_c[\mathbf{E}_s] \leq \left(\sqrt{\mathcal{E}_c[\mathbf{E}_*]} + \eta^{1/4}\sqrt{8C(\|\mathbf{E}_*\|_{\mathbf{\Xi}} + \|\mathbf{E}_s\|_{\mathbf{\Xi}})}\right)^2,$$

$$(ii)\quad \mathcal{E}_c[\mathbf{E}_\circ] \leq \left(\sqrt{\mathcal{E}_c[\mathbf{E}_*]} + \eta^{1/4}\sqrt{8C(\|\mathbf{E}_*\|_{\mathbf{\Xi}} + \|\mathbf{E}_\circ\|_{\mathbf{\Xi}})} + \delta^{1/2}\right)^2.$$

Proof. First, observe that if $\mathbf{E} \in \mathcal{H}_{\mathbf{\Xi}}$ then we have due to the Jensen inequality

$$|\mathbb{E}_X\mathbb{E}[\mathbf{E}^*k(x,\cdot)|x] - \mathbb{E}_X\langle \mathbf{E}^*k(x,\cdot), \mathbf{E}_*^*k(x,\cdot)\rangle_l| \leq \|\mathbf{E}^*k(x,\cdot)\|_l\mathbb{E}_X\left|\mathbb{E}\left[\frac{\mathbf{E}^*k(x,\cdot)}{\|\mathbf{E}^*k(x,\cdot)\|_l}\middle| x\right] - \langle\frac{\mathbf{E}^*k(x,\cdot)}{\|\mathbf{E}^*k(x,\cdot)\|_l}, \mathbf{E}_*^*k(x,\cdot)\rangle_l\right|$$

$$\leq \|\mathbf{E}^*k(x,\cdot)\|_l\sqrt{\mathbb{E}_X\left(\mathbb{E}\left[\frac{\mathbf{E}^*k(x,\cdot)}{\|\mathbf{E}^*k(x,\cdot)\|_l}\middle| x\right] - \langle\frac{\mathbf{E}^*k(x,\cdot)}{\|\mathbf{E}^*k(x,\cdot)\|_l}, \mathbf{E}_*^*k(x,\cdot)\rangle_l\right)^2} = \|\mathbf{E}^*k(x,\cdot)\|_l\sqrt{\eta}.$$

We can now reproduce the proof of Lemma E.1 with an approximation error. For any $\mathbf{E} \in \mathcal{H}_{\mathbf{\Xi}}$ we have

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

In particular,

$$|\mathbb{E}_{X\times Y}\langle l(y,\cdot), \mathbf{E}_*^*k(x,\cdot)\rangle_l - \mathbb{E}_{X\times Y}\|\mathbf{E}_*^*k(x,\cdot)\|_l^2| \leq \|\mathbf{E}_*^*k(x,\cdot)\|_l\sqrt{\eta}.$$

Like in the proof of Theorem E.1 we have for any $\mathbf{E}$ that

$$\begin{aligned}&\mathcal{E}_s[\mathbf{E}] - \mathcal{E}_s[\mathbf{E}_*] = \mathbb{E}_{X\times Y}\|l(y,\cdot) - \mathbf{E}^*k(x,\cdot)\|_l^2 - \mathbb{E}_{X\times Y}\|l(y,\cdot) - \mathbf{E}_*^*k(x,\cdot)\|_l^2\\ &\geq \mathbb{E}_X\|\mathbf{E}^*k(x,\cdot)\|_l^2 - 2\mathbb{E}_X\langle \mathbf{E}_*^*k(x,\cdot), \mathbf{E}^*k(x,\cdot)\rangle_l + \mathbb{E}_X\|\mathbf{E}_*^*k(x,\cdot)\|_l^2 - 2\|\mathbf{E}_*^*k(x,\cdot)\|_l\sqrt{\eta} - 2\|\mathbf{E}^*k(x,\cdot)\|_l\sqrt{\eta}\\ &= \mathbb{E}_X\|\mathbf{E}_*^*k(x,\cdot) - \mathbf{E}^*k(x,\cdot)\|_l^2 - 2\sqrt{\eta}(\|\mathbf{E}_*^*k(x,\cdot)\|_l + \|\mathbf{E}^*k(x,\cdot)\|_l).\end{aligned} \tag{11}$$

In particular, $|\mathcal{E}_s[\mathbf{E}] - \mathcal{E}_s[\mathbf{E}_*]| \geq \mathcal{E}_s[\mathbf{E}] - \mathcal{E}_s[\mathbf{E}_*] \geq \mathbb{E}_X\|\mathbf{E}_*^*k(x,\cdot) - \mathbf{E}^*k(x,\cdot)\|_l^2 - 2\sqrt{\eta}(\|\mathbf{E}_*^*k(x,\cdot)\|_l + \|\mathbf{E}^*k(x,\cdot)\|_l)$ and hence

$$\mathbb{E}_X\|\mathbf{E}_*^*k(x,\cdot) - \mathbf{E}^*k(x,\cdot)\|_l^2 \leq |\mathcal{E}_s[\mathbf{E}] - \mathcal{E}_s[\mathbf{E}_*]| + 2\sqrt{\eta}(\|\mathbf{E}_*^*k(x,\cdot)\|_l + \|\mathbf{E}^*k(x,\cdot)\|_l). \tag{12}$$

We can now bound the error $\mathcal{E}_c[\mathbf{E}]$ in dependence of how similar $\mathbf{E}$ is to $\mathbf{E}_*$ in the surrogate cost function $\mathcal{E}_s$:

$$\sqrt{\mathcal{E}_c[\mathbf{E}]} \leq \sqrt{\mathcal{E}_c[\mathbf{E}_*]} + \sup_{\|h\|_l\leq 1}\sqrt{\mathbb{E}_X\left[\langle h, \mathbf{E}_*^*k(x,\cdot) - \mathbf{E}^*k(x,\cdot)\rangle_l\right]^2} \leq \sqrt{\mathcal{E}_c[\mathbf{E}_*]} + \sqrt{\mathbb{E}_X\left(\frac{\|\mathbf{E}_*^*k(x,\cdot) - \mathbf{E}^*k(x,\cdot)\|_l^2}{\|\mathbf{E}_*^*k(x,\cdot) - \mathbf{E}^*k(x,\cdot)\|_l}\right)^2}$$

$$\leq \sqrt{\mathcal{E}_c[\mathbf{E}_*]} + \eta^{1/4}\sqrt{2(\|\mathbf{E}_*^*k(x,\cdot)\|_l + \|\mathbf{E}^*k(x,\cdot)\|_l)} + \sqrt{|\mathcal{E}_s[\mathbf{E}] - \mathcal{E}_s[\mathbf{E}_*]|}, \tag{13}$$

where we used the triangular inequality, we used that $\langle\frac{\mathbf{E}_*^*k(x,\cdot)-\mathbf{E}^*k(x,\cdot)}{\|\mathbf{E}_*^*k(x,\cdot)-\mathbf{E}^*k(x,\cdot)\|_l}, \mathbf{E}_*^*k(x,\cdot)-\mathbf{E}^*k(x,\cdot)\rangle_l \geq \langle h, \mathbf{E}_*^*k(x,\cdot)-\mathbf{E}^*k(x,\cdot)\rangle_l$ for any h with $\|h\|_l \leq 1$ and eq. 12.

Now, for $\mathbf{E} := \mathbf{E}_s$ observe that $\mathcal{E}_s[\mathbf{E}_s] + 2\sqrt{\eta}(\|\mathbf{E}_*^*k(x,\cdot)\|_l + \|\mathbf{E}_s^*k(x,\cdot)\|_l) \geq \mathcal{E}_s[\mathbf{E}_*]$ follows from eq. (11) and as $\mathbf{E}_s$ is a $\mathcal{E}_s$ minimiser we have $|\mathcal{E}_s[\mathbf{E}_*] - \mathcal{E}_s[\mathbf{E}_s]| \leq 2\sqrt{\eta}(\|\mathbf{E}_*^*k(x,\cdot)\|_l + \|\mathbf{E}^*k(x,\cdot)\|_l)$ and from eq. 13 we get

$$\sqrt{\mathcal{E}_c[\mathbf{E}_s]} \leq \sqrt{\mathcal{E}_c[\mathbf{E}_*]} + \eta^{1/4}\sqrt{8(\|\mathbf{E}_*^*k(x,\cdot)\|_l + \|\mathbf{E}_s^*k(x,\cdot)\|_l)}.$$

Furthermore, with $\|\mathbf{E}^*k(x,\cdot)\|_l \leq \|\mathbf{E}\|_{\mathbf{\Xi}}\|\mathbf{A}^{1/2}\|_{op}\|\mathbf{B}\|_{op}^{1/2}\sqrt{k(x,x)}$ we have

$$\sqrt{\mathcal{E}_c[\mathbf{E}_s]} \leq \sqrt{\mathcal{E}_c[\mathbf{E}_*]} + \eta^{1/4}\sqrt{8C(\|\mathbf{E}_*\|_{\mathbf{\Xi}} + \|\mathbf{E}_s\|_{\mathbf{\Xi}})}.$$

Similarly, for $\mathbf{E} := \mathbf{E}_\circ$ we have

$$\sqrt{\mathcal{E}_c[\mathbf{E}_\circ]} \leq \sqrt{\mathcal{E}_c[\mathbf{E}_*]} + \eta^{1/4}\sqrt{8C(\|\mathbf{E}_*\|_{\mathbf{\Xi}} + \|\mathbf{E}_\circ\|_{\mathbf{\Xi}})} + \delta^{1/2}.$$

□

E.2.2. Convergence Rates for the Surrogate Risk

The surrogate risk is a standard vector-valued risk function for which convergence rates are known under certain assumptions (Caponnetto & De Vito, 2007). This was used in Grünewälder et al. (2012a) to derive rates for conditional expectation estimates. We can do the same in our setting. With the n-sample estimate being denoted with $\mathbf{E}_n$ we have:

Theorem E.3. *Under assumptions E.1 we have that for every $\epsilon > 0$ there exists a constant C such that*

$$\limsup_{n\to\infty} \sup_{\mathbb{P}\in\mathfrak{P}} \mathbb{P}^n \left[\mathcal{E}_s[\mathbf{E}_n] - \mathcal{E}_s[\mathbf{E}_s] > Cn^{-1}\right] < \epsilon.$$

Proof. We only need to verify the assumptions in Caponnetto & De Vito (2007) to apply Theorem 2 from the same paper. Most of the verifications below are generic, however, there is one important point. The input space for the regression problem needs to be bounded in a suitable sense. If we use the full space $\mathcal{H}_X$ here then this is obviously not bounded. However, for the conditional expectation estimate we do not observe arbitrary $\mathcal{H}_X$ functions, but only functions $k(x,\cdot)$ and, due to our assumptions, $k(x,\cdot)$ is bounded. We hence use a bounded and closed ball $B_X \subset \mathcal{H}_X$, which contains all $k(x,\cdot)$, as the input space.

(a) The first assumption concerns the space B_X. B_X must be a Polish space, that is a separable completely metrizable topological space. $\mathcal{H}_X$ is finite dimensional, hence separable and a Polish space and B_X as a closed subset is too.

(b) $\mathcal{H}_Y$ must be a separable Hilbert space. Like in (a) this is fulfilled.

We continue with *Hypothesis 1* from Caponnetto & De Vito (2007).

(c) The space $\mathcal{H}_{\mathbf{\Xi}}$ is separable. **P** Let $\{e_i\}_{i=1}^n$ be a basis of $\mathcal{H}_X$ and $\{g_i\}_{i=1}^m$ be a basis of $\mathcal{H}_Y$. Now, for any $N \in \mathbb{N}$ we have that

$$\sum_{t=1}^N \langle\cdot, \mathbf{A}f_t\rangle_k \mathbf{B}h_t = \sum_{t=1}^N \langle\cdot, \mathbf{A}\sum_{i=1}^n a_{it}e_i\rangle_k \mathbf{B}\sum_{j=1}^m b_{jt}g_j = \sum_{i=1}^n\sum_{j=1}^m \langle\cdot, \mathbf{A}e_i\rangle_k \mathbf{B}g_j \left(\sum_{t=1}^N a_{it}b_{jt}\right) = \sum_{i=1}^n\sum_{j=1}^m c_{ij}\langle\cdot, \mathbf{e}_i\rangle_k \mathbf{g}_j,$$

where $f_t = \sum_{i=1}^n a_{it}e_i, h_t = \sum_{j=1}^m b_{jt}g_j$, $\mathbf{e}_i = \mathbf{A}e_i, \mathbf{g}_j = \mathbf{B}g_j$ and $c_{ij} = \sum_{t=1}^N a_{it}b_{jt} \in \mathbb{R}$.

We have for two such finite sums $\mathbf{F} = \sum_{i=1}^n\sum_{j=1}^m c_{ij}\langle\cdot, \mathbf{e}_i\rangle_k \mathbf{g}_j, \mathbf{G} = \sum_{i=1}^n\sum_{j=1}^m d_{ij}\langle\cdot, \mathbf{e}_i\rangle_k \mathbf{g}_j \in \mathcal{H}_{\mathbf{\Xi}}$ that

$$\|\mathbf{F}-\mathbf{G}\|_{\mathbf{\Xi}} = \|\sum_{i=1}^n\sum_{j=1}^m (c_{ij}-d_{ij})\langle\cdot, \mathbf{e}_i\rangle_k \mathbf{g}_j\|_{\mathbf{\Xi}} \leq \sum_{i=1}^n\sum_{j=1}^m \|\langle\cdot, \mathbf{e}_i\rangle_k \mathbf{g}_j\|_{\mathbf{\Xi}} |c_{ij}-d_{ij}| = \sum_{i=1}^n\sum_{j=1}^m w_{ij}|c_{ij}-d_{ij}|$$

with $w_{ij} = \|\langle\cdot, \mathbf{e}_i\rangle_k \mathbf{g}_j\|_{\mathbf{\Xi}} \geq 0$. Now $|\sum_{i=1}^n\sum_{j=1}^m w_{ij}|c_{ij}-d_{ij}| \leq \max_{i,j} w_{ij} \sum_{i=1}^n\sum_{j=1}^m |c_{ij}-d_{ij}|$ and we can use a countable cover of $\mathbb{R}^{nm}$ to approximate arbitrary operators represented as finite sums. As these operators are dense in $\mathcal{H}_{\mathbf{\Xi}}$ we also gain a countable cover for $\mathcal{H}_{\mathbf{\Xi}}$ and $\mathcal{H}_{\mathbf{\Xi}}$ is separable. **Q** Restricting to B_X instead of $\mathcal{H}_X$ does not change the argument.

(d) The next assumption concerns point evaluation. There exists for every $f \in B_X \subset \mathcal{H}_X$ an operator $(\mathbf{\Xi}_f)^* : \mathcal{H}_{\mathbf{\Xi}} \to \mathcal{H}_Y$ such that for any $\mathbf{F} \in \mathcal{H}_{\mathbf{\Xi}}$ it holds that $\mathbf{F}f = (\mathbf{\Xi}_f)^*\mathbf{F}$. This operator is the adjoint of the operator $\mathbf{\Xi}_f$ that we defined in Section 2.4. We have that this operator $(\mathbf{\Xi}_f)^*$ is a Hilbert-Schmidt operator. **P** We have that $(\mathbf{\Xi}_f)^*$ is a Hilbert-Schmidt operator if $\mathbf{\Xi}_f$ is and in this case both have the same Hilbert-Schmidt norm which is for a given basis $\{e_i\}_{i=1}^m$ of $\mathcal{H}_Y$

$$\sum_{i=1}^m \|\mathbf{\Xi}_f e_i\|_{\mathbf{\Xi}}^2 = \sum_{i=1}^m \langle e_i, \mathbf{\Xi}(f,f)e_i\rangle_l = \langle f, \mathbf{A}f\rangle_k \sum_{i=1}^m \langle e_i, \mathbf{B}e_i\rangle_l$$

finite as $\mathbf{A}$ and $\mathbf{B}$ are bounded. Hence, both operators are Hilbert-Schmidt operators. **Q**

(e) The trace of $(\mathbf{\Xi}_f)^*\mathbf{\Xi}_f$ must have a common upper bound for all $f \in B_X$. This is the point where we need the boundedness assumption of $k(x,\cdot)$. For a basis $\{e_i\}_{i=1}^m$ of $\mathcal{H}_Y$ we have that

$$\mathrm{Tr}[(\mathbf{\Xi}_f)^*\mathbf{\Xi}_f] = \sum_{i=1}^m \langle \mathbf{\Xi}_f e_i, \mathbf{\Xi}_f e_i\rangle_l = \sum_{i=1}^m \langle e_i, \mathbf{\Xi}(f,f)e_i\rangle_l \leq \|\mathbf{A}^{1/2}f\|_k^2 \sum_{i=1}^m \langle e_i, \mathbf{B}e_i\rangle_l$$

which is bounded as f is bounded.

The final assumptions we need to verify are the ones in *Hypothesis 2* from Caponnetto & De Vito (2007).

(f) The output data for this regression problem is concentrated on the set $\{l(y,\cdot) : y \in Y\}$ for which we have $\|l(y,\cdot)\|_l^2 = l(y,y) < \infty$, and, as by assumption $\|l(y,\cdot)\|_l^2$ is measurable we have that $\|l(y,\cdot)\|_l^2$ is integrable.

(g) The final assumption concerns the conditional distribution $\mathbb{P}_{Y|x}$. We have $\|l(y,\cdot) - \mathbf{E}_s^*[k(x,\cdot)]\|_l \leq \sqrt{l(y,y)} + \|\mathbf{E}_s^*[k(x,\cdot)]\|_l \leq \sqrt{l(y,y)} + C\|k(x,\cdot)\|_k = \sqrt{l(y,y)} + C\sqrt{k(x,x)}$ with a constant C as $\mathbf{E}_s^*$ is a bounded operator. This norm is hence bounded by assumption that the kernels are bounded. As we assumed also that all our $\mathcal{H}_X$ and $\mathcal{H}_Y$ functions are integrable we have that the following expectation is well defined

$$\mathbb{E}_{Y|x}\left[\exp \|l(y,\cdot) - \mathbf{E}_s^*[k(x,\cdot)]\|_l - \|l(y,\cdot) - \mathbf{E}_s^*[k(x,\cdot)]\|_l - 1\right]$$

and bounded. This implies Assumption 9 in Caponnetto & De Vito (2007) as our observations are concentrated on $\{l(y,\cdot) : y \in Y\}$. □

E.2.3. Convergence Rates for the Natural Risk

We now combine the upper bound argument with the convergence rate for the upper bound.

Theorem E.4. *Let $C = \|\mathbf{A}^{1/2}\|_{op}\|\mathbf{B}\|_{op}^{1/2} \sup_{x\in X}\sqrt{k(x,x)}$ and assume that there exists an $\eta > 0$ and $\mathbf{E}_* \in \mathcal{H}_{\boldsymbol{\Xi}}$ such that $\sup_{\|\mathbf{E}\|_{\boldsymbol{\Xi}}\leq 1} \mathbb{E}_X\left[\mathbb{E}[\mathbf{E}^* k(x,\cdot)|x] - \langle \mathbf{E}^* k(x,\cdot), \mathbf{E}_*^* k(x,\cdot)\rangle_l\right]^2 = \eta < \infty$. Under assumptions E.1 we have that for every $\epsilon > 0$ there exists a constant D such that*

$$\limsup_{n\to\infty} \sup_{\mathbb{P}\in\mathfrak{P}} \mathbb{P}^n\left[\mathcal{E}_c[\mathbf{E}_n] > \left(\sqrt{\mathcal{E}_c[\mathbf{E}_*]} + \eta^{1/4}\sqrt{8C(\|\mathbf{E}_*\|_{\boldsymbol{\Xi}} + \|\mathbf{E}_n\|_{\boldsymbol{\Xi}})} + Dn^{-1/2}\right)^2\right] < \epsilon.$$

The theorem tells us that we essentially have a rate of n^{-1} up to an approximation error which we suffer if we can not represent the conditional expectation exactly with our RKHS $\mathcal{H}_{\boldsymbol{\Xi}}$.

Also note that the term $\mathcal{E}_c[\mathbf{E}_*]$ is closely related to η. So if we can represent the true conditional expectation then both η and $\mathcal{E}_c[\mathbf{E}_*]$ will be 0 and we have a $O(n^{-1})$ convergence to the true conditional expectation.

E.2.4. Convergence Rates for the Approximate Sum Rule

We can apply these rates now directly to the approximate sum rule with the help of Theorem 4.1. The theorem uses a mean estimate together with an estimate for the conditional expectation. We therefore need samples from $\mathbb{Q}_X$ and $\mathbb{P}_{X\times Y}$. We use the notation $\mathbb{Q}\otimes\mathbb{P}$ to denote the product measure over $X\times(X\times Y)$ and $(\mathbb{Q}\otimes\mathbb{P})^n$ denotes the product measure over n samples, whereas we assume that all the samples are iid.

Theorem E.5. *Let $C = \|\mathbf{A}^{1/2}\|_{op}\|\mathbf{B}\|_{op}^{1/2} \sup_{x\in X}\sqrt{k(x,x)}$ and assume that there exists an $\eta > 0$ and $\mathbf{E}_* \in \mathcal{H}_{\boldsymbol{\Xi}}$ such that $\sup_{\|\mathbf{E}\|_{\boldsymbol{\Xi}}\leq 1} \mathbb{E}_X\left[\mathbb{E}[\mathbf{E}^* k(x,\cdot)|x] - \langle \mathbf{E}^* k(x,\cdot), \mathbf{E}_*^* k(x,\cdot)\rangle_l\right]^2 = \eta < \infty$. Furthermore, assume the mean estimate m_X^n fulfills eq. 10 with an $\alpha \in]0,1]$. Under assumptions E.1 and if $\mathbb{Q}_X \ll \mathbb{P}_X$ with a Radon-Nikodým derivative that is a.e. upper bounded by b we have that for every $\epsilon > 0$ exist constants A and D such that*

$$\limsup_{n\to\infty} \sup_{\mathbb{P}\in\mathfrak{P}} (\mathbb{P}\otimes\mathbb{Q})^n\left[\mathcal{E}_m[m_Y^n] > b\left(\sqrt{\mathcal{E}_c[\mathbf{E}_*]} + \eta^{1/4}\sqrt{8C(\|\mathbf{E}_*\|_{\boldsymbol{\Xi}} + \|\mathbf{E}_n\|_{\boldsymbol{\Xi}})} + Dn^{-1/2}\right)^2 + A\|\mathbf{E}_n\|_{\boldsymbol{\Xi}}^2 n^{-\alpha}\right] < \epsilon.$$

We restate the theorem in a more readable form. For this we combine the approximation error terms:

$$\mathcal{E}_A[\mathbf{E}_*] := \max\{\sqrt{\mathcal{E}_c[\mathbf{E}_*]}, \eta^{1/4}\}$$

and we simplify the theorem to:

Theorem E.6. *Let $\mathbf{E}_*$ be a minimiser of the approximation error $\mathcal{E}_A$. Under assumptions E.1 and if $\mathbb{Q}_X \ll \mathbb{P}_X$ with a bounded Radon-Nikodým derivative we have that for every $\epsilon > 0$ exist constants a, b, c, d such that*

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

F. Measure & Integration Assumptions

In this paper we have essentially three different sorts of expectation operations: Expectations over functions $f \in \mathcal{H}_X$ on a space X, expectations over functions $g \in \mathcal{H}_{X\times Y}$ on product spaces $X \times Y$ and conditional expectations of functions $h \in \mathcal{H}_Y$ given a $x \in X$. We use Lebesgue integrals based on Fremlin (2000).

For the simple expectation $\mathbb{E}f$ we assume that all $f \in \mathcal{H}_X$ are integrable wrt. the corresponding probability measure $\mathbb{P}$ on X. This is not a very restrictive assumption and most kernels one will consider in practice will imply this assumption (see also Berlinet & Thomas-Agnan (2004)).

Expectations over product spaces are similar. Given two measure spaces $(X, \Sigma, \mathbb{P}_X)$ and $(Y, \mathrm{T}, \mathbb{P}_Y)$ we use the product measure $(X \times Y, \Lambda, \mathbb{P}_{X\times Y})$ from Fremlin (2001)[Def. 251A]. For this product measure we have that $\Sigma \,\widehat{\otimes}\, \mathrm{T} \subset \Lambda$ and for $E \in \Sigma, F \in \mathrm{T}$ we have that $\mathbb{P}_{X\times Y}(E \times F) = \mathbb{P}_X(E)\mathbb{P}_Y(F)$. In the cases where we have RKHS functions on the product space we assume that these functions are integrable wrt. $\mathbb{P}_{X\times Y}$.

The important theorem for product integrals is the Fubini theorem (Fremlin, 2000)[Thm. 252B] which guarantees us that for $\mathbb{P}_{X\times Y}$-integrable functions g the function $\mathbb{E}_Y g(x, y)$ is $\mathbb{P}_X$-integrable and $\mathbb{E}_{X\times Y} g = \mathbb{E}_X \mathbb{E}_Y g(x, y)$.

The final object of interest is the conditional expectation $\mathbb{E}[h|x]$. There are multiple ways to deal with conditioning. The easiest case is where we have densities $p(x, y)$ and $p(x)$ wrt. Lebesgue-measure for $\mathbb{P}_{X\times Y}$ and $\mathbb{P}_X$. We can then define a conditional expectation

$$\mathbb{E}[h|x] := \int h(y) \frac{p(x, y)}{p(x)} \mathbb{P}_Y(dy)$$

interpreting $0/0$ as 0. Densities are only defined up to a set of zero measure and, hence, also this conditional expectation is only unique up to a $\mathbb{P}_X$ zero measure set. If such densities exist and if these are integrable wrt. the relevant measure then the Fubini theorem guarantees us that $\int h(y)p(x, y)\mathbb{P}_Y(dy)$ is $\mathbb{P}_X$-integrable and also that $\int h(y)\frac{p(x,y)}{p(x)}\mathbb{P}_Y(dy)$ is $\mathbb{P}_X$-integrable.

For simplicity we assume that we have such a conditional expectation, however, the density assumption is not crucial and can be avoided by working with general conditional expectations as in Fremlin (2001)[chp. 233].