NONPARAMETRIC APPROXIMATION OF CONDITIONAL EXPECTATION OPERATORS

MATTES MOLLENHAUER AND PÉTER KOLTAI

Abstract. Given the joint distribution of two random variables X, Y on some second countable locally compact Hausdorff space, we investigate the statistical approximation of the L^2-operator P defined by $[Pf](x) := \mathbb{E}[f(Y) \mid X = x]$ under minimal assumptions. By modifying its domain, we prove that P can be arbitrarily well approximated in operator norm by Hilbert–Schmidt operators acting on a reproducing kernel Hilbert space. This fact allows to estimate P uniformly by finite-rank operators over a dense subspace even when P is not compact. In terms of modes of convergence, we thereby obtain the superiority of kernel-based techniques over classically used parametric projection approaches such as Galerkin methods. This also provides a novel perspective on which limiting object the nonparametric estimate of P converges to. As an application, we show that these results are particularly important for a large family of spectral analysis techniques for Markov transition operators. Our investigation also gives a new asymptotic perspective on the so-called kernel conditional mean embedding, which is the theoretical foundation of a wide variety of techniques in kernel-based nonparametric inference.

1. Introduction

We consider two random variables X, Y taking values in a measurable space $(E, \mathcal{F}_E)$ where E is a second countable locally compact Hausdorff space and $\mathcal{F}_E$ its Borel σ-field. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space with expectation operator $\mathbb{E}$. Let π denote the the pushforward of $\mathbb{P}$ under X, i.e., $X \sim \pi$ and let $L^2(E, \mathcal{F}_E, \pi; \mathbb{R}) = L^2(\pi)$ be the space of real-valued Lebesgue square integrable functions on $(E, \mathcal{F}_E)$ with respect to π. Analogously, define ν as the pushforward of $\mathbb{P}$ under Y on E, i.e, $Y \sim \nu$. Our goal is to perform a nonparametric estimation of the contractive *conditional expectation operator* $P\colon L^2(\nu) \to L^2(\pi)$ defined by

$$[Pf](x) := \mathbb{E}[f(Y) \mid X = x] = \int_E f(y)\, p(x, \mathrm{d}y),$$

where $p : E \times \mathcal{F}_E \to \mathbb{R}_+$ is the *Markov kernel* which describes a *regular version* of the distribution of Y conditioned on X in terms of

$$\mathbb{P}[Y \in \mathcal{A} \mid X = x] = \mathbb{E}[\mathbb{1}_{\mathcal{A}}(Y) \mid X = x] = \int_{\mathcal{A}} p(x, \mathrm{d}y) = p(x, \mathcal{A})$$

for all $x \in E$ and events $\mathcal{A} \in \mathcal{F}_E$. We will introduce additional notation and details as well as appropriate assumptions in Section 4.

2020 *Mathematics Subject Classification.* 46E22, 47A58, 46B28, 62J02, 62G05.

Key words and phrases. reproducing kernel Hilbert space, conditional mean embedding, conditional expectation operator.

We derive a natural and self-contained theory of the approximation of P over functions in a *reproducing kernel Hilbert space* (RKHS) which is densely embedded into the domain of P. Our analysis shows that the approximation of P is strongly connected to recently developed concepts in RKHS-based inference and statistical learning such as the *kernel mean embedding* (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Muandet et al., 2017), *maximum mean discrepancy* (Gretton et al., 2012a; Sejdinovic et al., 2013) and the *conditional mean embedding* (Song et al. 2009; Grünewälder et al. 2012; Klebanov et al. 2020a; Park and Muandet 2020a), which allows to extend our theory to several practical directions such as hypothesis testing, filtering and spectral analysis for Markov kernels.

We will focus on deriving approximation-theoretic results instead of a statistical analysis of convergence rates in our investigation. However, we show that convergence results can be carried over from the theory of nonparametric regularized least squares regression with vector-valued output variables (see for example Caponnetto and De Vito, 2007; Rastogi and Sampath, 2017; Rastogi et al., 2020; Park and Muandet, 2020b).

As a practical application, we argue that our theory provides a statistical model for a well-known family of numerical spectral analysis techniques for *Markov transition operators*, which we highlight in the following example.

A motivating example: Markov transition operators. The above scenario is of particular practical interest when $Y := X_{t+\tau}$ and $X := X_t$ for some stationary Markov process $(X_t)_{t\in\mathbb{R}}$ on the state space $(E, \mathcal{F}_E)$, as in this case $\pi = \nu$ and P given by

$$[Pf](x) = \mathbb{E}[f(X_{t+\tau}) \mid X_t = x] \tag{1.1}$$

is the Markov transition operator with respect to the time lag $\tau > 0$.

In the context of Markov processes and dynamical systems, it is known that the spectrum of P and the associated eigenfunctions determine crucial properties of the underlying dynamics such as ergodicity, speed of mixing, the decomposition of the state space into almost invariant (so-called *metastable*) components and many more (Davies, 1982a,b, 1983; Roberts et al., 1997; Roberts and Tweedie, 2001; Kontoyiannis and Meyn, 2003, 2005, 2017; Huisinga et al., 2004; Huisinga and Schmidt, 2006; Paulin, 2015).

As such, the operator P is often empirically approximated in various scientific disciplines by performing a projection onto finite-dimensional subspaces of $L^2(\pi)$ (see for example Li, 1976; Ding and Li, 1991; Dellnitz and Junge, 1999; Huisinga, 2001; Junge and Koltai, 2009; Schmid, 2010; Schütte and Sarich, 2013; Tu et al., 2014; Williams et al., 2015a; Klus et al., 2016, 2018; Korda and Mezić, 2018, and references therein). That is, given an n-dimensional subspace $V_n \subset L^2(\pi)$ spanned by a dictionary of basis elements, a Monte Carlo quadrature based on observational data is performed on V_n to obtain the empirical finite-rank operator $\widehat{P}_n$ as an estimate of the *Galerkin-approximation* $P_n := \Pi_n P \Pi_n$. Here, Π_n is the orthogonal projection operator onto V_n. Under the assumption of ergodicity, one typically obtains $\widehat{P}_n \to P_n$ almost surely by some version of Birkhoff's ergodic theorem (Klus et al., 2016). From a statistical perpective, these methods can be regarded as *parametric models*, the parameter choice being the fixed basis functions spanning the ansatz space V_n. By increasing the number of spanning elements, a convergence of the Galerkin approximation P_n to the real operator P in only the *strong operator topology* (i.e., pointwise on $L^2(\pi)$) as $n \to \infty$ can generally be obtained (Korda and Mezić, 2018).

In practice, the above methods are typically aimed at performing an empirical spectral analysis of P, i.e., spectral properties of $\widehat{P}_n$ are computed and used as an approximation of the spectral properties of P. It is well-known that most desirable spectral convergence results require a convergence in operator norm (Kato, 1980; Chatelin, 1983). The spectral convergence of the parametric approaches mentioned above is therefore ultimately limited by the pointwise convergence of numerical projection methods (Hackbusch, 1995).

As a *nonparametric* counterpart of the given parametric methods, there exist RKHS-based versions where the basis functions are adapted to the data (Williams et al., 2015b; Kawahara, 2016; Klus et al., 2020). For these methods, one may hope that they allow for stronger modes of convergence than the classical projection methods. However, it has not been shown yet which object is actually approximated in the infinite-sample limit, as the asymptotics are significantly more complicated in this case. Our theory solves this problem and confirms that the overall convergence is stronger than in the parametric case under mild assumptions. The strength of this result comes at the price of requiring to restrict the domain of the operator onto an RKHS. Whether relevant objects, such as eigenfunctions, of the original operator are contained in this space, is in general an open question.

Structure of this paper. This work is structured as follows. We delineate related theoretical work in the field of nonparametric statistical inference in Section 2. For better accessibility, we present our main results from a high-level viewpoint in Section 3. Section 4 contains the mathematical preliminaries and detailed assumptions. We prove our main results in Section 5 along some additional findings and elaborate on their implications from a theoretical perspective. In Section 6, we outline the empirical estimation in the context of inverse problems and regularization theory, which we investigate in detail for the Tikhonov–Phillips case in Section 7. We revisit our example of Markov transition operators in Section 8 and conclude with a brief outlook on potential future research in Section 9.

2. Related work

This work is inspired by recent development in RKHS-based statistical inference. Although our investigation is targeted at creating a more general mathematical perspective from an approximation viewpoint, we make use of the theoretical tools which were originally developed in this context. We therefore highlight the most important work which impacted our analysis.

Over the last years, the theory of RKHS-based inference and the kernel mean embedding (KME, see Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Muandet et al., 2017) spawned a vast variety of methods in various statistical disciplines. In this context, a nonparametric approximation of the conditional mean operation $(x, f) \mapsto \mathbb{E}[f(Y) \mid X = x]$ for functions f in some RKHS $\mathscr{H}$ over E was developed by Song et al. (2009) as a purely linear-algebraic concept under the name *conditional mean embedding* (CME). This idea has since been used as the theoretical backbone for methods in Bayesian analysis, graphical models, time series analysis, spectral analysis and dimensionality reduction, filtering, reinforcement learning and many more (see for example Muandet et al. 2017 for a non-exhaustive selection of applications).

Although the CME as described by Song et al. (2009) performs successfully in applications, the mathematical assumptions imposed in the original work are typically violated; this has been thoroughly examined by Klebanov et al. (2020a). The foundational problems in the theory of the CME led to an investigation of the approximation of RKHS-valued conditional Bochner expectations from a regression perspective. In particular, Grünewälder et al. (2012) show that the empirical Tikhonov–Phillips solution of a regularized least squares regression problem in a vector-valued reproducing kernel Hilbert space coincides with the empirical estimate derived by Song et al. (2009). Additionally, Grünewälder et al. (2013) propose to use the same estimate for the approximation of linear operators in a very broad sense but do not offer an asymptotic perspective of this idea.

Park and Muandet (2020a) extend the asymptotic regression theory of the CME in the framework of least-squares regression in a *vector-valued reproducing kernel Hilbert space* (vRKHS) and regularization theory (see for example Caponnetto and De Vito 2007). In this context, uniform convergence rates are proven under the assumption that the true CME is contained in the hypothesis space. Klebanov et al. (2020b) extend the operator-theoretic interpretation of the CME. In particular, they prove existence of an operator on an RKHS which expresses the conditional mean under the assumption that the true conditional mean function is a member of a corresponding tensor product space. In fact, our analysis shows that this assumption is equivalent to the assumption under which Park and Muandet (2020a) derive convergence rates.

Comparison to this work. Concluding the overall picture of the aforementioned work: while the regression perspective of the CME (Grünewälder et al., 2012; Park and Muandet, 2020a) allows to consider asymptotic interpretations and prove convergence results, it has the fundamental drawback that the algebraically interesting operator-theoretic perspective of P is not present. Even more so, the estimation of spectral properties of P (for example in the case of Markov operators or for dimensionality reduction purposes) is impossible. Conversely, the operator-theoretic formulation of the CME (Song et al., 2009; Klebanov et al., 2020a,b) lacks an asymptotic perspective and suffers from complex interdependencies of various assumptions (Klebanov et al., 2020a), severely impeding a theoretical mathematical analysis. Additionally, the approximation viewpoint in the L^2-operator context has not been investigated yet. We will see that this approximation admits a natural perspective in terms of the maximum mean discrepancy between the underlying Markov kernels.

Regarded in the context of the CME, our results can be interpreted as the missing link between the recent work of Klebanov et al. (2020b) and Park and Muandet (2020a). In particular, we provide an asymptotic approximation perspective in the operator-theoretic context of conditional expectations. On our way, we moreover improve a surrogate risk bound used by Grünewälder et al. (2012) and Park and Muandet (2020a) which serves as the theoretical foundation for the regression perspective of the CME. However, our results are formulated in a more general perspective in terms of the numerical approximation of linear operators and can certainly be regarded outside of the context of the previously mentioned work on the CME.

3. Main results

We will briefly outline some main results and the general content of our work. All discussed concepts, mathematical preliminaries, and assumptions used in this section will be introduced in more detail in Section 4.

As previously mentioned, we aim to approximate P over a separable reproducing kernel Hilbert space $\mathscr{H}$ consisting of functions from E to $\mathbb{R}$ generated by the canonical feature map $\varphi : E \to \mathscr{H}$. We will choose the space $\mathscr{H}$ such that is is a subset of $C_0(E)$, i.e., the space of continuous real-valued functions which vanish at infinity (Carmeli et al., 2010). Additionally, we choose $\mathscr{H}$ such that it can be continuously embedded into $L^2(\pi)$ as well as $L^2(\nu)$. That is, the inclusion operator $i_\pi : \mathscr{H} \to L^2(\pi)$ defined by $f \mapsto [f]_{\sim L^2(\pi)}$ and the analogously defined inclusion $i_\nu : \mathscr{H} \to L^2(\nu)$ are bounded (Steinwart and Christmann, 2008). Moreover, we will generally assume that $\mathscr{H}$ is dense in both $L^2(\pi)$ and $L^2(\nu)$. This property is called L^2-*universality* (Carmeli et al., 2010; Sriperumbudur et al., 2011). The results shown here are proven in Section 5 with a more detailed presentation of the assumptions.

Remark 3.1 (Inclusion operators and notation). We will sometimes suppress the inclusion operators i_π and i_ν in our notation when the context is clear. In particular, for $f \in \mathscr{H}$ we will simply write $\|f\|_{L^2(\nu)}$ instead of $\|i_\nu f\|_{L^2(\nu)}$. Furthermore, under the above assumptions, we may understand the operator $Pi_\nu : \mathscr{H} \to L^2(\pi)$ as a conditional expectation operator acting on functions of $\mathscr{H}$ via

$$[Pi_\nu f](x) = \mathbb{E}[f(Y) \mid X = x] \in L^2(\pi) \quad \text{for } f \in \mathscr{H} \tag{3.1}$$

and use the norm of $\mathscr{H}$ on its domain. By abuse of notation, we may write $P : \mathscr{H} \to L^2(\pi)$ instead of Pi_ν for the operator in (3.1). We will emphasize which version of P we refer to by simply distinguishing between $P : \mathscr{H} \to L^2(\pi)$ and $P : L^2(\nu) \to L^2(\pi)$. We write out the corresponding operator norms $\|P\|_{\mathscr{H} \to L^2(\pi)}$ and $\|P\|_{L^2(\nu) \to L^2(\pi)}$ to prevent confusion. Note that by boundedness of i_ν, we have $\|P\|_{\mathscr{H} \to L^2(\pi)} \leq \|i_\nu\| \, \|P\|_{L^2(\nu) \to L^2(\pi)}$. Similarly, for every bounded operator $A : \mathscr{H} \to \mathscr{H}$ we can consider the bounded operator $i_\pi A$ from $\mathscr{H}$ to $L^2(\pi)$, which we will also abbreviate as $A : \mathscr{H} \to L^2(\pi)$. At this point, it is worth mentioning that functions in $\mathscr{H}$ are generally defined pointwise, while elements of $L^2(\pi)$ are equivalence classes of π-a.e. equivalent functions.

It is known that under the assumptions above, the inclusions i_π and i_ν are Hilbert–Schmidt operators (Steinwart and Christmann, 2008, Chapter 4.3). Therefore, the operator $P : \mathscr{H} \to L^2(\pi)$ is Hilbert–Schmidt (and hence compact), independently of the fact whether $P : L^2(\nu) \to L^2(\pi)$ is Hilbert–Schmidt or not. Intuitively, the approximation of P over functions in $\mathscr{H}$ in operator norm is therefore generally possible with finite-rank operators from $\mathscr{H}$ to $L^2(\pi)$. Since we can not efficiently impose the class of Hilbert–Schmidt operators from $\mathscr{H}$ to $L^2(\pi)$ as a nonparametric hypothesis space in practical applications, we now provide a more suitable approximation theory for practical scenarios. The following result shows that we may actually restrict ourselves to the class of Hilbert–Schmidt operators mapping from the space $\mathscr{H}$ to itself and still expect an approximation of $P : \mathscr{H} \to L^2(\pi)$ up to an arbitrary degree of accuracy.

Theorem 3.2 (Approximation by Hilbert–Schmidt operators)**.** *If there exists a reproducing kernel Hilbert space $\mathscr{H} \subset C_0(E)$ which is densely and continuously embedded into both $L^2(\pi)$ and $L^2(\nu)$, then for every $\delta > 0$, there exists a Hilbert–Schmidt operator $A\colon \mathscr{H} \to \mathscr{H}$, such that*

$$\|A - P\|_{\mathscr{H} \to L^2(\pi)} < \delta. \tag{3.2}$$

Remark 3.3. Some remarks related to Theorem 3.2 are in order.

(1) We do not require $P : L^2(\nu) \to L^2(\pi)$ to be a Hilbert–Schmidt operator or compact in order for the above statement to hold. Our result is not a contradiction to the known fact that operator norm limits of Hilbert–Schmidt operators are compact. The reason for that is that the compactness property is given with respect to the norm $\|\cdot\|_{\mathscr{H}}$ on the domain, which is stronger than the norm $\|\cdot\|_{L^2(\nu)}$. Hence, the continuous extension to $A : L^2(\nu) \to \mathscr{H}$ via the known construction for bounded operators (Weidmann, 1980, Theorem 4.5) is generally not compact. This can equivalently be seen by the fact that i_ν does generally not admit a globally defined bounded inverse. We visualize Theorem 3.2 in Figure 1.

(2) The assumptions on $\mathscr{H}$ are not restrictive, as they are well examined in statistical learning theory and often satisfied for particular RKHSs used in practice. It is actually sufficient to only require that $\mathscr{H}$ is dense in $L^2(\rho)$ for any probability measure ρ on $(E, \mathcal{F}_E)$, as this implies denseness in both $L^2(\pi)$ and $L^2(\nu)$. We address these topics in detail in Section 4.

(3) We will later also see under which requirements there exists a Hilbert–Schmidt operator $A : \mathscr{H} \to \mathscr{H}$ such that $\|A - P\|_{\mathscr{H} \to L^2(\pi)} = 0$.

$$\begin{tikzcd}
L^2(\nu) \arrow[r, "P"] & L^2(\pi) \\
\mathscr{H} \arrow[u, "i_\nu"] \arrow[ur, shift left, "P i_\nu"] \arrow[ur, shift right, "i_\pi A"'] \arrow[r, "A \in \mathrm{S}_2(\mathscr{H})"'] & \mathscr{H} \arrow[u, "i_\pi"']
\end{tikzcd}$$

FIGURE 1. Nonparametric approximation of P over functions in $\mathscr{H}$ by a Hilbert–Schmidt operator $A \in \mathrm{S}_2(\mathscr{H})$. Theorem 3.2 shows that $P i_\nu \approx i_\pi A$ to arbitrary accuracy in the associated operator norm. The operator A is approximated by finite-rank operators on $\mathscr{H}$ in Corollary 3.4.

Corollary 3.4. *Under the assumptions of Theorem 3.2, there exists a sequence of finite-rank operators $(A_n)_{n\in\mathbb{N}}$ from $\mathscr{H}$ to $\mathscr{H}$ such that $\|A_n - P\|_{\mathscr{H} \to L^2(\pi)} \to 0$ as $n \to \infty$.*

As we will prove, such a sequence $(A_n)_{n\in\mathbb{N}}$ can be almost surely computed in practice by performing a nonparametric regression based on a linear space consisting of functions mapping from E to $\mathscr{H}$ given by

$$\mathscr{G} = \{A\varphi(\cdot) : E \to \mathscr{H} \mid A : \mathscr{H} \to \mathscr{H} \text{ is Hilbert–Schmidt}\}.$$

One can show that the space $\mathscr{G}$ is actually a vector-valued reproducing kernel Hilbert space (Carmeli et al. 2006, 2010) consisting of $\mathscr{H}$-valued Bochner square integrable functions. This fact connects our theory directly to the aforementioned work on conditional mean embeddings.
We show that the approximation of P in the norm $\|\cdot\|_{\mathscr{H}\to L^2(\pi)}$ admits a natural measure-theoretic interpretation in terms of the well-known *maximum mean discprepancy* (Gretton et al., 2012b; Sejdinovic et al., 2013), paving the way for nonparametric hypothesis tests based on P.

The next result is the theoretical foundation for Theorem 3.2 and will allow us to construct an estimator of $P\colon \mathscr{H} \to L^2(\pi)$ (see Section 6). It shows how the approximation of P is related to the approximation of a conditional Bochner expectation and improves surrogate risk bounds used by Grünewälder et al. (2012) and Park and Muandet (2020a) in the context of the CME (see Remark 5.9 for details).

Theorem 3.5 (Regression and conditional mean approximation)**.** *Under the assumptions of Theorem 3.2, we have for every Hilbert–Schmidt operator $A\colon \mathscr{H} \to \mathscr{H}$ that*

$$\|A - P\|^2_{\mathscr{H}\to L^2(\pi)} \le \mathbb{E}\left[\|F_p(X) - A^*\varphi(X)\|^2_{\mathscr{H}}\right] = \|F_p - A^*\varphi(\cdot)\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})},$$

where $F_p = \mathbb{E}[\varphi(Y) \mid X = \cdot] \in L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ is any regular version of the $\mathscr{H}$-valued conditional Bochner expectation $\mathbb{E}[\varphi(Y) \mid X] \in L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathscr{H})$. The given bound is sharp.

Remark 3.6. In fact, the above result actually holds under less strict assumptions, which we will see in Section 5.

As is well-known in statistical learning theory (see for example Cucker and Zhou 2007), the right hand side of the bound in Theorem 3.5 is exactly the so-called *excess risk* $R(F) - R(F_p)$ of the infinite-dimensional least squares regression problem of finding $\arg\min_{F\in\mathscr{G}} R(F)$, where

$$R(F) := \mathbb{E}\left[\|\varphi(Y) - F(X)\|^2_{\mathscr{H}}\right] \text{ for } F(\cdot) = A^*\varphi(\cdot).$$

In particular, the risk $R(F)$ allows for the decomposition

$$R(F) = \|F_p - F\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} + R(F_p)$$

with the irreducible error term $R(F_p)$. This puts the approximation of P in perfect line with the formalism developed for regularized least squares regression with reproducing kernels which was established in a series of highly influential papers (De Vito et al., 2005; Caponnetto and De Vito, 2007; Bauer et al., 2007; Yao et al., 2007) and its connection to inverse problems in Hilbert spaces.

In particular, by employing a generic *regularization strategy* g_λ for a regularization parameter $\lambda > 0$, such as for example Tikhonov–Phillips regularization, spectral cutoff or Landweber iteration (see Engl et al. 1996), we obtain a regularized solution to the above regression problem via

$$F_\lambda := g_\lambda(T)\mathcal{I}_\pi^* F_p \in \mathscr{G}, \tag{3.3}$$

where $T : \mathscr{G} \to \mathscr{G}$ is the *generalized kernel covariance operator* (see Section 4.6.2) of the space $\mathscr{G}$ associated with X and $\mathcal{I}_\pi : \mathscr{G} \to L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ is the *inclusion operator* of $\mathscr{G}$ into the space of Bochner square integrable functions $L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$.

Since T plays a crucial role in the underlying inverse problem, we also show that the action of T on $\mathscr{G}$ admits a dual interpretation in terms of composition operators acting on the class of Hilbert–Schmidt operators on $\mathscr{H}$. For the special case that g_λ describes Tikhonov–Phillips regularization, this theory lets us obtain a closed form expression of the regularized solution in terms of the *kernel covariance operators* C_{XX} and C_{XY} on $\mathscr{H}$. We confirm this solution to be the adjoint of the CME first derived by Song et al. (2009) given by $A_\lambda = (C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1} C_{XY}$ without the limiting assumptions imposed in the original work. Although this statement does not come as a surprise, it has never been proven in any of the aforementioned papers on the CME. Our results can be interpreted as the population analogue of a similar statement for the empirical case derived by Grünewälder et al. (2012) (see Section 7.2).

By performing the empirical discretization of the above operators and problem (3.3) based on a finite set of observations $\mathbf{z} = ((X_1, Y_1), \ldots, (X_n, Y_n))$ sampled iid from $\mathcal{L}(X, Y)$ in terms of the *sampling operator approach* (Smale and Zhou, 2005, 2007), we obtain a regularized empirical solution $F_{\lambda,\mathbf{z}}(\cdot) = A_{\lambda,\mathbf{z}}^* \varphi(\cdot)$. Theorem 3.5 shows that the convergence $F_{\lambda,\mathbf{z}} \to F_p$ in $L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ for $n \to \infty$ with a suitable regularization scheme $\lambda = \lambda(n)$ implies convergence of $A_{\lambda,\mathbf{z}} \to P$ in the norm $\|\cdot\|_{\mathscr{H} \to L^2(\pi)}$.

4. Preliminaries and Assumptions

We give a concise overview of the needed mathematical background.

4.1. **Measure, integration and Hilbert space operators.** We briefly introduce the main concepts from measure theory and linear operators and analysis in Hilbert spaces. We refer the reader to Diestel and Uhl (1977), Dunford and Schwartz (1988a,b) and Dudley (2002) for details.

For any topological space E, we will write $\mathcal{F}_E = \mathcal{B}(E)$ for its associated Borel field. For any collection of sets $\mathcal{M}$, $\sigma(\mathcal{M})$ denotes the intersection of all σ-fields containing $\mathcal{M}$. For any σ-field $\mathcal{F}$ and countable index set I, we write $\mathcal{F}^{\otimes I}$ as the product σ-field (i.e., the smallest σ-field with respect to which all coordinate projections on E^I are measurable). Note that since E is Polish (i.e., separable and completely metrizable), we have $\mathcal{B}(E^I) = \mathcal{B}(E)^{\otimes I}$, i.e. the Borel field on the product space generated by the product topology and the product of the individual Borel fields are equal. Put differently, the Borel field operator and the product field operator are compatible with respect to product spaces (Dudley, 2002, Proposition 4.1.17). Moreover, E^I equipped with the product topology is Polish.

In what follows, we write B for a separable real Banach space with norm $\|\cdot\|_B$, and H for a separable real Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$. The expression $\mathfrak{B}(B, B')$ stands for the Banach algebra of bounded linear operators from B to another Banach space B' and is equipped with the operator norm $\|\cdot\|$. For the case $B = B'$, we abbreviate $\mathfrak{B}(B, B') = \mathfrak{B}(B)$. We will also write $\|\cdot\| = \|\cdot\|_{B \to B'}$, if the choice of norms on the underlying spaces B, B' needs to be emphasized.

Let $(\Omega, \mathcal{F}, \pi)$ be a measure space. For any separable Banach space B, we let $L^p(\Omega, \mathcal{F}, \pi; B)$ denote the space of strongly $\mathcal{F} - \mathcal{F}_B$ measurable and Bochner p-integrable functions $f \colon \Omega \to B$ for $1 \le p \le \infty$. In the case of $B = \mathbb{R}$, we simply write $L^p(\pi) := L^p(\Omega, \mathcal{F}, \pi; \mathbb{R})$ for the standard space of real-valued Lebesgue p-integrable functions.

The expression $H' \otimes H$ denotes the tensor product of Hilbert spaces H, H'. The Hilbert space $H' \otimes H$ is the completion of the algebraic tensor product with respect to the inner product $\langle x_1' \otimes x_1,\, x_2' \otimes x_2 \rangle_{H' \otimes H} = \langle x_1',\, x_2' \rangle_{H'} \langle x_1,\, x_2 \rangle_H$ for $x_1, x_2 \in H$ and $x_1', x_2' \in H'$. We interpret the element $x' \otimes x \in H' \otimes H$ as the linear rank-one operator $x' \otimes x \colon H \to H'$ defined by $\tilde{x} \mapsto \langle \tilde{x},\, x \rangle_H x'$ for all $\tilde{x} \in H$. Whenever $(e_i)_{i \in I}$, $(e_j')_{j \in J}$ are complete orthonormal systems (CONSs) in H and H', $(e_j' \otimes e_i)_{i \in I, j \in J}$ is a CONS in $H' \otimes H$. Thus, when H and H' are separable, $H' \otimes H$ is separable.

For $1 \le p < \infty$, the *p-Schatten class* $S_p(H, H')$ consists of all compact operators A from H to H' such that the norm $\|A\|_{S_p(H)} := \|(\sigma_i(A))_{i \in J}\|_{\ell_p}$ is finite. Here $\|(\sigma_i(A))_{i \in J}\|_{\ell_p}$ denotes the ℓ_p sequence space norm of the sequence of the strictly positive singular values of A indexed by the countable set J, which we assume to be ordered nonincreasingly. We set $S_\infty(H, H')$ to be the class of compact operators from H to H' equipped with the operator norm and write $S_p(H) := S_p(H, H)$ for all $1 \le p \le \infty$. The spaces $S_p(H)$ are two-sided ideals in $\mathfrak{B}(H)$. Moreover $\|A\|_{S_q(H,H')} \le \|A\|_{S_p(H,H')}$ holds for $1 \le p \le q \le \infty$, i.e., $S_p(H, H') \subseteq S_q(H, H')$. For $p = 2$, we obtain the Hilbert space of *Hilbert–Schmidt operators* from H to H' equipped with the inner product $\langle A_1,\, A_2 \rangle_{\mathrm{S}_2(H,H')} = \operatorname{Tr}(A_1^* A_2)$. For $p = 1$, we obtain the Banach space of *trace class operators*. The Schatten classes are the completion of *finite-rank operators* (i.e., operators in $\operatorname{span}\{x' \otimes x \mid x \in H, x' \in H'\}$) with respect to the corresponding norm.

We will make frequent use of the fact that the tensor product space $H' \otimes H$ can be isometrically identified with the space of Hilbert–Schmidt operators from H to H', i.e., we have $\mathrm{S}_2(H, H') \simeq H' \otimes H$. For elements $x_1, x_2 \in H$, $x_1', x_2' \in H'$, we have the relation $\langle x_1' \otimes x_1,\, x_2' \otimes x_2 \rangle_{H' \otimes H} = \langle x_1' \otimes x_1,\, x_2' \otimes x_2 \rangle_{\mathrm{S}_2(H,H')}$, where the tensors are interpreted as rank-one operators as described above. This identification of tensors with as rank-one operators extends to $\operatorname{span}\{x' \otimes x \mid x \in H, x' \in H'\}$ by linearity and defines a linear isometric isomorphism between $H' \otimes H$ and $\mathrm{S}_2(H, H')$, which can also be seen by considering Hilbert–Schmidt operators in terms of their singular value decompositions. We will frequently switch in between these two viewpoints when considering Hilbert–Schmidt operators.

4.2. **Joint and regular conditional distributions.** In this paper, we will consider a second countable locally compact Hausdorff space $(E, \mathcal{F}_E)$ equipped with its Borel field. We need this technical setup to avoid dealing with measure-theoretic details later on.

We consider two random variables X, Y defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in E. We will assume without loss of generality that $(\Omega, \mathcal{F}, \mathbb{P})$ is rich enough to support all performed operations in this paper. For a finite number of random variables $X_1, \dots, X_n$ defined with values in E, we write $\mathcal{L}(X_1, \dots, X_n)$ for the *finite-dimensional law*, i.e., *pushforward measure* on $(E^n, \mathcal{B}(E^n))$. We write $X \overset{d}{=} Y$, if X and Y are equal in distribution, i.e., their laws are equal. Throughout this paper, we define $\pi := \mathcal{L}(X)$ and $\nu := \mathcal{L}(Y)$, i.e., we have $X \sim \pi$ and $Y \sim \nu$.

Let $p : E \times \mathcal{F}_E \to \mathbb{R}$ be a *Markov kernel*[1], i.e., $p(x, \cdot)$ is a probability measure on $(E, \mathcal{F}_E)$ for every $x \in E$ and the map $E \ni x \mapsto p(x, \mathcal{A})$ is an $\mathcal{F}_E - \mathbb{R}$ measurable function for every $\mathcal{A} \in \mathcal{F}_E$ such that

$$\mathbb{P}[Y \in \mathcal{A} \mid X = x] = \int_{\mathcal{A}} p(x, \mathrm{d}y) = p(x, \mathcal{A})$$

for all $x \in E$ and events $\mathcal{A} \in \mathcal{F}_E$. The Markov kernel p defines a so-called *regular version* of the above conditional distribution which allows to consider the fiberwise disintegration

$$\mathbb{P}[X \in \mathcal{A}, Y \in \mathcal{B}] = \int_{\mathcal{A}} p(x, \mathcal{B})\, \mathrm{d}\pi(x),$$

see Dudley (2002, Theorem 10.2.1). Such a Markov kernel p exists always in our scenario, since the space E is Polish (Dudley, 2002, Theorem 10.2.2). Additionally, two regular versions of the same conditional distribution with corresponding Markov kernels p, p' coincide almost everywhere, i.e., we have $p(x, \cdot) = p'(x, \cdot)$ for π-a.e. $x \in E$.

Our goal is to perform a nonparametric estimation of the *conditional expectation operator* $P\colon L^2(\nu) \to L^2(\pi)$ defined by

$$[Pf](x) := \mathbb{E}[f(Y) \mid X = x] = \int_E f(y)\, p(x, \mathrm{d}y),$$

which is a contractive linear map (and therefore bounded). In fact, this can easily be seen by making use of Jensen's inequality for conditional expectations and considering

$$\|Pf\|^2_{L^2(\pi)} = \mathbb{E}\left[\mathbb{E}[f(Y) \mid X]^2\right] \leq \mathbb{E}\left[\mathbb{E}[(f(Y))^2 \mid X]\right] = \mathbb{E}[f(Y)^2] = \|f\|^2_{L^2(\nu)}.$$

4.3. **Vector-valued reproducing kernel Hilbert spaces.** We will give a brief overview of the concept of a *vector-valued reproducing kernel Hilbert space* (vRKHS), i.e., a Hilbert space consisting of functions from a nonempty set E to a Hilbert space H. Since the construction of such a space is quite technical, we will not cover all mathematical details here but rather introduce the most important properties. For a rigorous treatment of this topic, we refer the reader to Carmeli et al. (2006) as well as Carmeli et al. (2010).

Definition 4.1 (Operator-valued psd kernel)**.** *Let E be a nonempty set and H be a real Hilbert space. A function $K : E \times E \to \mathfrak{B}(H)$ is called an* operator-valued

[1] We distinguish different notions of *kernels* in this paper. We will often refer to reproducing kernels/symmetric positive semidefinite kernels simply as *kernel*, while the kernel p defining a conditional distribution will always be called *Markov kernel*.

positive-semidefinite (psd) kernel, *if* $K(x,x') = K(x',x)^*$ *and all* $x, x' \in E$ *and additionally for all* $n \in \mathbb{N}$, $x_1, \dots, x_n \in E$ *and* $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, *we have*

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j \left\langle h,\, K(x_i, x_j)h \right\rangle_H \geq 0 \tag{4.1}$$

for all $h \in H$.

Let $K : E \times E \to \mathfrak{B}(H)$ be an operator-valued psd kernel. For a fixed $x \in E$ and $h \in H$, we obtain a function from E to H via

$$[K_x h](\cdot) := K(\cdot, x)h.$$

We can now consider the set

$$\mathscr{G}_{\text{pre}} := \operatorname{span}\{K_x h \mid x \in E, h \in H\} \tag{4.2}$$

and define an inner product on $\mathscr{G}_{\text{pre}}$ by linearly extending the expression

$$\langle K_x h,\, K_{x'} h' \rangle_{\mathscr{G}} := \langle h,\, K(x,x')h' \rangle_H \,. \tag{4.3}$$

Let $\mathscr{G}$ be the completion of $\mathscr{G}_{\text{pre}}$ with respect to this inner product. We call $\mathscr{G}$ the *H-valued reproducing kernel Hilbert space* or more generally the vRKHS induced by the kernel K.

The space $\mathscr{G}$ is a Hilbert space consisting of functions from E to H with the *reproducing property*

$$\langle F(x),\, h \rangle_H = \langle F,\, K_x h \rangle_{\mathscr{G}} \tag{4.4}$$

for all $F \in \mathscr{G}$, $h \in H$ and $x \in E$. Additionally, we have

$$\|F(x)\|_H \leq \|K(x,x)\|^{1/2}\, \|F\|_{\mathscr{G}}\,, \quad x \in E \tag{4.5}$$

for all $F \in \mathscr{G}$. When K_x is understood as a linear operator from H to $\mathscr{G}$ fixed $x \in E$, the inner product given by (4.3) implies that K_x is a bounded operator for all $x \in E$. As a result, we can rewrite the reproducing property (4.4) as

$$F(x) = K_x^* F \tag{4.6}$$

for all $F \in \mathscr{G}$ and $x \in E$. Therefore we obviously have

$$K_x^* K_{x'} = K(x, x'), \quad x, x' \in E \tag{4.7}$$

and the linear operators $K_x \colon \mathscr{H} \to \mathscr{G}$ and $K_x^* \colon \mathscr{G} \to \mathscr{H}$ are bounded with

$$\|K_x\| = \|K_x^*\| = \|K(x,x)\|^{1/2}\,. \tag{4.8}$$

In this paper, we will deal with two very specific examples of psd kernels, which we will introduce in what follows.

4.3.1. $\mathbb{R}$-*valued RKHS.* When we identify the space of linear operators on $\mathbb{R}$ with $\mathbb{R}$ itself and consider a scalar-valued psd kernel

$$k\colon E \times E \to \mathbb{R} \tag{4.9}$$

in the sense of Definition 4.1, we obtain the standard setting of the ($\mathbb{R}$-valued) reproducing kernel Hilbert space (RKHS; Aronszajn 1950). The kernel k satisfies $k(x,x') = k(x',x)$ for all $x, x' \in E$. We obtain a space $\mathscr{H}$ consisting of functions from E to $\mathbb{R}$ with the properties

(i) $\langle f,\, k(x,\cdot) \rangle_{\mathscr{H}} = f(x)$ for all $f \in \mathscr{H}$ (reproducing property), and

(ii) $\mathscr{H} = \overline{\operatorname{span}\{k(x,\cdot) \mid x \in E\}}$, where the completion is with respect to the RKHS norm.

It follows in particular that $k(x,x') = \langle k(x,\cdot),\, k(x',\cdot)\rangle_{\mathscr{H}}$. The so-called *canonical feature map* $\varphi\colon E \to \mathscr{H}$ is given by $\varphi(x) := k(x,\cdot)$.

The space $\mathscr{H}$ has been thoroughly examined over the last decades and has numerous applications in statistics, approximation theory and machine learning. For details, the reader may consult Berlinet and Thomas-Agnan (2004), Steinwart and Christmann (2008) and Saitoh and Sawano (2016).

Remark 4.2 (Notation)*.* In what follows, $\mathscr{H}$ will always denote the $\mathbb{R}$-valued RKHS induced by the kernel $k\colon E \times E \to \mathbb{R}$ with corresponding canonical feature map $\varphi\colon E \to \mathscr{H}$ as described in this section. We will write small letters $f, g, h \in \mathscr{H}$ for $\mathbb{R}$-valued RKHS functions.

4.3.2. *$\mathscr{H}$-valued vRKHS.* Let $\mathscr{H}$ be the $\mathbb{R}$-valued RKHS induced by the kernel $k : E \times E \to \mathbb{R}$ as decribed in Section 4.3.1. Let $\mathrm{Id}_{\mathscr{H}}$ be the identity operator on $\mathscr{H}$. We define the map $K\colon E \times E \to \mathfrak{B}(\mathscr{H})$ with

$$K(x,x') := k(x,x')\mathrm{Id}_{\mathscr{H}} \tag{4.10}$$

for all $x, x' \in E$. It is straightforward to show that K is a psd kernel and therefore induces an $\mathscr{H}$-valued vRKHS $\mathscr{G}$ (see also Carmeli et al., 2010, Example 3.3.(i)).

Remark 4.3 (Notation)*.* In what follows, $\mathscr{G}$ will always denote the $\mathscr{H}$-valued vRKHS induced by the kernel $K\colon E \times E \to \mathfrak{B}(\mathscr{H})$ given by $K(x,x') = k(x,x')\mathrm{Id}_{\mathscr{H}}$ as described in this section. We will write capital letters $F, G, H \in \mathscr{G}$ for $\mathscr{H}$-valued functions in order to distinguish them from real-valued functions $f, g, h \in \mathscr{H}$.

4.4. **Isomorphism between $\mathscr{G}$ and $\mathrm{S}_2(\mathscr{H})$.** The foundation of our approach is given by the fact that elements of the vRKHS $\mathscr{G}$ defined by the kernel $K(x,x') = k(x,x')\mathrm{Id}_{\mathscr{H}}$ can be interpreted as Hilbert–Schmidt operators on $\mathscr{H}$. We again recall that the space of Hilbert–Schmidt operators $\mathrm{S}_2(\mathscr{H})$ is isometrically isomorphic to the tensor product space $\mathscr{H} \otimes \mathscr{H}$ via an identification of rank-one operators as elementary tensors. We will use the latter to state the result, since a formulation in this way is more natural.

Theorem 4.4 ($\mathscr{G}$ is isomorphic to $\mathscr{H} \otimes \mathscr{H}$)**.** *Let $\mathscr{H}$ be a scalar RKHS with corresponding kernel k. Let $\mathscr{G}$ be the vector-valued RKHS induced by the kernel $K(x,x') := k(x,x')\mathrm{Id}_{\mathscr{H}}$. The map Θ defined on rank-one tensors in $\mathscr{H} \otimes \mathscr{H}$ defining an $\mathscr{H}$-valued function on E by the relation*

$$\left[\Theta(f \otimes h)\right](x) := h(x)f = (f \otimes h)\varphi(x) = \langle h,\, \varphi(x)\rangle_{\mathscr{H}}\, f \tag{4.11}$$

for all $x \in E$ and $f, h \in \mathscr{H}$ maps to $\mathscr{G}$. Furthermore, extending Θ to $\mathscr{H} \otimes \mathscr{H}$ via linearity and completion yields an isometric isomorphism between $\mathscr{H} \otimes \mathscr{H}$ and $\mathscr{G}$.

A proof of Theorem 4.4 can be found in Carmeli et al. (2010, Proposition 3.5 & Example 3.3(i)). The isometric isomorphism

$$\Theta : \mathscr{H} \otimes \mathscr{H} \to \mathscr{G}$$

defined by (4.11) seems technical but actually becomes quite intuitive when one examines how the inner products of both spaces are connected via the kernels k and K. We outline this connection briefly below.

Let $x, x' \in E$ and $h, h' \in \mathscr{H}$. We define $F := K_x h \in \mathscr{G}$ and $F' := K_{x'} h' \in \mathscr{G}$ and note that we can express the inner product in $\mathscr{G}$ as

$$\begin{aligned}\langle F,\, F' \rangle_{\mathscr{G}} &= \langle K_{x'}^* K_x h,\, h' \rangle_{\mathscr{H}} = \langle k(x', x) \mathrm{Id}_{\mathscr{H}} h,\, h \rangle_{\mathscr{H}} \\ &= \langle \varphi(x'),\, \varphi(x) \rangle_{\mathscr{H}} \, \langle h,\, h' \rangle_{\mathscr{H}} \\ &= \langle h \otimes \varphi(x),\, h' \otimes \varphi(x') \rangle_{\mathscr{H} \otimes \mathscr{H}} .\end{aligned}$$

This derivation can be extended straightforwardly to a correspondence of vector-valued functions $F, F' \in \operatorname{span}\{K_x h \,|\, x \in E, h \in \mathscr{H}\} \subseteq \mathscr{G}$ and linear combinations of tensors in $\{h \otimes \varphi(x) \,|\, x \in E, h \in \mathscr{H}\} \subseteq \mathscr{H} \otimes \mathscr{H}$ by using bilinearity of the respective inner products. Since both spans are dense in the associated spaces, this property can be extended to the full spaces via completion. We now restate Theorem 4.4 in a more accessible way for our scenario. The formulation below shows that pointwise evaluation of functions in $\mathscr{G}$ may be conducted by the action of the corresponding operator in $\mathrm{S}_2(\mathscr{H})$ on the canonical feature map φ. We will refer to this property as the *operator reproducing property*. We visualize the relations between $\mathscr{H} \otimes \mathscr{H}$, $\mathrm{S}_2(\mathscr{H})$ and $\mathscr{G}$ in Figure 2.

Corollary 4.5 (Operator reproducing property)**.** *For every function $F \in \mathscr{G}$ there exists an operator $A := \Theta^{-1}(F) \in \mathrm{S}_2(\mathscr{H})$ such that*

$$F(x) = A\varphi(x) \in \mathscr{H} \tag{4.12}$$

for all $x \in E$ with $\|A\|_{\mathrm{S}_2(\mathscr{H})} = \|F\|_{\mathscr{G}}$ and vice versa.

Conversely, for any pair $F \in \mathscr{G}$ and $A \in \mathrm{S}_2(\mathscr{H})$ satisfying property (4.12) we have $A = \Theta^{-1}(F)$.

Proof. The first assertion directly follows from Theorem 4.4 and the construction of Θ. It remains to prove the second assertion. Let $F \in \mathscr{G}$ and define $A := \Theta^{-1}(F)$. By the first assertion, A satifies (4.12). Assume there exists $B \in \mathrm{S}_2(\mathscr{H})$ satisfying (4.12). Then by linearity, A and B coincide on $\operatorname{span}\{\varphi(x) \mid x \in E\}$, which is dense in $\mathscr{H}$. By continuity, we therefore have $A = B$. The operator in $\mathrm{S}_2(\mathscr{H})$ satisfying (4.12) is therefore uniquely given by $\Theta^{-1}(F)$. ∎

$$\begin{tikzcd}
\mathrm{S}_2(\mathscr{H}) \arrow[r, leftrightarrow, "A = \sum_i \sigma_i(A)\, u_i \otimes v_i", "\text{SVD}"'] \arrow[rr, leftrightarrow, bend right, "A \leftrightarrow A\varphi(\cdot)", "\text{Corollary 4.5}"'] & \mathscr{H} \otimes \mathscr{H} \arrow[r, leftrightarrow, "\Theta(u_i \otimes v_i) = v_i(\cdot)\, u_i", "\text{Theorem 4.4}"'] & \mathscr{G}
\end{tikzcd}$$

FIGURE 2. Visualization of the isometric isomorphisms between $\mathrm{S}_2(\mathscr{H})$, $\mathscr{H} \otimes \mathscr{H}$ and $\mathscr{G}$. Here, *SVD* refers to the singular value decomposition of compact operators.

Remark 4.6 (Operator reproducing property). Not only does Corollary 4.5 describe how functions in $\mathscr{G}$ can be evaluated in terms of their operator analogue in $\mathrm{S}_2(\mathscr{H})$, it also implies the *implicit* construction of $\mathscr{G}$ via Hilbert–Schmidt operators acting on the RKHS $\mathscr{H}$. In particular, the above result shows that the space of Hilbert–Schmidt operators $\mathrm{S}_2(\mathscr{H})$ generates the vRKHS $\mathscr{G}$ via

$$\mathscr{G} = \{F : E \to \mathscr{H} \mid F = A\varphi(\cdot),\, A \in \mathrm{S}_2(\mathscr{H})\}.$$

Our previous considerations show that $\mathscr{G}$ is precisely the vRKHS associated with the vector-valued kernel $K := k\mathrm{Id}_{\mathscr{H}}$.

Corollary 4.5 will be of central importance for our approach. The identification of an $\mathscr{H}$-valued vRKHS function in $\mathscr{G}$ with a corresponding Hilbert–Schmidt operator acting on $\mathscr{H}$ will be used to bridge the gap between vector-valued statistical learning theory and the nonparametric estimation of linear operators (Grünewälder et al., 2013).

4.5. **Assumptions on $\mathscr{H}$.** We impose some technical requirements on the RKHS $\mathscr{H}$ and the corresponding kernel k. Our first three assumptions allow that we can perform Bochner integration without being caught up in measurability and integrability issues later on (Diestel and Uhl, 1977). The fourth and the fifth assumption are needed to ensure that $\mathscr{H}$ supplies the typically used approximation qualities in a function space context.

Assumption 1 (Separability)**.** The RKHS $\mathscr{H}$ is separable. Note that for a Polish space E, the RKHSs induced by a continuous kernel $k\colon E \times E \to \mathbb{R}$ is always separable (see Steinwart and Christmann, 2008, Lemma 4.33). For a more general treatment of conditions implying separability, see Owhadi and Scovel (2017).

Assumption 2 (Measurability)**.** The canonical feature map $\varphi\colon E \to \mathscr{H}$ is $\mathcal{F}_E - \mathcal{F}_{\mathscr{H}}$ measurable. This is the case when $k(x,\cdot)\colon E \to \mathbb{R}$ is $\mathcal{F}_E - \mathcal{F}_{\mathbb{R}}$ measurable for all $x \in E$. If this condition holds, then additionally all functions $f \in \mathscr{H}$ are $\mathcal{F}_E - \mathcal{F}_{\mathbb{R}}$ measurable and $k\colon E \times E \to \mathbb{R}$ is $\mathcal{F}_E^{\otimes 2} - \mathcal{F}_{\mathbb{R}}$ measurable (see Steinwart and Christmann, 2008, Lemmas 4.24 and 4.25).

Assumption 3 (Existence of second moments)**.** We have $\varphi \in L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ as well as $\varphi \in L^2(E, \mathcal{F}_E, \nu; \mathscr{H})$. Note that this is equivalent to $\mathbb{E}[\|\varphi(X)\|^2_{\mathscr{H}}] < \infty$ and $\mathbb{E}[\|\varphi(Y)\|^2_{\mathscr{H}}] < \infty$ which trivially holds for all probability measures π, ν on $(E, \mathcal{F}_E)$ case whenever $\sup_{x\in E} k(x,x) < \infty$.

Assumption 4 (C_0-kernel)**.** We assume that $\mathscr{H} \subseteq C_0(E)$, where $C_0(E)$ is the space of continuous real-valued functions on E vanishing at infinity. In particular, this is the case if $x \mapsto k(x,x)$ is bounded on E and $k(x,\cdot) \in C_0(E)$ for all $x \in E$ (Carmeli et al., 2010, Proposition 2.2).

Assumption 5 (L^2-universal kernel, see Section 4.6)**.** We assume that $\mathscr{H}$ is dense in $L^2(\pi)$. In this case, the kernel k and the RKHS $\mathscr{H}$ are called L^2-*universal* (Carmeli et al., 2010; Sriperumbudur et al., 2011).

Remark 4.7. Since not all of our results will need all of the above assumptions, we collect some general implications of the different assumptions here.

(1) Assumptions 1–3 ensures that $\mathscr{H}$ can be continuously embedded into both $L^2(\pi)$ and $L^2(\nu)$ (see Section 4.6).

(2) The combination of Assumption 4 and Assumption 5 implies that $\mathscr{H}$ is even dense in $L^2(\nu)$ for all probability measures ν on $(E, \mathcal{F}_E)$ (Carmeli et al., 2010, Theorem 4.1 and Corollary 4.2).

(3) Instead of Assumption 5, it is sometimes required in the literature that $\mathscr{H}$ is dense in $C_0(E)$ with respect to the supremum norm. This property is usually called *C_0-universality*. One can show that when Assumption 4 holds, C_0-universality is equivalent to L^2-universality (Sriperumbudur et al., 2011).

(4) When Assumptions 1–5 are satisfied, then the vRKHS $\mathscr{G}$ induced by the kernel $K = k\mathrm{Id}_{\mathscr{H}}$ is dense in both $L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ and $L^2(E, \mathcal{F}_E, \nu; \mathscr{H})$ (see Carmeli et al. 2010, Example 6.3 and Carmeli et al. 2010, Theorem 4.1). This is important for us, as we will make use of this fact later on.

Example 4.8. For $E \subseteq \mathbb{R}^d$, well-known translation invariant kernels such as the *Gaussian kernel* or *Laplacian kernel* satisfy all of the above assumptions for arbitrary probability measures π, ν on $(E, \mathcal{F}_E)$ (Sriperumbudur et al., 2011).

4.6. **Integral operators and L^2-inclusions.** The Assumptions 1–3 imply that $\mathscr{H}$ can be embedded into spaces of square integrable functions. This fact and its connections to integral operators defined by the corresponding kernels plays a fundamental role in learning theory.

4.6.1. *Real-valued RKHS.* We begin with general statements for the scalar kernel k (see for example Steinwart and Christmann, 2008, Chapter 4.3). Let the Assumptions 1–3 be satisfied. The *inclusion operator* $i_\pi : \mathscr{H} \to L^2(\pi)$ given by $f \mapsto [f]_\sim \in L^2(\pi)$ identifies $f \in \mathscr{H}$ with its equivalence class of π-a.e. defined functions in $L^2(\pi)$. It is bounded with $\|i_\pi\| \leq \|\varphi\|_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}$ and Hilbert–Schmidt. The adjoint of i_π is the integral operator $i_\pi^* \colon L^2(\pi) \to \mathscr{H}$ given by

$$[i_\pi^* f](x) = \int_E k(x, x') f(x') \,\mathrm{d}\pi(x'), \quad f \in L^2(\pi).$$

The kernel k is L^2-universal if and only if i_π^* is injective.

The operator $C_{XX} := i_\pi^* i_\pi \colon \mathscr{H} \to \mathscr{H}$ is the *kernel covariance operator* associated with the measure π given by

$$C_{XX} = \int_E \varphi(x) \otimes \varphi(x) \,\mathrm{d}\pi(x) = \mathbb{E}[\varphi(X) \otimes \varphi(X)],$$

where the integral converges in trace norm. We define all of the above concepts analogously for the measure ν and the corresponding random variable Y. The *kernel cross-covariance operator* (Baker, 1973) of X and Y is the trace class operator given by

$$C_{YX} := \iint_{E\times E} \varphi(y) \otimes \varphi(x) \, p(x, \mathrm{d}y) \mathrm{d}\pi(x) = \mathbb{E}[\varphi(Y) \otimes \varphi(X)].$$

Both operators satisfy $\langle h, C_{XX} f\rangle_{\mathscr{H}} = \langle h, f\rangle_{L^2(\pi)} = \mathbb{E}[f(X)h(X)]$ as well as $\langle h, C_{YX} f\rangle_{\mathscr{H}} = \mathbb{E}[f(X)h(Y)]$ for all $f, h \in \mathscr{H}$.

Remark 4.9 (Scalar RKHSs and integral operators). Although the presented operators $i_\pi^* : L^2(\pi) \to \mathscr{H}$, $i_\pi i_\pi^* : L^2(\pi) \to L^2(\pi)$ and $C_{XX} : \mathscr{H} \to \mathscr{H}$ have the same analytical expression as integral operators, they are fundamentally different objects since they operate on different spaces. However, $i_\pi i_\pi^*$ and C_{XX} share the same nonzero eigenvalues and their eigenfunctions can be related (see for example Rosasco et al., 2010).

4.6.2. *Vector-valued RKHS.* Similarly to the above operators defined for the scalar kernel k, we can define the above concepts for the vector-valued kernel $K = k\mathrm{Id}_{\mathscr{H}}$ in the context of Bochner spaces (Carmeli et al., 2006, 2010).

When Assumptions 1–3 are satisfied, the space $\mathscr{G}$ is separable. The elements of $\mathscr{G}$ are $\mathcal{F}_E - \mathcal{F}_{\mathscr{H}}$ measurable functions. Additionally, they are Bochner square integrable w.r.t. π. The inclusion operator $\mathcal{I}_\pi \colon \mathscr{G} \to L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ given by $F \mapsto [F]_\sim$ is bounded with $\|\mathcal{I}_\pi\| \leq \|\varphi\|_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}$.

The adjoint of $\mathcal{I}_\pi$ is the integral operator $\mathcal{I}_\pi^* \colon L^2(E, \mathcal{F}_E, \pi, \mathscr{H}) \to \mathscr{G}$ given by

$$[\mathcal{I}_\pi^* F](x) = \int_E K(x, x') F(x') \, \mathrm{d}\pi(x'), \quad F \in L^2(E, \mathcal{F}_E, \pi, \mathscr{H}).$$

The operator $T := \mathcal{I}_\pi^* \mathcal{I}_\pi \colon \mathscr{G} \to \mathscr{G}$ is the *generalized covariance operator* (also called *frame operator*, Carmeli et al. 2006) associated with the measure π given by

$$TF = \int_E K_x K_x^* F \, \mathrm{d}\pi(x) \tag{4.13}$$

for all $F \in \mathscr{G}$. T is bounded.

The following example shows that the generalized covariance operator T associated with $K(x, x') = k(x, x')\mathrm{Id}_{\mathscr{H}}$ is noncompact in general. This fact will be very important for us later on in the context of inverse problems.

Example 4.10 (Noncompact generalized covariance operator T). It is easy to see that for commonly used radial kernels k such as the Gaussian kernel on $E \subseteq \mathbb{R}^d$, the generalized covariance operator T is never compact.
Consider a measurable kernel $k : E \times E \to \mathbb{R}$ which induces an infinite-dimensional RKHS $\mathscr{H}$ satisfying Assumptions 1 and 2. Assume $k(x, y) > 0$ for all $x, y \in E$ and $k(x, x) = 1$ for all $x \in E$. Let $K = k\mathrm{Id}_{\mathscr{H}}$ and $(e_i)_{i\in\mathbb{N}} \subset \mathscr{H}$ be an ONS. We fix some $x' \in E$ and define $F_i := K_{x'} e_i \in \mathscr{G}$ for all $i \in \mathbb{N}$. Note that we have

$$\langle K_{x'} e_i,\, K_{x'} e_j \rangle_{\mathscr{G}} = \langle k(x', \cdot) e_i,\, k(x', \cdot) e_j \rangle_{\mathscr{G}} = k(x', x') \left\langle e_i,\, e_j \right\rangle_{\mathscr{H}} = \delta_{ij},$$

i.e., $(F_i)_{i\in\mathbb{N}}$ is an ONS in $\mathscr{G}$. Then it is possible to show that $(TF_i)_{i\in\mathbb{N}}$ consists of orthogonal elements of the same length:

$$\begin{aligned}
\langle TF_i,\, TF_j \rangle_{\mathscr{G}} &= \left\langle \int_E K_x F_i(x) \mathrm{d}\pi(x),\, \int_E K_x F_j(x) \mathrm{d}\pi(x) \right\rangle_{\mathscr{G}} \\
&= \left\langle \int_E k(x', x)\, K_x e_i \mathrm{d}\pi(x),\, \int_E k(x', x)\, K_x e_j \mathrm{d}\pi(x) \right\rangle_{\mathscr{G}} \\
&= \iint_{E^2} k(x', x) k(x', y) \left\langle K_y^* K_x e_i,\, e_j \right\rangle_{\mathscr{H}} \mathrm{d}[\pi \otimes \pi](x, y) \\
&= \iint_{E^2} k(x', x) k(x', y) k(x, y) \left\langle e_i,\, e_j \right\rangle_{\mathscr{H}} \mathrm{d}[\pi \otimes \pi](x, y) = M\delta_{ij}
\end{aligned}$$

with the constant $M := \iint_{E^2} k(x', x)k(x', y)k(x, y)\, \mathrm{d}[\pi \otimes \pi](x, y) > 0$, which is independent of $i, j \in \mathbb{N}$. Consequently, we have $\|TF_i - TF_j\|_{\mathscr{G}}^2 = \|TF_i\|_{\mathscr{G}}^2 + \|TF_j\|_{\mathscr{G}}^2 = 2M$ for all $i \neq j$, i.e., no subsequence of $(TF_i)_{i\in\mathbb{N}}$ can be Cauchy. We therefore have constructed a bounded sequence $(F_i)_{i\in\mathbb{N}}$ in $\mathscr{G}$ such that $(TF_i)_{i\in\mathbb{N}}$ does not contain a convergent subsequence in $\mathscr{G}$, implying that T is not compact.

4.7. **Conditional mean embeddings and regression function.** Under Assumptions 1–3, the Bochner integrability of the feature map $\varphi : E \to \mathscr{H}$ can be elegantly used in combination with the reproducing property of $\mathscr{H}$ to express expectation operations via simple linear algebra.

In particular, the *kernel mean embedding* (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Muandet et al., 2017) of the probability measure π defined by the Bochner expectation

$$\mu_\pi := \int_E \varphi(x)\, \mathrm{d}\pi(x) = \mathbb{E}[\varphi(X)] \in \mathscr{H} \tag{4.14}$$

naturally satisfies the expectation reproducing property

$$\mathbb{E}[f(X)] = \mathbb{E}\left[\langle f,\, \varphi(X)\rangle_{\mathscr{H}}\right] = \langle f,\, \mu_\pi\rangle_{\mathscr{H}} \quad \text{for all } f \in \mathscr{H}. \tag{4.15}$$

We call the RKHS $\mathscr{H}$ (or equivalently the corresponding kernel k) *characteristic*, if the *mean embedding map*

$$\pi \mapsto \int_E \varphi(x)\, \mathrm{d}\pi(x) = \mu_\pi \in \mathscr{H}$$

defined on all probability measures on $(E, \mathcal{F}_E)$ is injective.

Remark 4.11 (The RKHS $\mathscr{H}$ is characteristic). Our Assumptions 4 and 5 imply that $\mathscr{H}$ is characteristic (Carmeli et al., 2010; Sriperumbudur et al., 2010, 2011).

For two probability measures π, ν on $(E, \mathcal{F}_E)$, the *maximum mean discrepancy* (MMD) is defined by

$$d_k(\pi, \nu) := \sup_{\substack{f\in\mathscr{H} \\ \|f\|_{\mathscr{H}}\leq 1}} \left| \int_E f(x)\mathrm{d}\pi(x) - \int_E f(x)\mathrm{d}\nu(x) \right| = \|\mu_\pi - \mu_\nu\|_{\mathscr{H}}. \tag{4.16}$$

For characteristic kernels, the MMD constitutes a metric on the set of probability measures on $(E, \mathcal{F}_E)$. This fact has been used as a powerful tool in RKHS-based inference (Gretton et al., 2012b; Sejdinovic et al., 2013). The MMD can be interpreted as a so-called *integral probability metric* (Müller, 1997) and has been shown to metrize weak convergence of measures under some mild conditions (Simon-Gabriel et al., 2020).

Transferring (4.14) to a regular conditional distribution of Y given X, we define $\mathscr{H}$-valued *conditional mean embedding* (CME) function (Park and Muandet, 2020a)

$$F_p(x) := \int_E \varphi(y)\, p(x, \mathrm{d}y) = \mathbb{E}[\varphi(Y) \mid X = x] \in L^2(E, \mathcal{F}_E, \pi; \mathscr{H}) \tag{4.17}$$

and obtain a pointwise conditional version of the expectation reproducing property (4.15) as

$$\mathbb{E}[f(Y) \mid X = x] = \langle f,\, F_p(x)\rangle_{\mathscr{H}} \quad \text{for all } f \in \mathscr{H} \text{ and } x \in E. \tag{CME}$$

The fact that F_p (or analogously any other regular version of $\mathbb{E}[\varphi(Y) \mid X = \cdot])$ is a well-defined element in $L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ can be seen by using Jensen's inequality for conditional Bochner expectations as

$$\begin{aligned}\|F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} &= \int_E \|F_p(x)\|^2_{\mathscr{H}}\,\mathrm{d}\pi(x)\\ &\leq \iint_{E\times E} \|\varphi(y)\|^2_{\mathscr{H}}\, p(x, \mathrm{d}y)\mathrm{d}\pi(x) = \mathbb{E}[\|\varphi(Y)\|^2_{\mathscr{H}}] < \infty.\end{aligned}$$

together with Assumption 3.

The approximation of F_p is a key concept in a wide variety of models for kernel-based inference. If C_{XX} is injective, Song et al. (2009) and Fukumizu et al. (2013) show that under the assumption

$$\mathbb{E}[f(Y) \mid X = \cdot] = \langle f,\, F_p(\cdot)\rangle_{\mathscr{H}} \in \mathscr{H} \quad \text{for all } f \in \mathscr{H}, \tag{4.18}$$

we have a closed form expression of F_p via

$$F_p(x) = C_{YX} C_{XX}^{\dagger} \varphi(x) \tag{4.19}$$

for all $x \in E$ such that $\varphi(x) \in \operatorname{range}(C_{XX})$. Here, the (generally unbounded and not globally defined) operator $C_{XX}^{\dagger}\colon\ : \operatorname{range}(C_{XX}) + \operatorname{range}(C_{XX})^{\perp} \to \mathscr{H}$ is the *Moore–Penrose pseudoinverse* of C_{XX} (see Engl et al. 1996). The assumption (4.18) is generally not satisfied (see Klebanov et al. 2020a for a detailed investigation of arising problems). Grünewälder et al. (2012) and Park and Muandet (2020a) show that a Tikhonov–Phillips regularized version of the estimate of (4.19) can be understood as an empirical approximation of F_p with functions in $\mathscr{G}$ in a least squares regression context. However, no approximation qualities of the CME in the L^2-operator context are considered. We will now extend this theory and connect it to the CME regression model later on.

5. Nonparametric approximation of P

We now restate the main results from Section 3 with detailed assumptions and provide their proofs. Furthermore, we investigate the connections of the approximation of P over functions in $\mathscr{H}$ to the maximum mean discrepancy and regularized least squares regression.

5.1. **Proofs of main results.** We begin with the proof of Theorem 3.5, as it constitutes the theoretical foundation for our remaining work. We note that this result can also be interpreted as an improvement of a surrogate risk bound derived by Grünewälder et al. (2012, Section 3.1) and later on used by Park and Muandet (2020a) to approximate the CME. We will elaborate on this fact in more detail later on (see Section 5.3 and Remark 5.9 in particular). Our bound has a significant impact from an approximation viewpoint, which we will highlight in our following examination.

Theorem 3.5 (Regression and conditional mean approximation)**.** *Under the Assumptions 1–3, we have for every operator $A \in \mathrm{S}_2(\mathscr{H})$ that*

$$\|A - P\|^2_{\mathscr{H}\to L^2(\pi)} \leq \mathbb{E}\left[\|F_p(X) - A^*\varphi(X)\|^2_{\mathscr{H}}\right] = \|F_p - A^*\varphi(\cdot)\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}.$$

The given bound is sharp.

Proof. Let $A \in \mathrm{S}_2(\mathscr{H})$. We have

$$\begin{aligned}
\|A - P\|^2_{\mathscr{H}\to L^2(\pi)} &= \sup_{\|f\|_{\mathscr{H}}=1} \|Af - Pf\|^2_{L^2(\pi)} \\
&= \sup_{\|f\|_{\mathscr{H}}=1} \|[Af](\cdot) - \mathbb{E}[f(Y) \mid X = \cdot]\|^2_{L^2(\pi)} \\
&= \sup_{\|f\|_{\mathscr{H}}=1} \left\|\langle Af,\, \varphi(\cdot)\rangle_{\mathscr{H}} - \langle f,\, F_p(\cdot)\rangle_{\mathscr{H}}\right\|^2_{L^2(\pi)} \\
&= \sup_{\|f\|_{\mathscr{H}}=1} \left\|\langle f,\, A^*\varphi(\cdot) - F_p(\cdot)\rangle_{\mathscr{H}}\right\|^2_{L^2(\pi)} \\
&= \sup_{\|f\|_{\mathscr{H}}=1} \mathbb{E}\left[\langle f,\, A^*\varphi(X) - F_p(X)\rangle^2_{\mathscr{H}}\right] \\
&\leq \sup_{\|f\|_{\mathscr{H}}=1} \mathbb{E}\left[\|f\|^2_{\mathscr{H}}\, \|A^*\varphi(X) - F_p(X)\|^2_{\mathscr{H}}\right] \\
&= \mathbb{E}\left[\|A^*\varphi(X) - F_p(X)\|^2_{\mathscr{H}}\right] = \|A^*\varphi(\cdot) - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})},
\end{aligned}$$

where we use the reproducing property in $\mathscr{H}$ in the third equality and the Cauchy–Schwarz inequality. It is clear that the above bound is sharp by considering the case that we have $\mathbb{P}$-a.e. $A^*\varphi(X) - F_p(X) = h$ for some constant $h \in \mathscr{H}$. In this case the above bound is attained when we choose $f = h/\|h\|_{\mathscr{H}}$ in the supremum. ∎

Theorem 3.2 (Approximation by Hilbert–Schmidt operators)**.** *Let Assumptions 1-5 be satisfied. Then for every $\delta > 0$, there exists a Hilbert–Schmidt operator $A\colon \mathscr{H} \to \mathscr{H}$, such that*

$$\|A - P\|_{\mathscr{H}\to L^2(\pi)} < \delta. \tag{5.1}$$

Proof. By Corollary 4.5, every operator $A^* \in \mathrm{S}_2(\mathscr{H})$ corresponds to a function $F \in \mathscr{G}$ via $F(x) = A^*\varphi(x)$ for all $x \in E$ and vice versa. The space $\mathscr{G}$ is densely embedded into $L^2(E,\mathcal{F}_E,\pi;\mathscr{H})$ by Remark 4.7(4). For every $\delta > 0$ we therefore have an operator $A^* \in \mathrm{S}_2(\mathscr{H})$ such that the bound $\|A^*\varphi(\cdot) - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} = \|F - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} < \delta$ holds. Together with the bound obtained in Theorem 3.5, this proves the assertion. ∎

Corollary 3.4. *Let Assumptions 1-5 be satisfied. Then there exists a sequence of finite-rank operators $(A_n)_{n\in\mathbb{N}}$ from $\mathscr{H}$ to $\mathscr{H}$ such that $\|A_n - P\|_{\mathscr{H}\to L^2(\pi)} \to 0$ as $n \to \infty$.*

Proof. Let $\delta > 0$. By the fact that the finite-rank operators on $\mathscr{H}$ are dense in $\mathrm{S}_2(\mathscr{H})$ and Theorem 3.5, we can choose $A \in \mathrm{S}_2(\mathscr{H})$ as well as a finite-rank operator A_n on $\mathscr{H}$ such that

$$\begin{aligned} \|A_n - P\|_{\mathscr{H}\to L^2(\pi)} &\leq \|A - P\|_{\mathscr{H}\to L^2(\pi)} + \|i_\pi\| \, \|A_n - A\|_{\mathscr{H}\to\mathscr{H}} \\ &\leq \|A - P\|_{\mathscr{H}\to L^2(\pi)} + \|i_\pi\| \, \|A_n - A\|_{\mathrm{S}_2(\mathscr{H})} < \frac{\delta}{2} + \frac{\delta}{2}. \end{aligned}$$

■

5.2. **Measure-theoretic implications of the approximation of P.** When $\mathscr{H}$ is characteristic, $P : \mathscr{H} \to L^2(\pi)$ uniquely determines the conditional distribution $p(x, \cdot)$ for π-a.e. $x \in E$ (that is, up to a choice of a regular version of the underlying conditional expectation). This underlines that the conditional expectation operator P interpreted as an operator with the domain $\mathscr{H}$ instead of $L^2(\nu)$ still captures sufficient information about the underlying joint distribution of X and Y. More generally, an approximation of P naturally yields a weighted approximation of the associated Markov kernel p in the MMD. This may provide a foundation for the adaptation of MMD-based hypothesis tests for Markov kernels.

Theorem 5.1 (Equivalence to approximation in MMD)**.** *Let Assumptions 1–3 be satisfied. Let $P, P' : \mathscr{H} \to L^2(\pi)$ be two well-defined bounded conditional expectation operators associated with the Markov kernels $p, p' : E \times \mathcal{F}_E \to \mathbb{R}$. Then we have*

$$\|P - P'\|^2_{\mathscr{H}\to L^2(\pi)} = \int_E d_k(p(x,\cdot), p'(x,\cdot))^2 \, \mathrm{d}\pi(x). \tag{5.2}$$

Proof. We have

$$\begin{aligned} \|P - P'\|^2_{\mathscr{H}\to L^2(\pi)} &= \sup_{\substack{f\in\mathscr{H} \\ \|f\|_{\mathscr{H}}=1}} \|Pf - P'f\|^2_{L^2(\pi)} \\ &= \sup_{\substack{f\in\mathscr{H} \\ \|f\|_{\mathscr{H}}=1}} \left\| \int f(y)\, p(\cdot, \mathrm{d}y) - \int f(y)\, p'(\cdot, \mathrm{d}y) \right\|^2_{L^2(\pi)} \\ &= \sup_{\substack{f\in\mathscr{H} \\ \|f\|_{\mathscr{H}}=1}} \left\| \left\langle f, \, \int_E \varphi(y)\, p(\cdot, \mathrm{d}y) - \int_E \varphi(y)\, p'(\cdot, \mathrm{d}y) \right\rangle_{\mathscr{H}} \right\|^2_{L^2(\pi)} \\ &= \int_E \left\| \mu_{p(x,\cdot)} - \mu_{p'(x,\cdot)} \right\|^2_{\mathscr{H}} \mathrm{d}\pi(x) \\ &= \int_E d_k(p(x,\cdot), p'(x,\cdot))^2 \, \mathrm{d}\pi(x), \end{aligned}$$

where we use the reproducing property in $\mathscr{H}$ in the third equality. ■

Under more restrictive assumptions, the low-dimensional approximation of the adjoint of P by means of the MMD has been proposed in the context of random dynamical systems with a different estimation scheme (Tian and Wu, 2020).

Remark 5.2 (Assumptions of Theorem 5.1). For simplicity, we do not explicitly assume in Theorem 5.1 that the underlying random variables associated with P and P' are distributed with respect to the marginals π and ν. To show the above statement, it is sufficient that both operators are well-defined and bounded when the domain and image space and domain are chosen to be $\mathscr{H}$ and $L^2(\pi)$. The proof of Theorem 5.1 shows that $\|P - P'\|^2_{\mathscr{H} \to L^2(\pi)}$ can equivalently be interpreted as the squared $L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ distance between the two conditional mean embeddings $F_p(x) = \int_E \varphi(y)\, p(x, \mathrm{d}y)$ and $F_p'(x) = \int_E \varphi(y)\, p'(x, \mathrm{d}y)$.

Remark 5.3 (Approximation of MMD in L^q-norm). Whenever the conditional expectation operators P, P' are well-defined and bounded operators from $\mathscr{H}$ to $L^q(\pi)$ for $1 \leq q \leq \infty$, we can analogously obtain versions of (5.2) for the respective L^q norm. In particular, in this case we have

$$\|P - P'\|^q_{\mathscr{H} \to L^q(\pi)} = \int_E d_k(p(x, \cdot), p'(x, \cdot))^q \, \mathrm{d}\pi(x). \tag{5.3}$$

When $\mathscr{H}$ is characteristic, we immediately obtain the following result. It shows that conditional expectation operators on $\mathscr{H}$ determine the conditional distribution of the associated random variables uniquely up to a choice of a regular version.

Corollary 5.4. *Let Assumptions 1–3 be satisfied and $\mathscr{H}$ be characteristic. With the notation of Theorem 5.1, we have $\|P - P'\|_{\mathscr{H} \to L^2(\pi)} = 0$ if and only if $p(x, \cdot) = p'(x, \cdot)$ for π-a.e. $x \in E$.*

Moreover, Corollary 5.4 implies that the joint distributions for the class of pairs of random variables X, Y with a fixed marginal $X \sim \pi$ are uniquely determined by $P : \mathscr{H} \to L^2(\pi)$.

Corollary 5.5. *Let X, X', Y, Y' be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $(E, \mathcal{F}_E)$ such that $X \sim \pi$ and $X' \sim \pi$ and Assumptions 1–3 are satisfied for both pairs X, Y and X', Y'. Let $\mathscr{H}$ be characteristic and $P, P' : \mathscr{H} \to L^2(\pi)$ be bounded conditional expectation operators given by $Pf = \mathbb{E}[f(Y) \mid X = \cdot]$ and $P'f = \mathbb{E}[f(Y') \mid X' = \cdot]$ defined by some Markov kernels p and p' respectively. Then we have $\|P - P'\|_{\mathscr{H} \to L^2(\pi)} = 0$ if and only if $\mathcal{L}(X, Y) = \mathcal{L}(X', Y')$.*

Proof. Let $\|P - P'\|_{\mathscr{H} \to L^2(\pi)} = 0$. For any two events $\mathcal{A}, \mathcal{B} \in \mathcal{F}_E$, we perform the disintegration

$$\mathbb{P}[X \in \mathcal{A}, Y \in \mathcal{B}] = \int_{\mathcal{A}} p(x, \mathcal{B}) \, \mathrm{d}\pi(x) \tag{5.4}$$

and analogously for the pair X', Y'. We apply Corollary 5.4, leading to the π-a.e. equivalence $p(\cdot, \mathcal{B}) = p'(\cdot, \mathcal{B})$. This gives $\mathbb{P}[X \in \mathcal{A}, Y \in \mathcal{B}] = \mathbb{P}[X' \in \mathcal{A}, Y' \in \mathcal{B}]$. The converse implication follows analogously. ■

5.3. **Least squares regression and connection to the CME.** We now describe the theoretical foundation of estimating P based on Theorem 3.5. In the process, we will see that our concept is closely related to the CME.

By the operator reproducing property from Corollary 4.5 we may rewrite the vRKHS least squares regression problem

$$\operatorname*{arg\,min}_{F \in \mathscr{G}} R(F) \text{ with } R(F) := \mathbb{E}[\|\varphi(Y) - F(X)\|^2_{\mathscr{H}}] \tag{5.5}$$

equivalently as

$$\operatorname*{arg\,min}_{A^* \in \mathrm{S}_2(\mathscr{H})} \mathbb{E}[\|\varphi(Y) - A^*\varphi(X)\|^2_{\mathscr{H}}]. \tag{5.6}$$

As is well-known in statistical learning theory (see for example Cucker and Smale, 2002, Proposition 1), for all $F \in L^2(E, \mathcal{F}, \pi; \mathscr{H})$, the risk R allows for the decomposition

$$R(F) = \|F_p - F\|^2_{L^2(E,\mathcal{F},\pi;\mathscr{H})} + R(F_p), \tag{5.7}$$

where $R(F_p)$ represents the irreducible error term (see Theorem A.1 for a proof in the infinite-dimensional case). This reduces the regression problem (5.5) and equivalently problem (5.6) to an L^2-approximation of the conditional mean embedding F_p. In this context, F_p is often called *regression function*. Therefore, the so-called *excess risk* $R(F) - R(F_p) = \|F_p - F\|^2_{L^2(E,\mathcal{F},\pi;\mathscr{H})}$ of some estimate $F \in \mathscr{G}$ is typically investigated in nonparametric statistics.

The above formalism allows us to estimate the conditional mean operator P based on our previous results. By Theorem 3.5, we have

$$\|A - P\|^2_{\mathscr{H} \to L^2(\pi)} \leq \|F_p - A^*\varphi(\cdot)\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} \tag{5.8}$$

for all $A^* \in \mathrm{S}_2(\mathscr{H})$. We can now perform the vRKHS regression (5.6) and obtain an approximation of P in the norm $\|\cdot\|^2_{\mathscr{H} \to L^2(\pi)}$ in terms of $A \in \mathrm{S}_2(\mathscr{H})$, which we implicitly interpret as an operator from $\mathscr{H}$ to $L^2(\pi)$. Theorem 3.2 and Corollary 3.4 show that this is possible up to an arbitrary degree of accuracy.

Along the lines of the known work on least squares regression of the form (5.5) or equivalently (5.6), we can distinguish following two general cases (Szabó et al., 2016):

(1) The *well-specified case*, i.e., there exists a regular version of the conditional distribution of Y given X such that $F_p(\cdot) = \mathbb{E}[\varphi(Y) \mid X = \cdot] \in \mathscr{G}$. For the well-specified case, we below obtain the known properties of the conditional mean embedding which were derived from the linear-algebraic perspective (Song et al., 2009; Klebanov et al., 2020a,b).

(2) The *misspecified case*, i.e., $F_p \in L^2(\pi) \setminus \mathscr{G}$. This is clearly the more interesting setting, as the well-specified case does typically not occur in practice. From the operator-theoretic perspective, this case has not been investigated yet.

Our previous results allow to reformulate the well-specified case and establish a connection to the CME.

Corollary 5.6 (Well-specified case). *Let Assumption 1–3 be satisfied. Consider a fixed regular version of the distribution of Y conditioned on X given by some Markov kernel $p : E \times \mathcal{F}_E \to \mathbb{R}$. The following statements are equivalent:*

(i) We have $F_p(\cdot) = \mathbb{E}[\varphi(Y) \mid X = \cdot] \in \mathscr{G}$.

(ii) There exists an operator $A \in \mathrm{S}_2(\mathscr{H})$ such that

$$[Af](x) = \langle Af,\, \varphi(x)\rangle_{\mathscr{H}} = \langle f,\, A^*\varphi(x)\rangle_{\mathscr{H}} = \mathbb{E}[f(Y) \mid X = x] \tag{5.9}$$

for all $x \in E$ and $f \in \mathscr{H}$.

Both (i) and (ii) imply (iii):

(iii) There exists an operator $A \in \mathrm{S}_2(\mathscr{H})$ which satisfies $\|A - P\|_{\mathscr{H}\to L^2(\pi)} = 0$.

Proof. We show that (i) is equal to (ii). Let $F_p(\cdot) = \mathbb{E}[\varphi(Y) \mid X = \cdot] \in \mathscr{G}$. Let $A^* \in \mathrm{S}_2(\mathscr{H})$ be the unique operator such that $A^*\varphi(\cdot) = F_p(\cdot)$ by Corollary 4.5. By the reproducing property in $\mathscr{H}$, we can verify (5.9) immediately. For the converse implication, let (5.9) be satisfied for some operator $A^* \in \mathrm{S}_2(\mathscr{H})$. Then by Corollary 4.5, we have the function $F \in \mathscr{G}$ with $F(\cdot) = A^*\varphi(\cdot)$ such that

$$\langle f,\, F(x)\rangle_{\mathscr{H}} = \mathbb{E}[f(Y) \mid X = x] = \mathbb{E}[\langle f,\, \varphi(Y)\rangle_{\mathscr{H}} \mid X = x] \tag{5.10}$$

for all $f \in \mathscr{H}$. The right hand side of 5.10 is equal to $\langle f,\, \mathbb{E}[\varphi(Y)_{\mathscr{H}} \mid X = x]\rangle_{\mathscr{H}}$ for all $x \in E$ and $f \in \mathscr{H}$, we therefore have $F(\cdot) = \mathbb{E}[\varphi(Y) \mid X = \cdot] = F_p(\cdot) \in \mathscr{G}$ as claimed. The last statement follows from Theorem 3.5 by inserting A^* into the right hand side of the bound, giving $\|A - P\|_{\mathscr{H}\to L^2(\pi)} = 0$. ■

Remark 5.7 (Connection to CME and well-specified case). By comparing (5.9) to the expectation reproducing property (CME), we see that in the well-specified case, the operator A^* satisfying (5.9) is exactly the operator which was introduced by Song et al. (2009) as the original conditional mean embedding. *That is, we obtain the approximation of P from $\mathscr{H}$ to $L^2(\pi)$ as the adjoint of the CME.* A similar connection was established by Klus et al. (2020) under the restrictive assumptions of Song et al. (2009) in the context of Markov operators.

Remark 5.8 (Well-specified case closed form solution)*.* Klebanov et al. (2020b, Theorem 5.8) prove in a slightly different context of tensor product spaces without explicitly using vRKHSs, that in the well-specified case the operator A^* satisfying (5.9) can be expressed in terms of the covariance operators as $A^* = (C_{XX}^{\dagger} C_{XY})^*$. In fact, this proves that $(C_{XX}^{\dagger} C_{XY})^*$ is Hilbert–Schmidt in this case.

Remark 5.9 (Surrogate risk bound for the CME)*.* In the well-specified case, Park and Muandet (2020a) investigate the estimation of the CME in terms of (5.5). Their results build upon the surrogate risk bound

$$\|A - P\|^2_{\mathscr{H}\to L^2(\pi)} \le R(A^*\varphi(\cdot)),$$

originally formulated by Grünewälder et al. (2012). Our Theorem 3.5 improves this bound and eliminates the need for additional approximation results (see for example Grünewälder et al., 2012, Theorem 3.2) for the analysis of the misspecified case. By (5.7), our bound from Theorem 3.5 equals to

$$\|A - P\|^2_{\mathscr{H}\to L^2(\pi)} \le R(A^*\varphi(\cdot)) - R(F_p),$$

which allows the approximation up to an arbitrary accuracy and removes the excess term $R(F_p)$.

We have seen that in the well-specified case, our results align with prior work on the CME. In the practically more relevant misspecified case however, the bound given by Theorem 3.5 significantly simplifies the theory of approximating the CME. For the remainder of the paper, we will focus on the empirical estimation of P without restricting ourselves to the well-specified case.

6. Empirical estimation and regularization theory

We now connect our previous results to the theory of supervised learning and derive empirical estimators of P. To this end, we will briefly review how the regression problem (5.5) can be formulated in terms of an inverse problem. The decomposition of R in (5.7) allows to obtain a solution by approximating F_p with functions in $\mathscr{G}$. This framework allows to derive the well-known formalism for supervised learning and regularization theory which will yield estimates of P. We refer to the seminal work for least squares regression with vRKHSs (De Vito and Caponnetto, 2005; Caponnetto and De Vito, 2007) for more details. This section contains the reformulation of our setting in terms of known results, making the theory of vRKHS regression applicable for the estimation of P. We use this framework to derive new results in Section 7.

6.1. **Inverse problem.** In the misspecified case, it is not necessarily clear that the minimizer of R over $\mathscr{G}$ exists. The analytical nature of this question can be naturally expressed in terms of an inverse problem. For the necessary background on inverse problems in Hilbert spaces and regularization theory, we refer to Engl et al. (1996). We will formulate (5.5) a bit more verbosely in terms of the inclusion $\mathcal{I}_\pi : \mathscr{H} \to L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$, so that the connection to the inverse problem becomes clear.

If $F \in \mathscr{G}$, we have by (5.7) that

$$R(F) = \|\mathcal{I}_\pi F - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} + R(F_p).$$

Finding $F_{\mathscr{G}} := \arg\min_{F\in\mathscr{G}} R(F)$ is therefore equivalent to finding $F_{\mathscr{G}} \in \mathscr{G}$ such that

$$\|\mathcal{I}_\pi F - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}$$

is minimal. As is well-known from the theory of inverse problems, this is equivalent to finding the optimal solution $F_{\mathscr{G}}$ of the potentially ill-posed inverse problem

$$\mathcal{I}_\pi F = F_p, \quad F \in \mathscr{G}. \tag{6.1}$$

The inverse problem (6.1) is again equivalent to finding the solution of the so-called *normal equation* (Engl et al., 1996, Theorem 2.6) given by

$$(\mathcal{I}_\pi^* \mathcal{I}_\pi)F = TF = \mathcal{I}_\pi^* F_p, \quad F \in \mathscr{G}.$$

In particular, we obtain the following solution.

Theorem 6.1 (Regression solution)**.** *Let Assumptions 1–3 be satisfied. The optimal solution*

$$F_{\mathscr{G}} = \operatorname*{arg\,min}_{F\in\mathscr{G}} R(F) = \operatorname*{arg\,min}_{F\in\mathscr{G}} \|\mathcal{I}_\pi F - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}$$

exists if and only if $\mathcal{I}_\pi^* F_p \in \operatorname{range}(T) + \operatorname{range}(T)^\perp =: \operatorname{dom}(T^\dagger)$,[2] *where the operator* $T^\dagger : \operatorname{range}(T) + \operatorname{range}(T)^\perp \to \mathscr{G}$ *is the pseudoinverse of* T*. In this case,* $F_\mathscr{G}$ *is given by the solution to the normal equation*

$$TF = \mathcal{I}_\pi^* F_p, \quad F \in \mathscr{G} \tag{6.2}$$

in terms of $F_\mathscr{G} = T^\dagger \mathcal{I}_\pi^* F_p$.

Remark 6.2 (Limitations of existing literature). Theorem 6.1 and the resulting normal equation (6.2) show that our surrogate problem is essentially a (potentially ill-posed) inverse problem with the following technical features:

(i) both the forward operator T and the right-hand side $\mathcal{I}_\pi^* F_p$ are unknown and must be discretized by sampling from $\mathcal{L}(Y, X)$ and

(ii) the forward operator T is in general not compact but only bounded, as Example 4.10 shows.

However, results on uniform upper and lower bounds for convergence rates in a vector-valued learning scenario are typically investigated in the case where the forward operator T of the problem (6.2) is trace class (Caponnetto and De Vito, 2007; Rastogi and Sampath, 2017; Rastogi et al., 2020, and references therein). In particular, the aforementioned authors assume $K_x K_x^* \in S_1(\mathscr{H})$ for all $x \in E$ and use the *effective dimension*

$$\mathcal{N}(\lambda) := \operatorname{Tr}\left((T + \lambda \mathrm{Id}_\mathscr{G})^{-1} T\right) \quad \text{ for } \lambda > 0$$

as the central tool in order to analyze the convergence of kernel-based regression problems (the reader may also refer to Blanchard and Mücke 2018; Lu et al. 2020; Lin et al. 2020 for the scalar case). Example 4.10 shows that $\mathcal{N}(\lambda)$ is generally not finite in our setting.
Moreover, most results on statistical inverse problems with noncompact forward operators seem to be derived under the assumption that the forward operator is known (see for example Cavalier, 2006; Bissantz et al., 2007) and do therefore not directly transfer to our scenario. Adapting these results in our setting would need a thorough perturbation analysis of the continuous spectrum of T. Moreover, discretizing $\mathscr{G}$ in the noncompact case may introduce additional difficulties, see Remark 6.3.
To the best of our knowledge, Park and Muandet (2020a,b) are the only authors who address the estimation under assumptions which are satisfied in our case (see Remark 7.1). As these problems require a deeper analysis in the context of inverse problems, they are out of the scope of this work.

Remark 6.3 (Discretization of T). Note that due to the noncompactness of $\mathscr{G}$, a bit of caution is required when discussing its discretization. In particular, a naive estimate of T would be the Monte Carlo sum

$$T_n := \frac{1}{n} \sum_{i=1}^{n} K_{X_i} K_{X_i}^*$$

[2] An equivalent condition is $\Pi F_p \in \operatorname{range}(\mathcal{I}_\pi)$, where $\Pi \colon L^2(E, \mathcal{F}_E, \pi; \mathscr{H}) \to L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ is the orthogonal projection onto the closure of $\operatorname{range}(\mathcal{I}_\pi)$.

for iid $X_i \sim \pi$ and one would think that some strong law of large numbers and concentration results in Banach spaces would lead to the desired convergence results $T_n \to T$ in operator norm. Note that the Banach space of bounded operators on $\mathscr{G}$ is not separable, even if $\mathscr{G}$ itself is separable.[3] This fact may lead to measurability issues of the $\mathfrak{B}(\mathscr{G})$-valued object

$$\xi := K_X K_X^*.$$

Because of this fact, we defined the operator T in (4.13) pointwise as

$$TF = \int_E K_x K_x^* F \,\mathrm{d}\pi(x), \quad F \in \mathscr{G}$$

instead of an integral over the object ξ as defined above – which would need to converge in operator norm. As previously mentioned, available literature on vector-valued regression imposes the assumption $K_x K_x^* \in S_1(\mathscr{H})$, which is not satisfied in our scenario. In addition, versions of the strong law of large numbers in Banach spaces typically require additional properties (Ledoux and Talagrand, 1991, Section 7). For simplicity, we will therefore consider the strong operator topology formulation

$$TF = \int_E K_x K_x^* F \,\mathrm{d}\pi(x) = \mathbb{E}[\xi F] \in \mathscr{G} \tag{6.3}$$

for every $F \in \mathscr{G}$ instead of the norm topology on $\mathfrak{B}(\mathscr{G})$.

6.2. **Regularization and empirical estimation.** For simplicity, we assume that the optimal solution $F_{\mathscr{G}} = \arg\min_{\mathscr{G}} R(F)$ exists, i.e., we have $\mathcal{I}_\pi^* F_p \in \operatorname{dom}(T^\dagger)$. We wish to compute a solution of the normal equation

$$TF = \mathcal{I}_\pi^* F_p, \quad F \in \mathscr{G}. \tag{6.4}$$

in terms of $F_{\mathscr{G}} = T^\dagger \mathcal{I}_\pi^* F_p$ based on an empirical realization of $(X_t)_{t\in\mathbb{Z}}$.

In order to do this, we must discretize T as well as the right-hand side $\mathcal{I}_\pi^* F_p$. We now face the problem that (6.4) may be *ill-posed* in the sense that the solution does not continously depend on $\mathcal{I}_\pi^* F_p$ (and of course on T as well). To still be able to perform an estimation, a *regularization strategy* (Engl et al., 1996) is needed to ensure well-posedness in practice.

Let $\{g_\lambda(T) : \mathscr{G} \to \mathscr{G} \mid \lambda \in (0, \infty]\}$ be a regularization strategy.[4] For a fixed regularization parameter $\lambda > 0$, we define the regularized solution

$$F_\lambda := g_\lambda(T)\mathcal{I}_\pi^* F_p \in \mathscr{G}. \tag{6.5}$$

We now discretize the regularized problem (6.5) based on the empirical data $\mathbf{z} := ((X_1, Y_1), \ldots, (X_n, Y_n))$, where we assume iid $(X_i, Y_i) \sim \mathcal{L}(X, Y)$. We generalize the *sampling operator approach* (Smale and Zhou, 2005) from the scalar setting to the vector-valued scenario and derive an empirical estimate of F_λ. Given the data above, we define the *sampling operator* $S_{\mathbf{x}} \colon \mathscr{G} \to \mathscr{H}^n$ given by $S_{\mathbf{x}} F := (F(X_t))_{t=1}^n =$

[3]This can be proven with the fact that the sequence space ℓ^∞ – which is not separable – can be isometrically embedded into $\mathfrak{B}(\mathscr{G})$.

[4]We require $\{g_\lambda(T) : \mathscr{G} \to \mathscr{G} \mid \lambda \in (0, \infty]\}$ to be a parametrized family of *globally defined bounded operators* satisfying $g_\lambda(T)F \to T^\dagger F$ for all $F \in \operatorname{dom}(T^\dagger)$ as $\lambda \to 0$.

$(K_{X_t}^* F)_{t=1}^n$. Here, we consider $\mathscr{H}^n$ as a Hilbert space equipped with the inner product

$$\langle \mathbf{f},\, \mathbf{h} \rangle_{\mathscr{H}^n} := \frac{1}{n} \sum_{i=1}^{n} \langle f_i,\, h_i \rangle_{\mathscr{H}}$$

for $\mathbf{f} = (f_1, \dots, f_n) \in \mathscr{H}^n$ and $\mathbf{h} = (h_1, \dots, h_n) \in \mathscr{H}^n$. It is easy to see that the adjoint of $S_{\mathbf{x}}$ is the operator $S_{\mathbf{x}}^* \colon \mathscr{H}^n \to \mathscr{G}$ given by

$$S_{\mathbf{x}}^* \mathbf{h} = \frac{1}{n} \sum_{i=1}^{n} K_{X_i} h_i$$

for all $\mathbf{h} \in \mathscr{H}^n$ and the operator $T_{\mathbf{x}} := S_{\mathbf{x}}^* S_{\mathbf{x}} \colon \mathscr{G} \to \mathscr{G}$ satisfies

$$T_{\mathbf{x}} F = S_{\mathbf{x}}^* S_{\mathbf{x}} F = \frac{1}{n} \sum_{i=1}^{n} K_{X_i} K_{X_i}^* F$$

for all $F \in \mathscr{G}$. Based on these considerations, we will use $S_{\mathbf{x}}^*$ and $T_{\mathbf{x}}$ as empirical estimates for $\mathcal{I}_\pi^*$ and T respectively based on the data $\mathbf{x}$. We define the *target data vector* $\Upsilon := (\varphi(Y_1), \dots, \varphi(Y_n)) \in \mathscr{H}^n$ and obtain the empirical regularized solution

$$F_{\lambda,\mathbf{z}} := g_\lambda(T_{\mathbf{x}}) S_{\mathbf{x}}^* \Upsilon \in \mathscr{G} \tag{6.6}$$

as the discretized analogue of the analytical regularized solution (6.5).

Via the identification of F_λ and $F_{\lambda,\mathbf{z}}$ with operators through the isomorphism Θ in Corollary 4.5, we obtain the *analytical regularized operator solution*

$$A_\lambda := [\Theta^{-1}(F_\lambda)]^* \in \mathrm{S}_2(\mathscr{H})$$

as well as the *empirical regularized operator solution*

$$A_{\lambda,\mathbf{z}} := [\Theta^{-1}(F_{\lambda,\mathbf{z}})]^* \in \mathrm{S}_2(\mathscr{H}),$$

i.e., $F_\lambda(x) = A_\lambda \varphi(x)$ and $F_{\lambda,\mathbf{z}}(x) = A_{\lambda,\mathbf{z}} \varphi(x)$ for all $x \in E$.

7. Tikhonov-Phillips regularization

For the remainder of this paper, we will restrict ourselves to the Tikhonov–Phillips regularization approach (Phillips, 1962; Tikhonov and Arsenin, 1977) to solve the (potentially ill-posed) inverse problem given by Theorem 6.1 in order to obtain the optimal solution $F_{\mathscr{G}}$ in $\mathscr{G}$ of the surrogate problem (assuming it exists).

7.1. **General framework.** Tikhonov–Phillips regularization corresponds to the regularization strategy $g_\lambda(T) := (T + \lambda \mathrm{Id}_{\mathscr{G}})^{-1} \in \mathfrak{B}(\mathscr{G})$ for $\lambda > 0$. We replace the risk R with the *regularized risk*

$$R_\lambda(F) := R(F) + \lambda \left\| F \right\|_{\mathscr{G}}^2 \tag{7.1}$$

with a regularization parameter $\lambda > 0$. The unique minimizer of (7.1) exists for all $\lambda > 0$ and is exactly given by the regularized solution $F_\lambda = (T + \lambda \mathrm{Id}_{\mathscr{G}})^{-1} \mathcal{I}_\pi^* F_p$, which is a standard result in inverse problems (Engl et al., 1996, Theorem 5.1). Based on the data $\mathbf{z}$, we define the *regularized empirical risk*

$$R_{\lambda,\mathbf{z}}(F) := \frac{1}{n} \sum_{i=1}^{n} \left\| \varphi(Y_i) - F(X_i) \right\|_{\mathscr{H}}^2 + \lambda \left\| F \right\|_{\mathscr{G}}^2 \tag{7.2}$$

for all $F \in \mathscr{G}$. We can reformulate (7.2) in terms of the sampling operator equivalently as $R_{\lambda,\mathbf{z}}(F) = \left\| S_{\mathbf{x}} F - \Upsilon \right\|_{\mathscr{H}^n}^2 + \lambda \left\| F \right\|_{\mathscr{G}}^2$ for all $F \in \mathscr{G}$. Therefore,

$R_{\lambda,\mathbf{z}}$ admits a unique minimizer in $\mathscr{G}$ given by the regularized empirical solution $F_{\lambda,\mathbf{z}} = (T_{\mathbf{x}} + \lambda \mathrm{Id}_{\mathscr{G}})^{-1} S_{\mathbf{x}}^* \Upsilon$, which we will consider from now on as the estimate of F_λ.

Remark 7.1 (Uniform convergence rates). As mentioned previously in Remark 6.2, uniform convergence rates of $F_{\lambda,\mathbf{z}}$ can only be achieved under additional smoothness assumptions on F_p. Park and Muandet (2020a,b) investigate Tikhonov–Phillips regularization in the well specified case, i.e., $F_p \in \mathscr{G}$ and show $R(F_{\lambda,\mathbf{z}}) - R(F_p) \in \mathcal{O}_p(n^{-1/4})$ for the regularization scheme $\lambda(n) \in \mathcal{O}(n^{-1/4})$ as $n \to \infty$ whenever the kernel k is bounded.

7.2. **Closed form Tikhonov–Phillips operator estimates.** We show that for the Tikhonov–Phillips estimate, the adjoint of the regularized analytical operator solution $A_\lambda^* = \Theta^{-1}(F_\lambda)$ which satisfies

$$A_\lambda^* = \operatorname*{arg\,min}_{A \in \mathrm{S}_2(\mathscr{H})} \mathbb{E}[\|\varphi(Y) - A^*\varphi(X)\|_{\mathscr{H}}^2] + \lambda \|A\|_{\mathrm{S}_2(\mathscr{H})}^2$$

admits a closed form representation in terms of covariance operators associated with the kernel k. In fact, we prove that A_λ^* has the known form which Song et al. (2009) originally identified as the conditional mean embedding under the previously mentioned restrictive assumptions.

While this result does not come as a surprise at this point, we emphasize that this has not been proven before. Although Grünewälder et al. (2012) establish a connection between the *empirical* regularized solution $F_{\lambda,\mathbf{z}}$ and a version of the *empirical* conditional mean embedding with a *rescaled regularization parameter*, a population analogue was never derived. A simple asymptotic argument via convergence in the infinite-data limit is hampered by the rescaling of the regularization parameter in this derivation. Interestingly, the population expression of A_λ which we derive here is sometimes taken for granted in the literature (see for example Fukumizu et al. 2013), even if it was never proven in the original work.

Our analysis offers a view on the beautiful duality between the generalized covariance operator T acting on $\mathscr{G}$, composition operators acting on $\mathrm{S}_2(\mathscr{H})$ and the kernel covariance operator C_{XX}.

Remark 7.2. While our analysis is purely aimed at a theoretical understanding at this point, we expect that the following results will have a practical benefit, as they allow an asymptotic discussion of the spectral properties of the given estimates (see also Section 8).

For an operator $B \in \mathfrak{B}(\mathscr{H})$, define the *right-composition operator*

$$\Xi_B \colon \mathrm{S}_2(\mathscr{H}) \to \mathrm{S}_2(\mathscr{H}), \tag{7.3}$$

$$A \mapsto AB. \tag{7.4}$$

It is easy to see that Ξ_B is a well-defined bounded operator since $\mathrm{S}_2(\mathscr{H})$ is an ideal in $\mathfrak{B}(\mathscr{H})$ and we have $\|\Xi_B A\|_{\mathrm{S}_2(\mathscr{H})} \leq \|A\|_{\mathrm{S}_2(\mathscr{H})} \|B\|$. Furthermore, if B is invertible then Ξ_B is invertible and we have $\Xi_{B^{-1}} = \Xi_B^{-1}$.

The following result describes the connection between $\mathscr{G}$ and C_{XX} in terms of the composition operator $\Xi_{C_{XX}}$. In fact, it shows that $T\colon \mathscr{G} \to \mathscr{G}$ describes exactly the action of $\Xi_{C_{XX}} : \mathrm{S}_2(\mathscr{H}) \to \mathrm{S}_2(\mathscr{H})$ under the isomorphism $\Theta\colon \mathrm{S}_2(\mathscr{H}) \to \mathscr{G}$.

$$
\begin{tikzcd}
\mathscr{G} \arrow[r, "\Theta^{-1}"] \arrow[d, "T"'] & \mathrm{S}_2(\mathscr{H}) \arrow[d, "\Xi_{C_{XX}}"] \\
\mathscr{G} \arrow[r, "\Theta^{-1}"'] & \mathrm{S}_2(\mathscr{H})
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathscr{G} \arrow[r, "\Theta^{-1}"] \arrow[d, "T+\lambda \mathrm{Id}_{\mathscr{G}}"'] & \mathrm{S}_2(\mathscr{H}) \arrow[d, "\Xi_{C_{XX}+\lambda \mathrm{Id}_{\mathscr{H}}}"] \\
\mathscr{G} \arrow[r, "\Theta^{-1}"'] & \mathrm{S}_2(\mathscr{H})
\end{tikzcd}
$$

FIGURE 3. Correspondence of T and $\Xi_{C_{XX}}$.

Theorem 7.3. *Let $F \in \mathscr{G}$ and $A := \Theta^{-1}(F) \in \mathrm{S}_2(\mathscr{H})$. Then the diagrams in Figure 3 are both commutative diagrams, i.e., we have*

$$\Theta^{-1}(TF) = AC_{XX}$$

as well as

$$\Theta^{-1}[(T + \lambda \mathrm{Id}_{\mathscr{G}})F] = A(C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}}).$$

Proof. Let $F \in \mathscr{G}$ and $A = \Theta^{-1}(F) \in \mathrm{S}_2(\mathscr{H})$. We have $F(\cdot) = A\varphi(\cdot)$ by Corollary 4.5. From the definition of $\mathscr{G}$, we get

$$
\begin{aligned}
TF &= \int_E K_x F(x)\mathrm{d}\pi(x) = \int_E K_x[A\varphi(x)]\mathrm{d}\pi(x) \\
&= \int_E A[k(\cdot, x)\varphi(x)]\mathrm{d}\pi(x) = A\int_E k(\cdot, x)\varphi(x)\mathrm{d}\pi(x) \\
&= A\int_E [\varphi(x) \otimes \varphi(x)]\varphi(\cdot)\mathrm{d}\pi(x) = AC_{XX}\varphi(\cdot),
\end{aligned}
$$

where we use the fact that for every fixed $x' \in E$, the map $x \mapsto k(x', x)\varphi(x)$ is an element of $L^1(E, \mathcal{F}_E, \pi; \mathscr{H})$ due to Assumption 3 and Hölder's inequality. Because of this, the integration and the operator A commute (Diestel and Uhl, 1977, Chapter II.2, Theorem 6). The operator AC_{XX} is Hilbert–Schmidt and $TF = AC_{XX}\varphi(\cdot)$ confirms the operator reproducing property under Θ^{-1} from Corollary 4.5, hence we have $\Theta^{-1}(TF) = AC_{XX}$. Using this fact, we obtain

$$(T + \lambda \mathrm{Id}_{\mathscr{G}})F = AC_{XX}\varphi(\cdot) + \lambda A\varphi(\cdot) = A(C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})\varphi(\cdot),$$

confirming the same relation for the second assertion of the theorem. ■

Theorem 7.3 allows us to easily derive the expression for the Tikhonov–Phillips estimate F_λ under Θ^{-1} in terms of its corresponding operator in $\mathrm{S}_2(\mathscr{H})$ in terms of C_{XX} and C_{YX}.

Corollary 7.4 (Closed form analytical operator solution)**.** *We have*

$$\Theta^{-1}(F_\lambda) = A_\lambda^* = C_{YX}(C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1},$$

i.e., the analytical regularized operator solution can be represented as

$$\Theta^{-1}(F_\lambda)^* = A_\lambda = (C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1}C_{XY}. \tag{7.5}$$

Proof. By definition, we have $F_\lambda = g_\lambda(T)\mathcal{I}_\pi^* F_p = (T + \lambda \mathrm{Id}_{\mathscr{G}})^{-1}\mathcal{I}_\pi^* F_p$. We can rearrange

$$\begin{aligned}\mathcal{I}_\pi^* F_p &= \int_E K(\cdot, x) F_p(x) \mathrm{d}\pi(x) = \int_E k(\cdot, x) \int_E \varphi(y) p(x, \mathrm{d}y) \mathrm{d}\pi(x) \\ &= \iint_{E^2} \varphi(y) \left\langle \varphi(x), \varphi(\cdot) \right\rangle_{\mathscr{H}} p(x, \mathrm{d}y) \mathrm{d}\pi(x) \\ &= \left[\int_E \varphi(Y) \otimes \varphi(X) \mathrm{d}\mathbb{P} \right] \varphi(\cdot) = C_{YX} \varphi(\cdot).\end{aligned}$$

We have thus shown that $C_{YX} = \Theta^{-1}(\mathcal{I}_\pi^* F_p)$ by the operator reproducing property from Corollary 4.5. Theorem 7.3 implies that the operator $(T + \lambda \mathrm{Id}_{\mathscr{G}})^{-1}$ acting on $\mathscr{G}$ may be represented under Θ^{-1} as by the right composition operator $\Xi_{(C_{XX}+\lambda \mathrm{Id}_{\mathscr{H}})^{-1}}$ acting on on $\mathrm{S}_2(\mathscr{H})$, leading to

$$\Theta^{-1}(F_\lambda) = \Xi_{(C_{XX}+\lambda \mathrm{Id}_{\mathscr{H}})^{-1}} C_{YX} = C_{YX}(C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1}$$

as claimed. ■

Analogously we obtain a closed form representation for the empirical regularized operator solution $A_{\lambda,\mathbf{z}}$, in terms of the emprical covariance operators

$$\widehat{C}_{XX} := \frac{1}{n} \sum_{i=1}^n \varphi(X_i) \otimes \varphi(X_i) \text{ and } \widehat{C}_{XY} := \frac{1}{n} \sum_{i=1}^n \varphi(Y_i) \otimes \varphi(X_i).$$

Theorem 7.5 (Closed form empirical operator solution)**.** *We have*

$$\Theta^{-1}(F_{\lambda,\mathbf{z}}) = A_{\lambda,\mathbf{z}}^* = \widehat{C}_{YX}(\widehat{C}_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1},$$

i.e., the empirical regularized operator solution can be represented as

$$\Theta^{-1}(F_{\lambda,\mathbf{z}})^* = A_{\lambda,\mathbf{z}} = (\widehat{C}_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1} \widehat{C}_{XY}. \tag{7.6}$$

Theorem 7.5 can be proven by simply replacing T with the sample-based operator $T_{\mathbf{x}}$ in the proof of Theorem 7.3, leading to $\Theta^{-1}[(T_{\mathbf{x}} + \lambda \mathrm{Id}_{\mathscr{G}})F] = \Theta^{-1}(F)(\widehat{C}_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})$ for all $F \in \mathscr{G}$. Furthermore replacing $\mathcal{I}_\pi^*$ with $S_{\mathbf{x}}^*$ in the proof of Corollary 7.4 yields $\Theta^{-1}(S_{\mathbf{x}}^* \Upsilon) = \widehat{C}_{YX}$, thereby confirming the claim when applying both results to $A_{\lambda,\mathbf{z}} = \Theta^{-1}(F_{\lambda,\mathbf{z}}) = \Theta^{-1}[(T_{\mathbf{x}} + \lambda \mathrm{Id}_{\mathscr{G}})^{-1} S_{\mathbf{x}}^* \Upsilon]$.

8. Application: kernel-based extended dynamic mode decomposition

The derivation of the closed form for the regularized operator solution from the previous section allows to connect our theory to known spectral analysis techniques used in practice.

Klus et al. (2020) and Mollenhauer et al. (2020b) show that the eigenfunctions of the regularized empirical estimate $A_{\lambda,\mathbf{z}} = (\widehat{C}_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1} \widehat{C}_{XY}$ can be computed by solving a matrix eigenproblem. In the case that P is the Markov transition operator from (1.1), it is furthermore shown by Klus et al. (2020) that this empirical eigenproblem coincides exactly with the regularized eigenproblem given by the well-known kernel-based version of *extended dynamic mode decomposition* (EDMD, Tu et al. 2014; Williams et al. 2015a,b; Kutz et al. 2016). Hence, the asymptotic viewpoint derived in our analysis proves that kernel EDMD essentially approximates $P : \mathscr{H} \to L^2(\pi)$ in the infinite-sample limit with a suitable regularization scheme, thereby providing a statistical model for kernel EDMD. A theory of the spectral convergence of kernel EDMD could now be developed by investigating the spectral perturbation under the convergence $\|A_{\lambda,\mathbf{z}} - P\|_{\mathscr{H} \to L^2(\pi)} \to 0$ for an admissible regularization scheme $\lambda(n)$ and $n \to \infty$ with suitable mixing assumptions of the underlying process along the lines of Mollenhauer et al. (2020a). In particular, our approximation results from Section 5 may be used to show that kernel EDMD overcomes the weak spectral convergence of standard EDMD which was proven by Korda and Mezić (2018). The details of this theory are not in the scope of this work and are subject to further research.

9. Outlook

This work provides the theoretical framework for the nonparametric approximation of the conditional expectation operator P over the RKHS embedded in its domain $\mathscr{H} \subset L^2(\nu)$ from an approximation viewpoint. As a core result, we prove that convergence takes place in the operator norm with respect to the RKHS $\mathscr{H}$, therefore allowing for a stronger mode of convergence than classically used numerical projection methods.

We establish the connection to recent topics in statistical learning theory, in particular least squares regression problems with vector-valued kernels and the maximum mean discrepancy. These connections may allow to extend our theory to practical applications such as nonparametric hypothesis tests for Markov kernels.

Although a large part of computational questions can be answered in terms of inverse problems and regularization theory, our work shows that the derivation of uniform convergence rates needs new theoretical approaches in this context. In particular, the inverse problem associated with the estimation of P violates typical assumptions used in statistical learning theory (see Remark 6.2).

In the case that P is a Markov transition operator, our analysis provides a statistical model for kernel-based EDMD. However, in this case there remain open questions from a theoretical perspective. In particular,

(i) convergence behaviour of the estimators need to be derived in terms of properties of the underlying Markov process such as a spectral gap, ergodicity rates and mixing;

(ii) a spectral analysis of the estimators is needed in the context of classical perturbation theory in order to understand details of the spectral convergence.

Acknowledgements

We greatfully acknowledge funding from Germany's Excellence Strategy (MATH+: The Berlin Mathematics Research Center, EXC-2046/1, project ID:390685689, project AA1-2).

References

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

C. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52 – 72, 2007. doi: https://doi.org/10.1016/j.jco.2006.07.001. URL http://www.sciencedirect.com/science/article/pii/S0885064X06000781.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.*, 45(6):2610–2636, 2007. doi: ://doi.org/10.1137/060651884. URL https://doi.org/10.1137/060651884.

G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.*, 18:971–1013, 2018. doi: 10.1007/s10208-017-9359-7. URL https://doi.org/10.1007/s10208-017-9359-7.

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

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 04(04):377–408, 2006. doi: 10.1142/S0219530506000838. URL https://doi.org/10.1142/S0219530506000838.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 08(01):19–61, 2010. doi: 10.1142/S0219530510001503. URL https://doi.org/10.1142/S0219530510001503.

L. Cavalier. Inverse problems with non-compact operators. *Journal of Statistical Planning and Inference*, 136(2):390 – 400, 2006. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2004.06.063. URL http://www.sciencedirect.com/science/article/pii/S0378375804003039.

F. Chatelin. *Spectral approximation of linear operators*. Academic Press, New York,, 1983.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

F. Cucker and D.-X. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007. doi: 10.1017/CBO9780511618796. URL https://doi.org/10.1017/CBO9780511618796.

E. B. Davies. Metastable states of symmetric Markov semigroups I. *Proceedings of the London Mathematical Society*, 3(1):133–150, 1982a.

E. B. Davies. Metastable states of symmetric Markov semigroups II. *Journal of the London Mathematical Society*, 2(3):541–556, 1982b.

E. B. Davies. Spectral properties of metastable Markov semigroups. *Journal of functional analysis*, 52(3):315–329, 1983.

E. De Vito and A. Caponnetto. Risk bounds for the regularized least-squares algorithm with operator-valued kernels. Technical report, Massachusetts Institute of Technology, Cambridge, May 2005. CBCL Paper #249/AI Memo #2005-015.

E. De Vito, L. Rosasco, A. Caponetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:

883–904, 2005.

M. Dellnitz and O. Junge. On the approximation of complicated dynamical behavior. *SIAM Journal on Numerical Analysis*, 36(2):491–515, 1999.

J. Diestel and J. Uhl. *Vector Measures*. American Mathematical Society, 1977.

J. Ding and T.-Y. Li. Markov finite approximation of the Frobenius-Perron operator. *Nonlin. Anal., Theory, Meth. & Appl.*, 17:759–772, 1991.

R. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

N. Dunford and J. T. Schwartz. *Linear operators. Part I*. Wiley Classics Library. John Wiley & Sons, Inc., New York, 1988a. ISBN 0-471-60848-3. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication.

N. Dunford and J. T. Schwartz. *Linear operators. Part II*. Wiley Classics Library. John Wiley & Sons, Inc., New York, 1988b. ISBN 0-471-60847-5. Spectral theory. Selfadjoint operators in Hilbert space, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1963 original, A Wiley-Interscience Publication.

H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012a.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012b. URL http://jmlr.org/papers/v13/gretton12a.html.

S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *International Conference on Machine Learing*, volume 5, 2012.

S. Grünewälder, G. Arthur, and J. Shawe-Taylor. Smooth operators. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1184–1192, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/grunewalder13.html.

W. Hackbusch. *Integral equations: theory and numerical treatment*. Birkhäuser, 1995.

W. Huisinga. *Metastability of Markovian systems: A transfer operator based approach in application to molecular dynamics*. PhD thesis, 2001.

W. Huisinga and B. Schmidt. Metastability and dominant eigenvalues of transfer operators. In *New algorithms for macromolecular simulation*, pages 167–182. Springer, 2006.

W. Huisinga, S. Meyn, and C. Schütte. Phase transitions and metastability in Markovian and molecular systems. *Ann. Appl. Probab.*, 14(1):419–458, 02 2004. doi: 10.1214/aoap/1075828057. URL https://doi.org/10.1214/aoap/1075828057.

O. Junge and P. Koltai. Discretization of the Frobenius–Perron operator using a sparse Haar tensor basis: The Sparse Ulam method. *SIAM Journal on Numerical Analysis*, 47:3464–3485, 2009.

T. Kato. *Perturbation Theory for Linear Operators*. Springer, Berlin, 1980.

Y. Kawahara. Dynamic mode decomposition with reproducing kernels for Koopman spectral analysis. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 911–919. Curran Associates, Inc., 2016.

I. Klebanov, I. Schuster, and T. J. Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583 – 606, 2020a. doi: 10.1137/19M1305069.

I. Klebanov, B. Sprungk, and T. Sullivan. The linear conditional expectation in Hilbert space. *arXiv preprint arXiv:2008.12070*, 2020b.

S. Klus, P. Koltai, and C. Schütte. On the numerical approximation of the Perron–Frobenius and Koopman operator. *Journal of Computational Dynamics*, 3(1): 51–79, 2016.

S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 2018. doi: 10.1007/s00332-017-9437-7.

S. Klus, I. Schuster, and K. Muandet. Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. *J. Nonlinear Sci.*, 30(1):283–315, 2020. ISSN 0938-8974. doi: 10.1007/s00332-019-09574-z. URL https://doi.org/10.1007/s00332-019-09574-z.

I. Kontoyiannis and S. Meyn. Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes. *Electron. J. Probab.*, 10:61–123, 2005. doi: 10.1214/EJP.v10-231. URL https://doi.org/10.1214/EJP.v10-231.

I. Kontoyiannis and S. Meyn. Approximating a diffusion by a finite-state hidden Markov model. *Stochastic Processes and their Applications*, 127(8):2482 – 2507, 2017. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2016.11.004. URL http://www.sciencedirect.com/science/article/pii/S0304414916302149.

I. Kontoyiannis and S. P. Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab.*, 13(1):304–362, 01 2003. doi: 10.1214/aoap/1042765670. URL https://doi.org/10.1214/aoap/1042765670.

M. Korda and I. Mezić. On convergence of Extended Dynamic Mode Decomposition to the Koopman operator. *J. Nonlinear Sci.*, 28:687–710, 2018. doi: 10.1007/s00332-017-9423-0. URL https://doi.org/10.1007/s00332-017-9423-0.

J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. SIAM, 2016.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer, 1991.

T.-Y. Li. Finite approximation for the Frobenius-Perron operator. A solution to Ulam's conjecture. *J. Approx. Theory*, 17:177–186, 1976.

J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Appl. Comput. Harmon. Anal.*, 48(3):868–890, 2020. ISSN 1063-5203. doi: 10.1016/j.acha.2018.09.009. URL https://doi.org/10.1016/j.acha.2018.09.009.

S. Lu, P. Mathé, and S. V. Pereverzev. Balancing principle in supervised learning for a general regularization scheme. *Applied and Computational Harmonic Analysis*, 48(1):123–148, 2020.

M. Mollenhauer, S. Klus, C. Schütte, and P. Koltai. Kernel autocovariance operators of stationary processes: Estimation and convergence. *arXiv preprint arXiv:2004.00891*, 2020a.

M. Mollenhauer, I. Schuster, S. Klus, and C. Schütte. Singular value decomposition of operators on reproducing kernel hilbert spaces. In O. Junge, O. Schütze, G. Froyland, S. Ober-Blobaum, and K. Padberg-Gehle, editors, *Advances om Dynamics, Optimization and Computation. Series: Studies in Systems, Decision and Control. A volume dedicated to Michael Dellnitz on his 60th birthday*, volume 304, pages 109 – 131. 2020b. ISBN 978-3-030-51264-4.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141, 2017.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. URL http://www.jstor.org/stable/1428011.

H. Owhadi and C. Scovel. Separability of reproducing kernel spaces. *Proc. Amer. Math. Soc.*, 145:2131–2138, 2017.

J. Park and K. Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020a.

J. Park and K. Muandet. Regularised least-squares regression with infinite-dimensional output space. *arXiv preprint arXiv:2010.10973*, 2020b.

D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20:32 pp., 2015. doi: 10.1214/EJP.v20-4039. URL https://doi.org/10.1214/EJP.v20-4039.

D. L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *J. ACM*, 9(1):84–97, Jan. 1962. ISSN 0004-5411. doi: 10.1145/321105.321114. URL https://doi.org/10.1145/321105.321114.

A. Rastogi and S. Sampath. Optimal rates for the regularized learning algorithms under general source condition. *Frontiers in Applied Mathematics and Statistics*, 3:3, 2017. ISSN 2297-4687. doi: 10.3389/fams.2017.00003. URL https://www.frontiersin.org/article/10.3389/fams.2017.00003.

A. Rastogi, G. Blanchard, and P. Mathé. Convergence analysis of Tikhonov regularization for non-linear statistical inverse problems. *Electron. J. Statist.*, 14(2):2798–2841, 2020. doi: 10.1214/20-EJS1735. URL https://doi.org/10.1214/20-EJS1735.

G. Roberts, J. Rosenthal, et al. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997.

G. O. Roberts and R. L. Tweedie. Geometric L2 and L1 convergence are equivalent for reversible Markov chains. *Journal of Applied Probability*, pages 37–41, 2001.

L. Rosasco, M. Belkin, and E. D. Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

S. Saitoh and Y. Sawano. *Theory of Reproducing Kernels and Applications*. Springer, 2016.

P. J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010. doi: 10.1017/S0022112010001217.

C. Schütte and M. Sarich. *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches*. Number 24 in Courant Lecture Notes. American Mathematical Society, 2013.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013. URL http://www.jstor.org/stable/23566550.

C.-J. Simon-Gabriel, A. Barp, and L. Mackey. Metrizing weak convergence with maximum mean discrepancies. *arXiv preprint arXiv:2006.09268*, 2020.

S. Smale and D.-X. Zhou. Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285 – 302, 2005. doi: 10.1016/j.acha.2005.03.001. URL http://www.sciencedirect.com/science/article/pii/S1063520305000540.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, 2007.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009. doi: 10.1145/1553374.1553497.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 773–780, 13–15 May 2010. URL http://proceedings.mlr.press/v9/sriperumbudur10a.html.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011. URL http://jmlr.org/papers/v12/sriperumbudur11a.html.

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

Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016. URL http://jmlr.org/papers/v17/14-510.html.

W. Tian and H. Wu. Kernel embedding based variational approach for low-dimensional approximation of dynamical systems. *arXiv preprint arXiv:2008.02962*, 2020.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill Posed Problems*. W. H. Winston, 1977.

J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, 1(2), 2014.

J. Weidmann. *Linear Operators in Hilbert Spaces*. Springer, 1980.

M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015a.

M. O. Williams, C. W. Rowley, and I. G. Kevrekidis. A kernel-based method for data-driven Koopman spectral analysis. *Journal of Computational Dynamics*, 2 (2):247–265, 2015b.

Y. Yao, L. Rosasco, and A. Caponnetto. On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, 26:289–315, 2007. doi: 10.1007/s00365-006-0663-2. URL https://doi.org/10.1007/s00365-006-0663-2.

Appendix A.

It is well-known that for $F \in L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$, the standard least squares risk

$$R(F) := \mathbb{E}\left[\|\varphi(Y) - F(X)\|^2_{\mathscr{H}}\right],$$

can be rewritten in terms of the *regression function* F_p. We report the proof here for completeness (see also Cucker and Smale, 2002, Proposition 1 for a proof in the scalar case).

Theorem A.1 (Risk and regression function)**.** *Under Assumptions 1–3, the risk R can equivalently be rewritten as*

$$R(F) = \|F - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} + R(F_p) \tag{A.1}$$

for all $F \in L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$.

Proof. We have

$$\begin{aligned}
R(F) &= \mathbb{E}\left[\|\varphi(Y) - F(X)\|^2_{\mathscr{H}}\right] \\
&= \mathbb{E}\left[\|\varphi(Y) - F_p(X) + F_p(X) - F(X)\|^2_{\mathscr{H}}\right] \\
&= \mathbb{E}\left[\|\varphi(Y) - F_p(X)\|^2_{\mathscr{H}}\right] \\
&+ 2\mathbb{E}\left[\langle \varphi(Y) - F_p(X),\, F_p(X) - F(X)\rangle_{\mathscr{H}}\right] \\
&+ \mathbb{E}\left[\|F(X) - F_p(X)\|^2_{\mathscr{H}}\right],
\end{aligned}$$

where we see that the first summand equals to $R(F_p)$. The second summand which contains the mixed terms vanishes since we have

$$\begin{aligned}
&\mathbb{E}\left[\langle \varphi(Y) - F_p(X),\, F_p(X) - F(X)\rangle_{\mathscr{H}}\right] \\
= &\int_E \left\langle \underbrace{\int_E \varphi(y)\, p(x, \mathrm{d}y)}_{=F_p(x)} - F_p(x),\, F_p(x) - F(x) \right\rangle_{\mathscr{H}} \mathrm{d}\pi(x).
\end{aligned}$$

The last summand can be rewritten as

$$\mathbb{E}\left[\|F(X) - F_p(X)\|^2_{\mathscr{H}}\right] = \|F - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}$$

by change of measure, proving the assertion. ■

Freie Universität Berlin

Email address: mattes.mollenhauer@fu-berlin.de

Freie Universität Berlin

Email address: peter.koltai@fu-berlin.de