Journal of Machine Learning Research 25 (2024) 1-35 Submitted 2/23; Revised 7/24; Published 8/24

Recursive Estimation of Conditional Kernel Mean Embeddings

Ambrus Tamás[1,2] AMBRUS.TAMAS@SZTAKI.HU
Balázs Csanád Csáji[1,2] BALAZS.CSAJI@SZTAKI.HU

[1] *Institute for Computer Science and Control (SZTAKI)*
Hungarian Research Network (HUN-REN),
Kende utca 13-17, Budapest, H-1111, Hungary

[2] *Department of Probability Theory and Statistics*
Institute of Mathematics, Eötvös Loránd University (ELTE)
Pázmány Péter Sétány 1 / C, Budapest, H-1117, Hungary

Editor: Kenji Fukumizu

Abstract

Kernel mean embeddings, a widely used technique in machine learning, map probability distributions to elements of a reproducing kernel Hilbert space (RKHS). For supervised learning problems, where input-output pairs are observed, the conditional distribution of outputs given the inputs is a key object. The input dependent conditional distribution of an output can be encoded with an RKHS valued function, the conditional kernel mean map. In this paper we present a new recursive algorithm to estimate the conditional kernel mean map in a Hilbert space valued L_2 space, that is in a Bochner space. We prove the weak and strong L_2 consistency of our recursive estimator under mild conditions. The idea is to generalize Stone's theorem for Hilbert space valued regression in a locally compact Polish space. We present new insights about conditional kernel mean embeddings and give strong asymptotic bounds regarding the convergence of the proposed recursive method. Finally, the results are demonstrated on three application domains: for inputs coming from Euclidean spaces, Riemannian manifolds and locally compact subsets of function spaces.

Keywords: conditional kernel mean embeddings, recursive estimation, reproducing kernel Hilbert spaces, strong consistency, nonparametric inference, Riemannian manifolds

1. Introduction

In *statistical learning* we need to work with probability distributions defined on a wide variety of objects, which may not even come from a linear space. *Kernel mean embeddings* (KMEs) offer a neat representation for dealing with distributions by mapping them into a *reproducing kernel Hilbert space* (RKHS). These embeddings were defined in (Berlinet and Thomas-Agnan, 2004) and (Smola et al., 2007). In machine learning often the conditional distribution of an output w.r.t. some input plays a key role. In these problems the concept of *conditional kernel mean embeddings* (CKME) is more adequate. The notion of CKME was first recognized in (Song et al., 2009), then a refined definition was given in (Song et al., 2013). Its rigorous measure theoretic background was presented in (Klebanov et al.,

2020) and in (Park and Muandet, 2020) as an input dependent random element in an RKHS. These concepts strongly rely on the notions of Bochner integral and conditional expectation in Banach spaces of which theory is presented in (Pisier, 2016; Hytönen et al., 2016) and Appendix A. Under some stringent assumptions Song et al. (2009) estimate the conditional kernel mean map with an empirical variant of the regularized inverse cross-covariance operator. It is shown in (Grünewälder et al., 2012) that this estimate can also be deduced from a regularized risk minimization scheme in a *vector-valued* RKHS based on (Micchelli and Pontil, 2005). In (Park and Muandet, 2020) a consistency theorem is proved when the conditional kernel mean map is included in a vector-valued RKHS.

Related works In this paper we follow the general *distribution-free* framework presented in (Györfi et al., 2002). It is recognized that estimating the conditional kernel mean map is a particular *vector-valued regression* problem. Vector-valued regression was already studied in (Ramsay and Silverman, 2005; Bosq and Delecroix, 1985; Dabo-Niang and Rhomari, 2009; Forzani et al., 2012; Brogat-Motte et al., 2022). Nevertheless, it is still challenging to provide sufficient conditions for strong L_2 (mean square) consistency that can be applied in practice. We also build on the theory of *stochastic approximation* (Kushner and Yin, 2003; Ljung et al., 2012). Our core algorithm uses a stochastic update in each iteration. The presented method was motivated by the celebrated *stochastic gradient descent* algorithm, see for example (Bertsekas and Tsitsiklis, 1996, Proposition 4.2). The update term of our recursive algorithm is a modified gradient of an *empirical surrogate loss* function.

Contributions Our main contribution is a *weakly and strongly consistent recursive algorithm* for estimating the conditional kernel mean map under general structural assumptions. We present and prove a *generalized version of Stone's theorem* (Stone, 1977), motivated by the recursive form in (Györfi et al., 2002, Theorem 25.1). Our generalization is twofold. We deal with general *locally compact Polish input spaces* and allow the output space to be *Hilbertian*. Our main assumption is that the *measure of metric balls* goes to zero slowly w.r.t. the radius. In (Forzani et al., 2012) a similar consistency result is proved, however the authors of that paper do not deal with functional outputs and also require the so-called Besicovitch condition to hold on the regression function. In (Dabo-Niang and Rhomari, 2009) sufficient conditions for L_2 consistency are given for vector-valued regression, however, in this paper a differentiation condition is required for the regression function and also the rate of the small ball probabilities is bounded from below when the radius goes to zero.

First, in Section 2 we present a recursive scheme for (unconditional) kernel mean embeddings. A stochastic approximation based theorem (Ljung et al., 2012) is used to prove its strong consistency. Then, in Section 3 we introduce a recursive estimator for the conditional kernel mean map. Our algorithm is formulated in a flexible offline manner in such a way that the online algorithm is also covered. We prove the weak L_2 consistency for the general *local averaging* scheme in locally compact Polish spaces under very mild assumptions. We apply this consistency result to prove the strong consistency of local averaging kernel estimates with a (measurable) nonnegative, bounded and symmetric *smoother* sequence. This scheme requires a smoother sequence with appropriate shrinking properties w.r.t. the unknown marginal measure. We satisfy these requirements by a general mother smoother function induced sequence parameterized by two shrinkage rates. In the final theorem only a *structural condition* on the probability of small balls remains. Finally, in Section 4 we

apply our results for three arch-typical cases. We deduce universal consistency for *Euclidean spaces*, *Riemannian manifolds* and locally compact subsets of *Hilbert spaces*.

Applications Here, we highlight some potential application domains of the ideas. Based on the theory of KMEs and CKMEs, a rich variety of applications were introduced. Kernel mean embeddings can be efficiently used, for example, in graphical models (Song et al., 2013), Bayesian inference (Fukumizu et al., 2013), dynamical systems (Song et al., 2009), reinforcement learning (Grünewälder et al., 2012), causal inference (Schölkopf et al., 2015) and hypothesis testing for independence (Gretton et al., 2008; Gretton and Györfi, 2008; Gretton et al., 2012), conditional independence (Fukumizu et al., 2007; Zhang et al., 2011), and binary classification (Csáji and Tamás, 2019) and (Tamás and Csáji, 2022).

Recursive algorithms are crucial to handle data streams or fixed, but very large datasets, cf. divide-and-conquer. Iterative methods play a key role in, e.g., reinforcement learning, system identification, signal processing and adaptive control. Therefore, developing recursive estimators for conditional kernel mean embeddings is beneficial for many domains.

2. Kernel Mean Embeddings of Distributions

Kernel methods offer an elegant approach to deal with abstract sample spaces. One only needs to specify a *similarity measure* on the space via the *kernel* function to provide a geometrical structure for the learning problem. In fact, every symmetric, positive definite function, a.k.a. kernel, defines a reproducing kernel Hilbert space. The user-chosen kernel function also determines a natural projection from the sample space into the RKHS. The projected values of sample points are called feature vectors, or *features*. This projection map can be easily extended to (probability) measures defining kernel mean embeddings. In this section we present the notion of KMEs in a statistical learning framework.

2.1 Statistical Framework

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, where Ω is the sample space, $\mathcal{A}$ is the σ-algebra of events, and $\mathbb{P}$ is the probability measure. We consider a random pair (X, Y) from a measurable product space $\mathbb{X} \times \mathbb{Y}$ with some product σ-algebra $\mathcal{X} \otimes \mathcal{Y}$. Let us consider a metric on $\mathbb{X}$ and let us denote its induced Borel σ-algebra as $\mathcal{X}$. We denote the joint distribution of (X, Y) by Q (or $Q_{X,Y}$) and the corresponding marginal distributions by Q_X and Q_Y. In practice, distribution Q is usually *unknown*, however we assume that:

A1 *An independent identically distributed* (*i.i.d.*) *sample* $\{(X_i, Y_i)\}_{i=1}^n$ *is given.*

In the next section we present the definition of kernel mean embeddings of probability distributions into an RKHS associated with kernel $\ell : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$. With a somewhat imprecise terminology, we also talk about the kernel mean embedding of a random variable into an RKHS, by which we mean taking the KME of its distribution. Explanatory variable X will only be used for the *conditional* variant of KME to generate the conditioning σ-algebra, consequently, for the *unconditional* KME, we will only use one variable, Y.

In the subsequent section we focus on conditional kernel mean embeddings. The conditional kernel mean embedding of Y with respect to X into a given RKHS is the conditional expectation of the random feature defined by the kernel and variable Y with respect to the

σ-algebra generated by X, i.e., a conditional kernel mean embedding is a random feature vector in an RKHS measurable to some explanatory variable. The main object for CKME is the *conditional kernel mean map* which maps the explanatory points into RKHS $\mathcal{G}$. This function maps (almost) every input point to the kernel mean embedding of the conditional distribution given that input. For a linear kernel one can recover the regression function as a particular conditional kernel mean map. When X and Y are independent, the conditional kernel map is a constant function mapping every input into the unconditional KME of Y.

2.2 Reproducing Kernel Hilbert Spaces

Let ℓ be a (real-valued, symmetric, positive definite) kernel on $\mathbb{Y} \times \mathbb{Y}$, where $\mathbb{Y}$ is a nonempty set, i.e., for all $n \in \mathbb{N} \doteq \{1, 2, \dots\}$, $y_1, \dots, y_n \in \mathbb{Y}$ and $a_1, \dots, a_n \in \mathbb{R}$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \, \ell(y_i, y_j) \;\geq\; 0, \tag{1}$$

or equivalently kernel matrix $L \in \mathbb{R}^{n \times n}$, where $L_{i,j} \doteq \ell(y_i, y_j)$ for $i, j \in [n] \doteq \{1, \dots, n\}$, is required to be positive semidefinite. A Hilbert space of functions, $\mathcal{G} = \{g : \mathbb{Y} \to \mathbb{R}\}$, is called a *reproducing kernel Hilbert space* if every point evaluating (linear) functional, $\delta_y : \mathcal{G} \to \mathbb{R}$ defined by $\delta_y(g) = g(y)$ for $y \in \mathbb{Y}$, is bounded (or equivalently continuous). In (Aronszajn, 1950) it is proved that for every (symmetric and positive definite) kernel ℓ there uniquely (up to isometry) exists an RKHS $\mathcal{G}$, such that $\ell(\cdot, y) \in \mathcal{G}$ and

$$\langle\, g, \ell(\cdot, y) \rangle_{\mathcal{G}} \;=\; g(y) \tag{2}$$

holds for all $g \in \mathcal{G}$ and $y \in \mathbb{Y}$. In particular, $\langle \ell(\cdot, y_1), \ell(\cdot, y_2) \rangle_{\mathcal{G}} = \ell(y_1, y_2)$ is true. In the reverse direction, by the Riesz representation theorem (Lax, 2002, Theorem 4), for every RKHS $\mathcal{G}$ there uniquely exists a (symmetric and positive definite) kernel $\ell : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$ such that (2) holds for all $g \in \mathcal{G}$ and $y \in \mathbb{Y}$. Function ℓ is called the *reproducing kernel* of space $\mathcal{G}$. For further details, see (Schölkopf and Smola, 2002; Berlinet and Thomas-Agnan, 2004; Shawe-Taylor and Cristianini, 2004; Steinwart and Christmann, 2008).

2.3 Kernel Mean Embeddings of Marginal Distributions

In this paper we will assume that space $\mathcal{G}$ is *separable*. One can observe that if $\mathbb{Y}$ is a separable topological space and kernel ℓ is continuous, then the separability of $\mathcal{G}$ follows immediately, see (Steinwart and Christmann, 2008, Lemma 4.33). Specifically, we assume

B1 *Kernel ℓ is measurable, $\mathbb{E}[\sqrt{\ell(Y,Y)}\,] < \infty$, and RKHS $\mathcal{G}$ is separable.*

In this case, by the theorem of Pettis (Hytönen et al., 2016, Theorem 1.1.20), $\ell(\cdot, Y) : \Omega \to \mathcal{G}$ is also measurable. Then, the kernel mean embedding of μ_Y exists if and only if $B1$ holds (Hytönen et al., 2016, Proposition 1.2.2). Thus, one can embed Q_Y into RKHS $\mathcal{G}$ by

$$\mu_Y \doteq \mathbb{E}[\, \ell(\cdot, Y)\,], \tag{3}$$

where the expectation is a *Bochner integral*. It is known (Smola et al., 2007) that if $B1$ holds, then by the Riesz representation theorem $\mu_Y \in \mathcal{G}$ and for all $g \in \mathcal{G}$, we have

$$\langle \mu_Y, g \rangle_{\mathcal{G}} \;=\; \mathbb{E}[\, g(Y)\,], \tag{4}$$

i.e., μ_Y is the representer of the *expectation functional* on $\mathcal{G}$ w.r.t. Q_Y.

Given an i.i.d. sample $\{Y_i\}_{i=1}^n$, having distribution Q_Y, we can estimate the kernel mean embedding of Q_Y with the empirical average (a.k.a. sample mean), that is

$$\widehat{\mu}_n \doteq \frac{1}{n}\sum_{i=1}^{n} \ell(\cdot, Y_i). \tag{5}$$

By the strong law of large numbers, $\widehat{\mu}_n \to \mu_Y$ almost surely (Taylor, 1978). We refer to (Tolstikhin et al., 2017, Proposition A.1.) to quantify the convergence rate of this estimate:

Theorem 1 *Let $\{Y_i\}_{i=1}^n$ be an i.i.d. sample from distribution Q_Y on a separable topological space $\mathbb{Y}$. Assume that $\ell(\cdot, y)$ is continuous and there exists $C_\ell > 0$ such that $\ell(y, y) \leq C_\ell$ for all $y \in \mathbb{Y}$. Then, for all $\delta \in (0, 1)$ with probability at least $1 - \delta$, we have*

$$\|\mu_Y - \widehat{\mu}_n\|_{\mathcal{G}} \leq \sqrt{\frac{C_\ell}{n}} + \sqrt{\frac{2\, C_\ell \log\left(\frac{1}{\delta}\right)}{n}}. \tag{6}$$

It is also proved that this rate is minimax for a broad class of distributions. Note that only the measurability of $\ell(\cdot, y)$ is used in the proof of Tolstikhin et al. (2017).

2.4 Recursive Estimation of Kernel Mean Embeddings

As in the scalar case, these empirical estimates can be computed recursively by

$$\begin{aligned} \widehat{\mu}_1 &\doteq \ell(\cdot, Y_1), \quad \text{and} \\ \widehat{\mu}_n &\doteq \frac{n-1}{n}\,\widehat{\mu}_{n-1} + \frac{1}{n}\,\ell(\cdot, Y_n) = \widehat{\mu}_{n-1} + \frac{1}{n}\left(\ell(\cdot, Y_n) - \widehat{\mu}_{n-1}\right), \end{aligned} \tag{7}$$

for $2 \leq n \in \mathbb{N}$. More generally, one can use some possibly random stepsize (or learning rate) sequence $(a_n)_{n\in\mathbb{N}}$, taking values in $(0, 1)$, and apply the refined recursion defined by

$$\begin{aligned} \widehat{\mu}_1 &\doteq \ell(\cdot, Y_1), \quad \text{and} \\ \widehat{\mu}_n &\doteq \widehat{\mu}_{n-1} - a_n\left(\widehat{\mu}_{n-1} - \ell(\cdot, Y_n)\right), \end{aligned} \tag{8}$$

for $n \geq 2$. We prove the strong consistency of $(\widehat{\mu}_n)_{n\in\mathbb{N}}$ by applying Theorem 1.9 of Ljung et al. (2012), which is restated (for convenience) as Theorem 24 in Appendix D.

Theorem 2 *Let $Y_1, Y_2, \ldots$ be an i.i.d. sequence, $C_\ell > 0$ be a constant such that $\ell(y, y) \leq C_\ell$, for all $y \in \mathbb{Y}$, and assume B1. Let $(\widehat{\mu}_n)_{n\in\mathbb{N}}$ be the estimate sequence defined in (8). If $a_n \geq 0$, $\sum_n a_n^2 < \infty$ and $\sum_n a_n = \infty$ (a.s.), then $\widehat{\mu}_n \to \mu_Y$, as $n \to \infty$, almost surely.*

Proof It is sufficient to check the conditions of Theorem 24 for $\Lambda(g) \doteq g - \mu_Y$, $\mu \doteq \mu_Y$, $\mu_n \doteq \widehat{\mu}_n$ and $W_n \doteq \ell(\cdot, Y_n) - \mu_Y$. The conditions on sequence $(a_n)_{n\in\mathbb{N}}$ are assumed, hence we only need to verify (i), (ii) and (iii) of Theorem 24. Condition (i) holds with $c \doteq \max\left(\|\mu_Y\|_{\mathcal{G}}, 1\right)$. Condition (ii) is again straigthforward, because $\langle \Lambda(g), g - \mu_Y\rangle_{\mathcal{G}} = \|g - \mu_Y\|_{\mathcal{G}}^2$. For condition (iii) let $H_n \doteq 0$ and $V_n \doteq W_n$. Using the independence of $\{Y_i\}_{i=1}^n$ yields

$$\begin{aligned} &\mathbb{E}\big[V_n \,|\, \widehat{\mu}_1, V_1, \ldots, \widehat{\mu}_{n-1}, V_{n-1}\big] = \mathbb{E}\big[\ell(\cdot, Y_n) - \mu_Y \,|\, Y_1, \ldots, Y_{n-1}\big] \\ &= \mathbb{E}\big[\ell(\cdot, Y_n)\big] - \mu_Y = 0, \end{aligned} \tag{9}$$

hence $(V_n)_{n\in\mathbb{N}}$ is a martingale difference sequence. Finally

$$\begin{aligned}\sqrt{\mathbb{E}\Big[\|V_n\|_{\mathcal{G}}^2\Big]} &\leq \sqrt{\mathbb{E}\Big[\,\|\ell(\cdot,Y_n)\|_{\mathcal{G}}^2\,\Big]} + \sqrt{\mathbb{E}\Big[\,\|\mu_Y\|_{\mathcal{G}}^2\,\Big]} \\ &= \sqrt{\mathbb{E}\Big[\ell(Y_n,Y_n)\Big]} + \|\mu_Y\|_{\mathcal{G}} \leq \sqrt{C_\ell} + \|\mu_Y\|_{\mathcal{G}}\,,\end{aligned} \tag{10}$$

thus $\sum_n a_n^2 \mathbb{E}[\,\|V_n\|_{\mathcal{G}}^2\,] < \infty$. Consequently, Theorem 2 follows from Theorem 24. ■

3. Conditional Kernel Mean Embeddings

If the kernel mean embedding of Q_Y exists, then for a given σ-algebra one can take the conditional expectation of $\ell(\cdot, Y)$. The conditional kernel mean embedding of Y given X in RKHS $\mathcal{G}$ is an X measurable random element of $\mathcal{G}$ defined by

$$\mu_{Y|X} \doteq \mathbb{E}[\,\ell(\cdot,Y)\,|\,X\,]. \tag{11}$$

For more details on conditional expectations in Banach spaces, see (Pisier, 2016) and Appendix A. It is known that the conditional expectation is well-defined if $\mathbb{E}[\,\ell(\cdot,Y)\,]$ exists, thus we assume that $\mathbb{E}[\,\sqrt{\ell(Y,Y)}\,] < \infty$ as above. By (Kallenberg, 2002, Lemma 1.13) there exists a measurable map $\mu_* : \mathbb{X} \to \mathcal{G}$ such that $\mu_{Y|X} = \mu_*(X)$ (a.s.). One can use the pointwise form of μ_* defined Q_X-a.s. by $\mu_*(x) = \mathbb{E}[\,\ell(\cdot,Y)\,|\,X = x\,]$. This function plays a key role, and we call it the *conditional kernel mean map*. By equation (94) we have that

$$\langle \mu_{Y|X}, g\rangle_{\mathcal{G}} = \langle \mu_*(X), g\rangle_{\mathcal{G}} = \langle\, \mathbb{E}[\ell(\cdot,Y)\,|\,X], g\rangle_{\mathcal{G}} = \mathbb{E}\Big[\langle \ell(\cdot,Y), g\rangle_{\mathcal{G}}\,|\,X\Big] = \mathbb{E}[\,g(Y)\,|\,X\,] \tag{12}$$

holds for all $g \in \mathcal{G}$. In addition, by the law of total expectation we have $\mathbb{E}[\,\mu_{Y|X}\,] = \mu_Y$.

Recovering the regression function Let us consider the special case when $\mathbb{X} \doteq \mathbb{R}^d$, $\mathbb{Y} \doteq \mathbb{R}$ and $\ell(y_1, y_2) = y_1 \cdot y_2$. Then, the RKHS generated by ℓ can be simply identified with $\mathbb{R}$ by mapping $\ell(y, a) \doteq y \cdot a$ to a and the conditional kernel mean map takes the form of $(\mu_*(X))(y) = \mathbb{E}[\,Y \cdot y\,|\,X\,] = y\,\mathbb{E}[\,Y\,|\,X\,]$. Therefore, by linearity, μ_* is equivalent to the regression function, i.e., $\mu_*(X)$ is the linear function determined by $\mathbb{E}[Y\,|\,X]$.

Recovering a family of real-valued regression For a function $g \in \mathcal{G}$, the regression of $g(Y)$ given X is a real-valued regression problem for which the results of Györfi et al. (2002) can be applied. One of the main advantages of dealing with CMKEs is to produce solutions for every $g \in \mathcal{G}$ simultaneously. Moreover, the information about the conditional distribution of the outputs, given an input, can be used for various types of inference.

3.1 Universal Consistency

The main challenge addressed by the paper is to construct an estimate sequence $(\mu_n)_{n\in\mathbb{N}}$ for μ_* with strong stochastic guarantees. Since the probability distribution of the given sample is unknown, in general our estimate will not be equal to μ_* in finite steps. In order to measure the loss of an estimate sequence, one may choose from several error criteria. In

this paper we use the vector valued L_2 risk with respect to Q_X. The main reason leading to this choice of L_2 criterion is that conditional expectation can be defined as a projection in an L_2 space. We strengthen our assumption to ensure that the examined functions are included in the appropriate Bochner spaces. Appendix A contains an overview on Bochner spaces and Bochner integrals, where our notations are precisely defined.

A2 *Kernel ℓ is measurable, $\mathbb{E}[\,\ell(Y,Y)\,] < \infty$, and RKHS $\mathcal{G}$ is separable.*

Lemma 3 *If A2 holds, then $\mu_* \in L_2(\mathbb{X}, Q_X; \mathcal{G})$.*

Proof We can see that from *A2* it follows immediately that $\ell(\cdot, Y) \in L_2(\Omega, \mathbb{P}; \mathcal{G})$ because

$$\int_\Omega \|\ell(\cdot,Y)\|_{\mathcal{G}}^2 \,\mathrm{d}\mathbb{P} = \mathbb{E}[\,\ell(Y,Y)\,] < \infty. \tag{13}$$

Moreover by (93), (Hytönen et al., 2016, Prop. 2.6.31.), we have that

$$\begin{aligned} &\int_{\mathbb{X}} \|\mu_*(x)\|_{\mathcal{G}}^2 \,\mathrm{d}Q_X(x) = \mathbb{E}[\,\|\mu_*(X)\|_{\mathcal{G}}^2\,] = \mathbb{E}[\,\|\mathbb{E}[\ell(\cdot,Y)\mid X]\|_{\mathcal{G}}^2\,] \\ &\leq \mathbb{E}\big[\mathbb{E}[\|\ell(\cdot,Y)\|_{\mathcal{G}}^2 \mid X]\big] = \mathbb{E}[\ell(Y,Y)] < \infty, \end{aligned} \tag{14}$$

hence $\mu_* \in L_2(\mathbb{X}, Q_X; \mathcal{G})$. ■

For scalar-valued regression, if $\mathbb{E}[Y^2] < \infty$ holds, then we have that the conditional expectation is an orthogonal projection to $L_2(\Omega, \sigma(X), \mathbb{P}|_{\sigma(X)})$, where $\sigma(X)$ is the σ-algebra generated by X and $\mathbb{P}|_{\sigma(X)}$ is the restriction of $\mathbb{P}$ into $\sigma(X)$. Similarly:

Lemma 4 *Under A2, $\mu_*(X)$ is an orthogonal projection of $\ell(\cdot, Y)$ to $L_2(\Omega, \sigma(X), \mathbb{P}|_{\sigma(X)}; \mathcal{G})$.*

Proof For any $\mu \in L_2(\mathbb{X}, Q_X; \mathcal{G})$ it holds (Park and Muandet, 2020, Theorem 4.2) that

$$\begin{aligned} \mathcal{E}(\mu) &\doteq \mathbb{E}[\,\|\ell(\cdot,Y) - \mu(X)\|_{\mathcal{G}}^2\,] = \mathbb{E}[\,\|\ell(\cdot,Y) - \mu_*(X) + \mu_*(X) - \mu(X)\|_{\mathcal{G}}^2\,] \\ &= \mathbb{E}[\,\|\ell(\cdot,Y) - \mu_*(X)\|_{\mathcal{G}}^2\,] + \mathbb{E}[\,\|\mu_*(X) - \mu(X)\|_{\mathcal{G}}^2\,] \\ &\quad + 2\,\mathbb{E}\Big[\mathbb{E}\big[\langle \ell(\cdot,Y) - \mu_*(X), \mu_*(X) - \mu(X)\rangle_{\mathcal{G}} \mid X\big]\Big] \\ &= \mathbb{E}[\,\|\ell(\cdot,Y) - \mu_*(X)\|_{\mathcal{G}}^2\,] + \mathbb{E}[\,\|\mu_*(X) - \mu(X)\|_{\mathcal{G}}^2\,] \\ &\quad + 2\,\mathbb{E}\big[\langle \mathbb{E}[\ell(\cdot,Y)\mid X] - \mu_*(X), \mu_*(X) - \mu(X)\rangle_{\mathcal{G}}\big] \\ &= \mathbb{E}[\,\|\ell(\cdot,Y) - \mu_*(X)\|_{\mathcal{G}}^2\,] + \mathbb{E}[\,\|\mu_*(X) - \mu(X)\|_{\mathcal{G}}^2\,]. \end{aligned} \tag{15}$$

Thus, μ_* minimizes $\mathcal{E}$ in $L_2(\mathbb{X}, Q_X; \mathcal{G})$. ■

Note that $\mathbb{E}[\,\|\ell(\cdot,Y) - \mu_*(X)\|_{\mathcal{G}}^2\,]$ is independent of μ, therefore it is reasonable to apply $\mathbb{E}[\,\|\mu_*(X) - \mu(X)\|_{\mathcal{G}}^2\,]$ as an error criterion. This argument motivates the following notions of $L_2(\mathbb{X}, Q_X; \mathcal{G})$ consistency. Recall that Q_X is the marginal distribution of Q w.r.t. X.

Definition 5 *Let $(\mu_n)_{n\in\mathbb{N}}$ be a (random) estimate sequence for μ_*. The (random) L_2 error (or more precisely $L_2(\mathbb{X}, Q_X; \mathcal{G})$ error) of μ_n is defined by*

$$\int_{\mathbb{X}} \|\mu_*(x) - \mu_n(x)\|_{\mathcal{G}}^2 \,\mathrm{d}Q_X(x). \tag{16}$$

We say that $(\mu_n)_{n\in\mathbb{N}}$ *is weakly consistent for the distribution of* (X,Y)*, if as* $n\to\infty$*,*

$$\mathbb{E}\Big[\int_{\mathbb{X}} \|\mu_*(x)-\mu_n(x)\|_{\mathcal{G}}^2 \,\mathrm{d}Q_X(x)\Big] \longrightarrow 0. \tag{17}$$

We say that $(\mu_n)_{n\in\mathbb{N}}$ *is strongly consistent for the distribution of* (X,Y)*, if as* $n\to\infty$*,*

$$\int_{\mathbb{X}} \|\mu_*(x)-\mu_n(x)\|_{\mathcal{G}}^2 \,\mathrm{d}Q_X(x) \xrightarrow{a.s.} 0. \tag{18}$$

One can see that strong consistency does not necessarily imply weak consistency, indeed neither of these two consistency notions imply the other. Nevertheless, knowing that an estimator is weakly consistent will be of great help to also prove its strong consistency.

Note that it can happen that an estimate sequence is consistent for some distributions of (X,Y) and inconsistent for other distributions of (X,Y). We say that $(\mu_n)_{n\in\mathbb{N}}$ is *universally weakly (strongly) consistent* if it is weakly (strongly) consistent for all possible distributions of (X,Y). It is a somewhat surprising fact that there exist such estimates for $\mathbb{R}^d\to\mathbb{R}$ type regression functions, see (Stone, 1977), for example, local averaging adaptive methods such as k-Nearest Neighbors (kNN), several kernel-based methods and various partitioning rules were proved to be universally consistent, see the book of Györfi et al. (2002).

Recently Hanneke et al. (2020) proved that there exists a universally strongly consistent classifier (OptiNet) for any separable metric space and Györfi and Weiss (2021) constructed a universally consistent classification rule called Proto-NN for multiclass classification in metric spaces whenever the space admits a universally consistent estimator. For Banach space valued regression in semi-metric spaces Dabo-Niang and Rhomari (2009) constructed estimates with asymptotic bounds under some differentiation condition on the conditional expectation function and Forzani et al. (2012) proved consistency under the Besicovitch condition. Nevertheless, it is still an open question whether there exist universally consistent estimates under more general conditions (e.g., for Hilbert space valued regression).

3.2 Generalization of Stone's Theorem

In this section, we present a generalization of Stone's theorem (Stone, 1977) for conditional kernel mean map estimates for locally compact Polish[1] input spaces. This theorem will serve as our main theoretical tool to prove consistency results for our new recursive kernel algorithm. We consider local averaging methods of the general form

$$\mu_n(x) \doteq \sum_{i=1}^{n} W_{n,i}(x)\,\ell(\cdot,Y_i), \tag{19}$$

for the conditional kernel mean map, $\mathbb{E}[\,\ell(\cdot,Y)\,|\,X=x\,]$, in a locally compact Polish space $(\mathbb{X},d)$. A Polish space gives a convenient setup for probability theory and locally compactness is required because of Lemma 26. We use the fact that the set of compactly supported continuous functions is dense in $L_2(\mathbb{X},Q_X)$ for every probability measure Q_X when $\mathbb{X}$ is

1. A Polish space is a separable, completely metrizable topological space; as the specific choice of the metric is unimportant for the problems studied in this paper, we fix an arbitrary metric d which is compatible with the topology of our space, and treat a Polish space as a (complete and separable) metric space.

locally compact, see Theorem 21. We deal with data-dependent probability weight functions $W_{n,i}(x)$ for $n \in \mathbb{N}$ and $i \in [n]$, i.e., we assume that for all $n \in \mathbb{N}$, weight $W_{n,i}(x)$ is nonnegative for all $i \in [n]$ and $\sum_{i=1}^{n} W_{n,i}(x) = 1$ for all $x \in \mathbb{X}$. These are somewhat stronger conditions compare to the original theorem of Stone, however, in practice they are usually satisfied and simplify the proof. For several well-known methods (e.g, kNN, partitioning rules, Nadaraya–Watson estimates) these assumptions hold immediately.

Theorem 6 *Let $(\mathbb{X}, d)$ be a locally compact Polish space and let $\mathcal{X}$ be the Borel σ-algebra on $\mathbb{X}$. Assume A1 and A2. Let $\mu_*(x) \doteq \mathbb{E}[\,\ell(\cdot, Y) \,|\, X = x\,]$ and*

$$\mu_n(x) \doteq \sum_{i=1}^{n} W_{n,i}(x)\, \ell(\cdot, Y_i). \tag{20}$$

Assume that:

(i) For all $n \in \mathbb{N}$, $i \in [n]$ and $x \in \mathbb{X}$, we have

$$0 \leq W_{n,i}(x) \leq 1, \qquad \text{and} \qquad \sum_{i=1}^{n} W_{n,i}(x) = 1. \tag{21}$$

(ii) There exists $c > 0$ such that for all $f : \mathbb{X} \to \mathbb{R}^+$ with $\mathbb{E}[f(X)] < \infty$, where $\mathbb{R}^+$ denotes the nonnegative real numbers, we have

$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}(X) f(X_i)\Big] \leq c\, \mathbb{E}[f(X)]. \tag{22}$$

(iii) For all $\delta > 0$, we have

$$\lim_{n\to\infty} \mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}(X)\, \mathbb{I}\big(d(X_i, X) > \delta\big)\Big] = 0. \tag{23}$$

(iv) The following convergence holds for the weight functions:

$$\lim_{n\to\infty} \mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}^2(X)\Big] = 0. \tag{24}$$

Then, the estimate sequence (μ_n) is weakly consistent, that is

$$\lim_{n\to\infty} \mathbb{E} \int_{\mathbb{X}} \|\mu_n(x) - \mu_*(x)\|_G^2 \, \mathrm{d}Q_X(x) = 0. \tag{25}$$

Condition (i) ensures that $\{W_{n,i}\}_{i=1}^n$ are probability weights for all $n \in \mathbb{N}$. Assumption (ii) intuitively says that the L_1 norm of the local averaging estimate is at most some positive constant times the L_1 norm of the estimated function whenever there is no noise on the observations, i.e., when $Y_i = f(X_i)$. Condition (iii) guarantees that asymptotically only those inputs affect the estimate in a point x which are close to x. Condition (iv) ensures that the individual weights vanish as the sample size goes to infinity.

Proof The proof is based on the proof of Stone (1977) presented in (Györfi et al., 2002). By first using $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\mathbb{E}\Big[\, \|\mu_n(X) - \mu_*(X)\|_{\mathcal{G}}^2 \,\Big] \tag{26}$$
$$\leq 2\,\mathbb{E}\bigg[\, \Big\| \sum_{i=1}^{n} W_{n,i}(X)\big(\ell(\cdot, Y_i) - \mu_*(X_i)\big) \Big\|_{\mathcal{G}}^2 \bigg] + 2\,\mathbb{E}\bigg[\, \Big\| \sum_{i=1}^{n} W_{n,i}(X)\big(\mu_*(X_i) - \mu_*(X)\big) \Big\|_{\mathcal{G}}^2 \bigg].$$

Then, the application of Lemma 7 and Lemma 8 completes the proof. ■

Lemma 7 *Under the conditions of Theorem 6, we have that, as* $n \to \infty$,

$$\mathbb{E}\bigg[\, \Big\| \sum_{i=1}^{n} W_{n,i}(X)\big(\mu_*(X_i) - \mu_*(X)\big) \Big\|_{\mathcal{G}}^2 \bigg] \to 0. \tag{27}$$

Proof Let us use the following notation

$$J_n \doteq \mathbb{E}\bigg[\, \Big\| \sum_{i=1}^{n} W_{n,i}(X)\big(\mu_*(X_i) - \mu_*(X)\big) \Big\|_{\mathcal{G}}^2 \bigg]. \tag{28}$$

By Theorem 21 there exists a function $\tilde{\mu}_*$ which is bounded and continuous and

$$\mathbb{E}\Big[\, \|\mu_*(X) - \tilde{\mu}_*(X)\|_{\mathcal{G}}^2 \,\Big] < \varepsilon. \tag{29}$$

Recall that by the theorem of Heine and Cantor, see Theorem 27 in Appendix D, a continuous function with compact support is always uniformly continuous. Using the convexity of $f(x) = \|x\|_{\mathcal{G}}$ and also the elementary fact that $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$ yield

$$\begin{aligned}
J_n &= \mathbb{E}\bigg[\Big\| \sum_{i=1}^{n} W_{n,i}(X)\big(\mu_*(X_i) - \mu_*(X)\big) \Big\|_{\mathcal{G}}^2 \bigg] \\
&\leq \mathbb{E}\bigg[\Big(\sum_{i=1}^{n} W_{n,i}(X)\, \|\mu_*(X_i) - \mu_*(X)\|_{\mathcal{G}} \Big)^2 \bigg] \\
&\leq \mathbb{E}\bigg[\sum_{i=1}^{n} W_{n,i}(X)\, \|\mu_*(X_i) - \mu_*(X)\|_{\mathcal{G}}^2 \bigg] \\
&\leq 3\mathbb{E}\bigg[\sum_{i=1}^{n} W_{n,i}(X)\, \|\mu_*(X_i) - \tilde{\mu}_*(X_i)\|_{\mathcal{G}}^2 \bigg] \\
&\quad + 3\mathbb{E}\bigg[\sum_{i=1}^{n} W_{n,i}(X)\, \|\tilde{\mu}_*(X_i) - \tilde{\mu}_*(X)\|_{\mathcal{G}}^2 \bigg] \\
&\quad + 3\mathbb{E}\bigg[\sum_{i=1}^{n} W_{n,i}(X)\, \|\tilde{\mu}_*(X) - \mu_*(X)\|_{\mathcal{G}}^2 \bigg] \\
&\doteq 3\, J_n^{(1)} + 3\, J_n^{(2)} + 3\, J_n^{(3)}.
\end{aligned} \tag{30}$$

We bound these three terms separately. For any $\delta > 0$ one can see that

$$\begin{aligned} J_n^{(2)} &= \mathbb{E}\Bigg[\sum_{i=1}^n W_{n,i}(X)\,\|\tilde{\mu}_*(X_i) - \tilde{\mu}_*(X)\|_{\mathcal{G}}^2\,\mathbb{I}\big(d(X_i,X) > \delta\big)\Bigg] \\ &\quad + \mathbb{E}\Bigg[\sum_{i=1}^n W_{n,i}(X)\,\|\tilde{\mu}_*(X_i) - \tilde{\mu}_*(X)\|_{\mathcal{G}}^2\,\mathbb{I}\big(d(X_i,X) \le \delta\big)\Bigg] \\ &\le \mathbb{E}\Bigg[\sum_{i=1}^n W_{n,i}(X) 2\Big(\,\|\tilde{\mu}_*(X_i)\|_{\mathcal{G}}^2 + \|\tilde{\mu}_*(X)\|_{\mathcal{G}}^2\,\Big)\mathbb{I}\big(d(X_i,X) > \delta\big)\Bigg] \\ &\quad + \mathbb{E}\Bigg[\sum_{i=1}^n W_{n,i}(X)\,\|\tilde{\mu}_*(X_i) - \tilde{\mu}_*(X)\|_{\mathcal{G}}^2\,\mathbb{I}\big(d(X_i,X) \le \delta\big)\Bigg] \\ &\le 4\sup_{u\in\mathbb{X}}\|\tilde{\mu}(u)\|_{\mathcal{G}}^2\;\mathbb{E}\Big[\sum_{i=1}^n W_{n,i}(X)\cdot\mathbb{I}\big(d(X_i,X) > \delta\big)\Big] \\ &\quad + \sup_{u,v\in\mathbb{X}\,:\,d(u,v)\le\delta}\|\tilde{\mu}_*(u) - \tilde{\mu}_*(v)\|_{\mathcal{G}}^2\,. \end{aligned} \tag{31}$$

holds. We know that function $\tilde{\mu}_*$ is uniformly continuous, hence for any $\varepsilon > 0$ one can choose $\delta > 0$ such that

$$\sup_{u,v\in\mathbb{X}\,:\,d(u,v)\le\delta}\|\tilde{\mu}_*(u) - \tilde{\mu}_*(v)\|_{\mathcal{G}}^2 \;<\; \frac{\varepsilon}{2}, \tag{32}$$

hence the second term is less than $\varepsilon/2$. In addition, by condition (iii), for a given δ if n is large enough, then the first term is also less than $\varepsilon/2$. In conclusion $J_n^{(2)}$ converges to 0.

For the third term we have that

$$\begin{aligned} |J_n^{(3)}| &= \mathbb{E}\Big[\sum_{i=1}^n W_{n,i}(X)\,\|\tilde{\mu}_*(X) - \mu_*(X)\|_{\mathcal{G}}^2\,\Big] \\ &= \mathbb{E}\big[\,\|\tilde{\mu}_*(X) - \mu_*(X)\|_{\mathcal{G}}^2\,\big] < \varepsilon. \end{aligned} \tag{33}$$

For $J_n^{(1)}$ the application of condition (ii) to function $f(x) \doteq \|\mu_*(x) - \tilde{\mu}_*(x)\|_{\mathcal{G}}^2$ yields

$$\begin{aligned} J_n^{(1)} &= \mathbb{E}\Bigg[\sum_{i=1}^n W_{n,i}(X)\,\|\mu_*(X_i) - \tilde{\mu}_*(X_i)\|_{\mathcal{G}}^2\,\Bigg] \\ &\le c\,\mathbb{E}\Big[\,\|\mu_*(X) - \tilde{\mu}_*(X)\|_{\mathcal{G}}^2\,\Big] = c\,\varepsilon. \end{aligned} \tag{34}$$

Therefore, we can conclude that J_n tends to 0, as $n \to \infty$. ■

Lemma 8 *Under the conditions of Theorem 6, we have that*

$$\mathbb{E}\Bigg[\,\Bigg\|\sum_{i=1}^n W_{n,i}(X)\big(\ell(\cdot,Y_i) - \mu_*(X_i)\big)\Bigg\|_{\mathcal{G}}^2\,\Bigg] \to 0. \tag{35}$$

Proof For simplicity let

$$\begin{aligned} I_n &\doteq \mathbb{E}\Big[\Big\| \sum_{i=1}^{n} W_{n,i}(X)\big(\ell(\cdot, Y_i) - \mu_*(X_i)\big) \Big\|_{\mathcal{G}}^2 \Big] \\ &= \mathbb{E}\Big[\sum_{i=1}^{n}\sum_{j=1}^{n} W_{n,i}(X) W_{n,j}(X) \, \langle \ell(\cdot, Y_i) - \mu_*(X_i), \ell(\cdot, Y_j) - \mu_*(X_j) \rangle_{\mathcal{G}} \Big]. \end{aligned} \tag{36}$$

For $i \neq j$ we have that

$$\begin{aligned} &\mathbb{E}\Big[W_{n,i}(X) W_{n,j}(X) \langle \ell(\cdot, Y_i) - \mu_*(X_i), \ell(\cdot, Y_j) - \mu_*(X_j) \rangle_{\mathcal{G}} \Big] \\ &= \mathbb{E}\Big[\mathbb{E}\Big[W_{n,i}(X) W_{n,j}(X) \, \langle \ell(\cdot, Y_i) - \mu_*(X_i), \ell(\cdot, Y_j) - \mu_*(X_j) \rangle_{\mathcal{G}} \,|\, X, X_1, \dots X_n, Y_i \Big] \Big] \\ &= \mathbb{E}\Big[W_{n,i}(X) W_{n,j}(X) \, \big\langle \ell(\cdot, Y_i) - \mu_*(X_i), \mathbb{E}\Big[\ell(\cdot, Y_j) \,|\, X, X_1, \dots X_n, Y_i \Big] - \mu_*(X_j) \big\rangle_{\mathcal{G}} \Big] \\ &= \mathbb{E}\Big[W_{n,i}(X) W_{n,j}(X) \, \big\langle \ell(\cdot, Y_i) - \mu_*(X_i), \mathbb{E}[\ell(\cdot, Y_j) \,|\, X_j] - \mu_*(X_j) \big\rangle_{\mathcal{G}} \Big] = 0. \end{aligned} \tag{37}$$

Furthermore, by $\mathbb{E}[\ell(Y, Y)] < \infty$ we have

$$\mathbb{E}[\sigma^2(X)] < \infty \tag{38}$$

for $\sigma^2(X) \doteq \mathbb{E}[\, \|\ell(\cdot, Y) - \mu_*(X)\|_{\mathcal{G}}^2 \,|\, X]$. By (37) we have

$$\begin{aligned} I_n &= \mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}^2(X) \, \|\ell(\cdot, Y_i) - \mu_*(X_i)\|_{\mathcal{G}}^2 \Big] \\ &= \mathbb{E}\Big[\mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}^2(X) \, \|\ell(\cdot, Y_i) - \mu_*(X_i)\|_{\mathcal{G}}^2 \,\Big|\, X, X_1, \dots, X_n \Big] \Big] \\ &= \mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}^2(X) \mathbb{E}\Big[\|\ell(\cdot, Y_i) - \mu_*(X_i)\|_{\mathcal{G}}^2 \,\Big|\, X, X_1, \dots, X_n \Big] \Big] \\ &= \mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}^2(X) \mathbb{E}\Big[\|\ell(\cdot, Y_i) - \mu_*(X_i)\|_{\mathcal{G}}^2 \,\Big|\, X_i \Big] \Big] \\ &= \mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}^2(X) \sigma^2(X_i) \Big]. \end{aligned} \tag{39}$$

If σ^2 is bounded, then I_n tends to zero almost surely by condition (iv). Otherwise let $\tilde{\sigma}^2 \leq L$ be such that

$$\mathbb{E}[|\tilde{\sigma}^2(X) - \sigma^2(X)|] < \varepsilon, \tag{40}$$

see (Györfi et al., 2002, Theorem A.1). Then

$$\begin{aligned} I_n &\leq \mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}^2(X) \tilde{\sigma}^2(X_i) \Big] + \mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}^2(X) \,|\, \sigma^2(X_i) - \tilde{\sigma}^2(X_i)| \Big] \\ &\leq L \, \mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}^2(X) \Big] + \mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}(X) \,|\, \sigma^2(X_i) - \tilde{\sigma}^2(X_i)| \Big], \end{aligned} \tag{41}$$

where the first term tends to zero by condition (iv) and for the second term we have

$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}(X)\big|\,\sigma^2(X_i) - \tilde{\sigma}^2(X_i)\big|\Big] \leq c\,\mathbb{E}\left[\,\big|\,\sigma^2(X) - \tilde{\sigma}^2(X)\,\big|\,\right] \leq c\,\varepsilon \tag{42}$$

by condition (ii). In conclusion, I_n converges to 0. ■

3.3 Recursive Estimation of Conditional Kernel Mean Embeddings

Let $(\mathbb{X}, d)$ be a locally compact Polish space, as before. In this section, we apply local averaging *smoothers* or "kernels" on the input space (Györfi et al., 2002). These "kernels" are not necessarily positive definite, therefore, to avoid misunderstanding, we call them smoothers. Nevertheless, many positive definite functions are included in the definition.

Definition 9 (smoother) *A function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is called a smoother, if it is measurable, non-negative, symmetric, and bounded.*

Now, we introduce a recursive local averaging estimator. Let $(k_n)_{n\in\mathbb{N}}$ be a smoother sequence and $(a_n)_{n\in\mathbb{N}}$ be a stepsize (learning rate or gain) sequence of positive numbers.

Let us consider the following recursive algorithm:

$$\begin{aligned} \mu_1(x) &\doteq \ell(\cdot, Y_1), \quad \text{and} \\ \mu_{n+1}(x) &\doteq (1 - a_{n+1}k_{n+1}(x, X_{n+1}))\mu_n(x) + a_{n+1}k_{n+1}(x, X_{n+1})\ell(\cdot, Y_{n+1}), \end{aligned} \tag{43}$$

for $n \geq 1$. By induction, one can see that μ_n is of the form defined in (19). It is clear that one only needs to store μ_n, because in each iteration, when a new observation pair is available, one can immediately update μ_n by the recursion defined in (43).

In practice, we often need to evaluate $\mu_*(x)$ on a function $g \in \mathcal{G}$, i.e., the conditional expectation $\mathbb{E}[\,g(Y)\,|\,X = x\,]$ is sought. Then, our recursion simplifies to

$$\begin{aligned} &\langle \mu_1(x), g\rangle_{\mathcal{G}} = \langle \ell(\cdot, Y_1), g\rangle_{\mathcal{G}} = g(Y_1), \\ &\langle \mu_{n+1}(x), g\rangle_{\mathcal{G}} = \langle (1 - a_{n+1}k_{n+1}(x, X_{n+1}))\mu_n(x) + a_{n+1}k_{n+1}(x, X_{n+1})\ell(\cdot, Y_{n+1}), g\rangle_{\mathcal{G}} \\ &= (1 - a_{n+1}k_{n+1}(x, X_{n+1}))\langle \mu_n(x), g\rangle_{\mathcal{G}} + a_{n+1}k_{n+1}(x, X_{n+1})g(Y_{n+1}), \end{aligned} \tag{44}$$

which can be computed in linear time with constant memory. The real-valued consistency of this method can be established, e.g., by (Györfi et al., 2002, Theorem 25.1), however, we usually do not know in advance which function $g \in \mathcal{G}$ is of our interest, hence estimating the CMKE is very useful. Moreover, knowing the whole conditional distribution could also be used for various inference tasks, e.g., hypothesis testing or building prediction regions.

Theorem 10 presents sufficient conditions for the weak and strong consistency of our recursive scheme. It is a generalization of Theorem 25.1 of Györfi et al. (2002). There are two novelties w.r.t. the result in (Györfi et al., 2002). We prove consistency for *Hilbert space valued regression* and we cover *locally compact Polish input spaces*. The smoother functions are controlled on an intuitive and flexible manner. They do not need to admit a particular form or be monotonous. Intuitively the smoother functions should shrink around the observed input points as n goes to infinity. The rate of this decrease is controlled by stepsizes (h_n) and (r_n). The proposed conditions are w.r.t. a general majorant function H.

Theorem 10 *Let $(\mathbb{X}, d)$ be a locally compact Polish space. Assume that A1, A2 and the following conditions are satisfied.*

(i) There exist sequences $(h_n)_{n\in\mathbb{N}}$ and $(r_n)_{n\in\mathbb{N}}$ of positive numbers tending to 0 and a non-negative, nonincreasing function H on $[0,\infty)$ with $1/r_n \cdot H(s/h_n) \to 0$ $(n\to\infty)$ for all $s \in (0,\infty)$.

(ii) For all $n \in \mathbb{N}$ it holds that

$$k_n(x,z) \;\leq\; \frac{1}{r_n} H\big(\, d(x,z)/h_n \,\big). \tag{45}$$

(iii) The supremum of additive weights is bounded by 1, i.e.,

$$\sup_{x,z\in\mathbb{X}, n\in\mathbb{N}} a_n k_n(x,z) \;\leq\; 1. \tag{46}$$

(iv) Measure Q_X does not diminish w.r.t. the smoother functions, that is

$$\liminf_{n\to\infty} \int_{\mathbb{X}} k_n(x,t)\,\mathrm{d}Q_X(t) \;>\; 0 \quad \textit{for } Q_X\textit{-almost all } x. \tag{47}$$

(v) The positive learning rates satisfy

$$\sum_{n=1}^{\infty} a_n \;=\; \infty, \qquad \textit{and} \qquad \sum_{n=1}^{\infty} \frac{a_n^2}{r_n^2} \;<\; \infty. \tag{48}$$

Then, the estimate sequence $(\mu_n)_{n\in\mathbb{N}}$ defined by (43) is weakly and strongly consistent.

Conditions (i) and (ii) ensure that $k_n(x,z)$ vanishes for $x \neq z$ as $n\to\infty$. The weight functions are probability weights because of assumption (iii). Condition (iv) is the strongest technical assumption in a general metric measure space, however, in several cases it is easy to verify. Condition (v) can be usually guaranteed by the choice of stepsizes.

Proof For the weak consistency, we apply Theorem 6, hence our main task is to verify its assumptions. Observe that with $W_{1,1}(x) = 1$ and

$$W_{n,i}(x) \;=\; \prod_{l=i+1}^{n} \big(1 - a_l k_l(x, X_l)\big)\, a_i k_i(x, X_i), \tag{49}$$

for $n \geq 2$ and $i \in [n]$ we have $\mu_n(x) = \sum_{i=1}^n W_{n,i}(x)\ell(\cdot, Y_i)$. By induction, it is easy to see that the weights are probability weights. We start the procedure with a constant 1 function, then an update clearly does not change the sum of the weights for any input point $x \in \mathbb{X}$.

Condition (22) is verified in the proof of Györfi et al. (2002, Theorem 25.1). Note that (22) only regards the weight functions, which are identical for real-valued regression. Condition (24) again follows from the scalar-valued case.

In order to prove that condition (23) holds, we follow the proof in (Györfi et al., 2002, Theorem 25.1) with necessary modifications. By using independence of $\{X_i\}_{i=1}^n$ and the inequality $(1-x) \leq e^{-x}$, we have for all $i \leq n$ that

$$\begin{aligned}\mathbb{E}\big[\,W_{n,i}(x)\,\big] &= \prod_{l=i+1}^{n} \mathbb{E}\big[\,(1-a_l k_l(x,X_l))\,\big]\,\mathbb{E}\big[\,a_i k_i(x,X_i)\,\big] \\ &\leq \frac{a_i\,H(0)}{r_i} e^{-\sum_{l=i+1}^{n} a_l \mathbb{E}\left[k_l(x,X_l)\right]} \xrightarrow{n\to\infty} 0,\end{aligned} \tag{50}$$

because $\sum_{l=i+1}^{n} a_l \mathbb{E}[k_l(x,X_l)] \xrightarrow{n\to\infty} \infty$ by condition (iv) and $\sum_{n=1}^{\infty} a_n = \infty$. Let us fix $\delta > 0$. By the dominated convergence theorem, it is sufficient to prove that for almost all x it holds that

$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{n,i}(x)\mathbb{I}(d(x,X_i)>\delta)\Big] \to 0. \tag{51}$$

Let $p_n(x) \doteq \mathbb{E}[k_n(x,X)]$. By condition (iv) we have that

$$p(x) \doteq \frac{1}{2}\liminf_{n\to\infty} p_n(x) > 0 \tag{52}$$

holds Q_X-almost everywhere implying that for Q_X-almost all x there exists $n_0(x)$ such that, for $n > n_0(x)$ we have $p_n(x) \geq p(x)$. Because of (50), for such $x \in \mathbb{X}$ it is sufficient to prove

$$\mathbb{E}\Big[\sum_{i=n_0(x)}^{n} W_{n,i}(x)\mathbb{I}(d(x,X_i)>\delta)\Big] \xrightarrow{n\to\infty} 0. \tag{53}$$

By independence one can observe that

$$\begin{aligned}&\mathbb{E}\Big[W_{n,i}(x)\mathbb{I}(d(x,X_i)>\delta)\Big] = \mathbb{E}\Big[\prod_{l=i+1}^{n}(1-a_l k_l(x,X_l))a_i k_i(x,X_i)\mathbb{I}(d(x,X_i)>\delta)\Big] \\ &= \mathbb{E}\Big[\prod_{l=i+1}^{n}(1-a_l k_l(x,X_l))\Big]\mathbb{E}[a_i k_i(x,X_i)\mathbb{I}(d(x,X_i)>\delta)] \\ &= \frac{\mathbb{E}[W_{n,i}(x)]}{\mathbb{E}[a_i k_i(x,X_i)]}\mathbb{E}[a_i k_i(x,X_i)\mathbb{I}(d(x,X_i)>\delta)],\end{aligned} \tag{54}$$

therefore, we have

$$\begin{aligned}&\mathbb{E}\Big[\sum_{i=n_0(x)}^{n} W_{n,i}(x)\mathbb{I}(d(x,X_i)>\delta)\Big] \\ &= \sum_{i=n_0(x)}^{n}\mathbb{E}[W_{n,i}(x)]\frac{\mathbb{E}[a_i k_i(x,X_i)\mathbb{I}(d(x,X_i)>\delta)]}{\mathbb{E}[a_i k_i(x,X_i)]} \\ &\leq \sum_{i=n_0(x)}^{n}\mathbb{E}[W_{n,i}(x)]\frac{H(\delta/h_i)}{r_i\,p(x)}.\end{aligned} \tag{55}$$

Since $\sum_{i=n_0(x)}^{n}\mathbb{E}[W_{n,i}(x)] \leq 1$, $\mathbb{E}[W_{n,i}(x)] \to 0$ and $1/r_n\,H(\delta/h_n) \to 0$ as $n \to \infty$ the application of Lemma 29 yields

$$\sum_{i=n_0(x)}^{n}\mathbb{E}\big[\,W_{n,i}(x)\,\big]\,\frac{H(\delta/h_i)}{r_i\,p(x)} \to 0. \tag{56}$$

Consequently, the weak consistency of (μ_n) follows from Theorem 6.

The strong consistency of $(\mu_n)_{n\in\mathbb{N}}$ can be proved based on the almost supermartingale convergence theorem of Robbins and Siegmund (1971), see Theorem 30. The technical details of the proof are presented in Appendix C. ■

We presented our recursive algorithm in a general setup. It remains to construct smoother sequences and stepsizes that satisfy the proposed conditions. In the theorem that follows we generate the smoothers with the help of an $\mathbb{R} \to \mathbb{R}$ type *mother kernel* K. In the next section we present several corollaries of this theorem to design consistent algorithms for many important machine learning applications.

Theorem 11 *Let $(\mathbb{X}, d)$ be a locally compact Polish space and $K : \mathbb{R} \to \mathbb{R}$ be a measurable, nonnegative and bounded function. Assume A1 and A2. Let $(h_n)_{n\in\mathbb{N}}$ and $(r_n)_{n\in\mathbb{N}}$ be positive sequences with zero limit. Let us define the following smoothers*

$$k_n(x, z) \doteq \frac{1}{r_n} K\Big(\frac{d(x, z)}{h_n}\Big), \tag{57}$$

for $n \in \mathbb{N}$. Let $(\mu_n)_{n\in\mathbb{N}}$ be defined as in (43). Assume that:

(i) There are positive constants R and b such that $K(t) \geq b\,\mathbb{I}(t \in B(0, R))$ for all $t \in \mathbb{R}$, where $B(0, R)$ denotes the open ball with center 0 and radius R.

(ii) There is a measurable, nonnegative and nonincreasing function H on $[0, \infty)$ such that $1/r_n\, H(s/h_n) \to 0$ as $n \to \infty$ and $K(s) \leq H(s)$ for all $s \in (0, \infty)$.

(iii) The nonnegative learning rates satisfy

$$\sum_{n=1}^{\infty} a_n = \infty, \qquad \text{and} \qquad \sum_{n=1}^{\infty} \frac{a_n^2}{r_n^2} < \infty. \tag{58}$$

(iv) The weights of the update term is bounded by 1, i.e.,

$$\sup_{t\in\mathbb{R}} K(t) \sup_{n\in\mathbb{N}} \frac{a_n}{r_n} \leq 1. \tag{59}$$

(v) The probability of small balls is bounded from below by

$$\liminf_{n\to\infty} \frac{Q_X(B(x, R\,h_n))}{r_n} > 0 \qquad Q_X\text{ - a.s.} \tag{60}$$

Then, $(\mu_n)_{n\in\mathbb{N}}$ is weakly and strongly consistent.

Proof It is sufficient to verify the conditions of Theorem 10. We assumed that

$$1/r_n\, H(s/h_n) \to 0, \tag{61}$$

for all $s \in (0, \infty)$. From (57) and condition (ii) it follows that

$$r_n k_n(x, z) = K(d(x, z)/h_n) \leq H(d(x, z)/h_n). \tag{62}$$

Condition (iv) implies that

$$\sup_{x,z\in\mathbb{X},n\in\mathbb{N}} a_n k_n(x,z) = \sup_{x,z\in\mathbb{X},n\in\mathbb{N}} \frac{a_n}{r_n} K\Big(\frac{d(x,z)}{h_n}\Big) \leq \sup_{t\in\mathbb{R}} K(t) \sup_{n\in\mathbb{N}} \frac{a_n}{r_n} \leq 1 \tag{63}$$

holds. Finally by condition (i) we have

$$\begin{aligned}
&\int_{\mathbb{X}} k_n(x,t)\,\mathrm{d}Q_X(t) = \int_{\mathbb{X}} \frac{1}{r_n} K\Big(\frac{d(x,t)}{h_n}\Big)\,\mathrm{d}Q_X(t) \\
&\geq \int_{\mathbb{X}} \frac{b}{r_n} \mathbb{I}\Big(\frac{d(x,t)}{h_n} \in B(0,R)\Big)\,\mathrm{d}Q_X(t) \\
&= \int_{\mathbb{X}} \frac{b}{r_n} \mathbb{I}\Big(t \in B(x, R\,h_n)\Big)\,\mathrm{d}Q_X(t) = \frac{b\,Q_X(B(x,R\,h_n))}{r_n}.
\end{aligned} \tag{64}$$

Taking the lim inf on both sides and applying condition (v) yield (47). The conditions on the stepsize sequence are assumed. Therefore, by applying Theorem 10, we can conclude that the estimate sequence, $(\mu_n)_{n\in\mathbb{N}}$, is weakly and strongly consistent. ■

4. Demonstrative Applications

In this section we demonstrate the applicability of Theorem 11 through some characteristic model classes important for statistical learning. We consider three general setups. The main difference of these applications lies in the input spaces. We show that the requirement on the probability of small balls is satisfied in these cases and provide constructions for the learning rates, (a_n), (r_n) and (h_n), such that they satisfy the conditions of Theorem 11.

First, *Euclidean* input spaces are considered, for which we prove the weak and strong universal consistency of the recursive estimator of the conditional kernel mean map. The main novelty of this corollary compared to the results of Györfi et al. (2002, Chapter 25) is that our method is applicable for special *Hilbert space valued regression* problems.

Second, we study abstract *Riemannian manifolds* as input spaces for which an embedding into an "ambient" Euclidean space may not be explicitly given. One of the motivations of this example comes from *manifold learning* (Tenenbaum et al., 2000; Lin and Zha, 2008), which is an actively studied area in machine learning, whose theoretical study is full of challenges. The presented recursive framework can offer a new technical tool for this direction.

Finally, we investigate the case of *functional inputs* (Ferraty and Vieu, 2006). This example is motivated by *signal processing* problems and methods for continuous-time systems in time-series analysis. We consider a wavelet-like approach leading to an input space that is a locally compact subset of $L_2(\mathbb{R})$ induced by translations of a suitable *mother function*.

4.1 Euclidean Spaces

When $\mathbb{X}$ is an Euclidean space, the weak and strong universal consistency of the recursive estimator of the conditional kernel mean map is a consequence of Theorem 11.

Corollary 12 *Let $\mathbb{X} = \mathbb{R}^p$ be a p-dimensional Euclidean space, with the Euclidean metric, and $K : \mathbb{R} \to \mathbb{R}$ be a measurable, nonnegative and bounded function. Let the stepsizes be*

defined as $a_n \doteq 1/n$, *and let* $r_n \doteq 1/n^{(1-\varepsilon)/2}$, *where* $\varepsilon \in (0,1)$, *and* $h_n \doteq \sqrt[p]{r_n}$ *and* k_n *be as in* (57) *for* $n \in \mathbb{N}$. *Let* $(\mu_n)_{n\in\mathbb{N}}$ *be defined as in* (43). *Assume A1, A2 and conditions* (i) *and* (ii) *from Theorem 11, then* $(\mu_n)_{n\in\mathbb{N}}$ *is weakly and strongly universally consistent.*

Proof We can assume (w.l.o.g.) that K is bounded by 1. A Euclidean space is a locally compact Polish space, therefore, we can apply Theorem 11. Conditions (i) and (ii) are assumed. Our choice of stepsizes clearly satisfies (58) and (59). For (60) we use (Devroye, 1981, Lemma 2.2) which proves that for every probability measure Q_X on $\mathbb{R}^p$ there exists a nonnegative function g satisfying $g(x) < \infty$ Q_X-almost surely, such that

$$\frac{h_n^p}{Q_X(B(x, R\,h_n))} \to g(x) \qquad Q_X - \text{a.s.} \tag{65}$$

for all $R > 0$. Therefore

$$\liminf_{n\to\infty} \frac{Q_X(B(x, R\,h_n))}{r_n} = \frac{1}{g(x)} > 0 \qquad Q_X - \text{a.s.} \tag{66}$$

and then the statement of corollary is proved by applying Theorem 11. ■

Remark 13 *Mother kernel* K *can be chosen based on prior knowledge on the problem. In the table that follows we list some of the possible candidates.*

Name	*Formula*
Box kernel	$\mathbb{I}(\lvert x\rvert < B)$, *with* $B > 0$
Gaussian kernel	$\exp\left(-\frac{x^2}{\sigma}\right)$, *with* $\sigma > 0$
Laplace kernel	$\exp\left(-\frac{\lvert x\rvert}{\sigma}\right)$, *with* $\sigma > 0$
Epanechnikov kernel	$3/4(1-x^2)\,\mathbb{I}(\lvert x\rvert < 1)$

One can easily verify that all of these functions are measurable, nonnegative and bounded by 1. *They also satisfy conditions* (i) *and* (ii) *from Theorem 11 with* $H = K|_{[0,\infty)}$.

4.2 Riemannian Manifolds

In our recursive scheme the power of h_n w.r.t. r_n is (at least) p for a p-dimensional Euclidean space. Manifolds can often be embedded in a high-dimensional Euclidean space, however, they are locally isomorphic to a lower dimensional one. This local isometry allows us to use the lower manifold dimension as the power of h_n w.r.t. r_n. Moreover, we also cover *abstract* p-dimensional manifolds, hence an ambient Euclidean space is not required.

Corollary 14 *Let* $\mathbb{X}$ *be a* p*-dimensional connected, geodesically complete Riemannian manifold, with geodesical distance* ϱ *and* $K : \mathbb{R} \to \mathbb{R}$ *be a measurable, nonnegative and bounded function. Let* $a_n \doteq 1/n$, $r_n \doteq 1/n^{(1-\varepsilon)/2}$, *where* $\varepsilon \in (0,1)$, *and* $h_n = \sqrt[p]{r_n}$ *and* k_n *be as in* (57) *for* $n \in \mathbb{N}$. *Let* $(\mu_n)_{n\in\mathbb{N}}$ *be defined as in* (43). *Assume A1, A2 and conditions* (i) *and* (ii) *from Theorem 11, then* $(\mu_n)_{n\in\mathbb{N}}$ *is weakly and strongly universally consistent.*

Proof Assume (w.l.o.g.) that K is bounded by 1. Every topological manifold is locally compact (Lee, 2013, Proposition 1.12). Connected Riemannian manifolds are metrizable, henceforth paracompact[2] (Stone, 1948). A connected paracompact metric space is second countable, thus it is separable and admits a countable atlas. Moreover, by the Hopf-Rinow theorem a geodesically complete Riemannian manifold is complete as a metric space.

Consider a countable smooth atlas $\mathcal{A}$ on $\mathbb{X}$. We are going to prove the assumption about small ball probabilities for every chart in the atlas almost surely, hence the claim for the whole space will follow almost surely. Consider a chart (U, f) in $\mathcal{A}$.

If $Q_X(U) = 0$, then the condition is satisfied automatically, otherwise consider the probability measure $Q_X/Q_X(U)$ on U. We push measure $Q_X/Q_X(U)$ to $\mathbb{R}^p$, i.e., consider the probability measure μ on $\mathbb{R}^p$, defined by $\mu(A) \doteq Q_X \circ f^{-1}[f(U) \cap A]/Q_X(U)$ for every Borel set A in $\mathbb{R}^p$. We have already seen that

$$\liminf_{n\to\infty} \frac{\mu(B_2(s, R\, h_n))}{h_n^p} > 0 \qquad \mu - \text{almost everywhere}, \tag{67}$$

that is there exists a measurable set $\Omega_1 \subseteq \mathbb{R}^p$ such that $\mu(\Omega_1) = 1$ and for all $s \in \Omega_1$ one has (67). Notice that one can assume that $\Omega_1 \subseteq f(U)$ (otherwise we can take the intersection).

Let $q \in \Omega_1$. We know that f^{-1} is a smooth diffeomorphism, therefore there exists $C > 0$ such that $\left\|Df^{-1}(s)\right\| \leq C$ for $s \in B_2(q, 1/C\, h_n)$, where D is the differential operator and $\|\cdot\|$ denotes the operator norm. In the next step, we prove that $B_\varrho(f^{-1}(q), h_n) \supseteq f^{-1}[B_2(q, 1/C\, h_n) \cap f(U)]$ for h_n small enough. Let $s \in B_2(q, 1/C\, h_n) \cap f(U)$. Then, there are $x = f^{-1}(q)$ and $\bar{x} \doteq f^{-1}(s)$. By Lemma 32, we have

$$\varrho(x, \bar{x}) = \varrho(f^{-1}(q), f^{-1}(s)) \leq C\, \|q - s\|_2 \leq h_n. \tag{68}$$

It follows that

$$Q_X(B_\varrho(f^{-1}(q), h_n)) \geq Q_X \circ f^{-1}(B_2(q, 1/C\, h_n) \cap f(U)) = \mu(B_2(q, 1/C\, h_n))\, Q_X(U). \tag{69}$$

Let $\Omega_0 = f^{-1}[\Omega_1] \subseteq \mathbb{X}$, for which $Q_X(\Omega_0)/Q_X(U) = 1$. Moreover, for $x \in \Omega_0$ we have that

$$\liminf_{n\to\infty} \frac{Q_X(B_\varrho(x, h_n))}{h_n^d} \geq \liminf_{n\to\infty} \frac{Q_X(U)\mu(B_2(f(x), 1/C\, h_n))}{h_n^d} > 0. \tag{70}$$

holds for every $x \in \Omega_0$. Hence (70) holds $Q_X/Q_X(U)$-almost surely and Q_X-almost everywhere. The rest of the proof is similar to that of the Euclidean case. ■

4.3 Locally Compact Subset of a Hilbert Space

Now, we consider *functional inputs*, which can be useful, e.g., in *signal processing*. The analyzed function set is motivated by *wavelets*. Our model class can be seen as how a signal is encoded in a computer by storing a finite number of coefficients for some basis functions at given knots, where the locations of these knots are flexible and part of the representation.

2. A topological space is paracompact, if every open cover has an open refinement that is locally finite.

Let us consider the Hilbert space $L_2(\mathbb{R})$. Let ψ be a function in a Sobolev space $W^{1,2}(\mathbb{R})$ (Berlinet and Thomas-Agnan, 2004), and τ_x be the translation operator on $L_2(\mathbb{R})$ defined pointwise by $\tau_x f(t) \doteq f(t-x)$. For given (constant) hyper-parameters $m \in \mathbb{N}, M > 0$, let

$$\mathcal{M} \doteq \left\{ f : \mathbb{R} \to \mathbb{R} \;\middle|\; f = \sum_{i=1}^{m} \lambda_i \, \tau_{x_i} \psi \text{ for some } \lambda_1, \ldots, \lambda_m, x_1, \ldots, x_m \in [-M, M] \right\}. \quad (71)$$

Consider the canonical (measurable) mapping ϕ from $[-M, M]^{2m}$ to $\mathcal{M}$ given by

$$\phi : \begin{bmatrix} \lambda \\ x \end{bmatrix} \mapsto \sum_{i=1}^{m} \lambda_i \, \tau_{x_i} \psi, \quad (72)$$

where $\lambda = [\lambda_1, \ldots, \lambda_m]^{\mathrm{T}}$ and $x = [x_1, \ldots, x_m]^{\mathrm{T}}$. Clearly, ϕ is surjective, however, it is not injective. We are going to consider probability measures of the form $Q_X \doteq P_X \circ \phi^{-1}$ for some P_X on $[-M, M]^{2m}$, i.e., we assume that Q_X is the pushforward of some measure on the bounded (locally compact) Polish space $[-M, M]^{2m}$ with the Euclidean metric.

Corollary 15 *Let $\psi \in W^{1,2}(\mathbb{R})$, $\mathbb{X} = \mathcal{M}$ with the L_2 metric, $K : \mathbb{R} \to \mathbb{R}$ be a measurable, nonnegative and bounded function. Let $a_n \doteq 1/n$, $r_n \doteq 1/n^{(1-\varepsilon)/2}$, for an $\varepsilon \in (0,1)$, and $h_n = \sqrt[2m]{r_n}$ and k_n be as in (57) for $n \in \mathbb{N}$. Let $(\mu_n)_{n\in\mathbb{N}}$ be defined as in (43). Assume (i)-(ii) from Theorem 11, then $(\mu_n)_{n\in\mathbb{N}}$ is weakly and strongly consistent for measures of the form $Q_X \doteq P_X \circ \phi^{-1}$ for any distribution P_X on $[-M, M]^{2m}$, where ϕ is given by (72).*

Proof We need to prove that $\mathcal{M}$ is locally compact and Polish. First, we show that $\mathcal{M}$ is complete with the L_2 metric. It is known that L_2 is complete. Hence, for any Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{M}$ we have a limit function $f \in L_2(\mathbb{R})$. We show that f is also in $\mathcal{M}$. Let $f_n = \sum_{i=1}^{m} \lambda_{n,i} \, \tau_{x_{n,i}} \psi$. We know that $[-M, M]$ is a compact set hence for $i = 1$ there is a subsequence $(n_k)_{k\in\mathbb{N}}$ of $\mathbb{N}$ such that $\lambda_{n_k,1} \to \lambda_1$. One can iteratively find subsequences of the resulting subsequences such that for all $i = 1, \ldots, m$ the convergences of $\lambda_{s_k,i} \to \lambda_i$ and $x_{s_k,i} \to x_i$ hold true for a subsequence $(s_k)_{k\in\mathbb{N}}$ of $\mathbb{N}$. Because of $\psi \in W^{1,2}(\mathbb{R})$ function ψ is (Lebesque) almost everywhere continuous. Hence, $f_{s_k} \to \hat{f}$ holds almost everywhere for $\hat{f} \doteq \sum_{i=1}^{n} \lambda_i \tau_{x_i} \psi$. It is known, see (Rudin, 1987, Theorem 3.12), that this is sufficient for $f = \hat{f}$ almost everywhere. A similar argument can be used to show that $\mathcal{M}$ is sequentially compact. Then, since $\mathcal{M}$ is a metric space, it is also compact; not just locally, but globally.

We also need to prove that for all $f \in \mathcal{M}$

$$\liminf_{n\to\infty} \frac{Q_X(B(f, R\,h_n))}{h_n^{2m}} > 0 \qquad Q_X\text{-a.s.} \quad (73)$$

We show that for $\tilde{f} \doteq \sum_{i=1}^{m} \tilde{\lambda}_i \tau_{\tilde{x}_i} \psi$ from

$$\left\| \begin{bmatrix} \lambda \\ x \end{bmatrix} - \begin{bmatrix} \tilde{\lambda} \\ \tilde{x} \end{bmatrix} \right\|_2 < h \quad (74)$$

it follows that $\|f - \tilde{f}\|_2 < C\,h$. Recall that on finite dimensional vector spaces all norms are equivalent, thus one can use $\|\cdot\|_\infty$ instead of $\|\cdot\|_2$, i.e., there exists $c > 0$ such that

$$\left\| \begin{bmatrix} \lambda \\ x \end{bmatrix} - \begin{bmatrix} \tilde{\lambda} \\ \tilde{x} \end{bmatrix} \right\|_\infty \leq c \left\| \begin{bmatrix} \lambda \\ x \end{bmatrix} - \begin{bmatrix} \tilde{\lambda} \\ \tilde{x} \end{bmatrix} \right\|_2 \quad (75)$$

holds. In conclusion, if (74) holds, then

$$\left\| \begin{bmatrix} \lambda \\ x \end{bmatrix} - \begin{bmatrix} \tilde{\lambda} \\ \tilde{x} \end{bmatrix} \right\|_\infty \leq c\, h. \tag{76}$$

We use the notion of moduli of continuity to prove that $\left\| f - \tilde{f} \right\|_2 < C\, h$. Recall that

$$\omega_\psi^{(2)}(h) \doteq \sup_{|x| \leq h} \int |\psi(t+x) - \psi(t)|^2 \, \mathrm{d}t. \tag{77}$$

It is known (Evans, 2010, Section 5.8,Theorem 3) that for $\psi \in W^{1,2}(\mathbb{R})$ there is a constant M_ψ with

$$\omega_\psi^{(2)}(h) \leq M_\psi h^2. \tag{78}$$

Using the triangle inequality and (78) one can gain

$$\begin{aligned}
\left\| f - \tilde{f} \right\|_2 &= \left\| \sum_{i=1}^m \lambda_i \tau_{x_i} \psi - \sum_{i=1}^m \tilde{\lambda}_i \tau_{\tilde{x}_i} \psi \right\|_2 \\
&\leq \left\| \sum_{i=1}^m \lambda_i \tau_{x_i} \psi - \sum_{i=1}^m \tilde{\lambda}_i \tau_{x_i} \psi + \sum_{i=1}^m \tilde{\lambda}_i \tau_{x_i} \psi - \sum_{i=1}^m \tilde{\lambda}_i \tau_{\tilde{x}_i} \psi \right\|_2 \\
&\leq \left\| \sum_{i=1}^m \lambda_i \tau_{x_i} \psi - \sum_{i=1}^m \tilde{\lambda}_i \tau_{x_i} \psi \right\|_2 + \left\| \sum_{i=1}^m \tilde{\lambda}_i \tau_{x_i} \psi - \sum_{i=1}^m \tilde{\lambda}_i \tau_{\tilde{x}_i} \psi \right\|_2 \\
&\leq \sum_{i=1}^m |\lambda_i - \tilde{\lambda}_i| \, \left\| \tau_{x_i} \psi \right\|_2 + \sum_{i=1}^m |\tilde{\lambda}_i| \, \left\| \tau_{x_i} \psi - \tau_{\tilde{x}_i} \psi \right\|_2 \\
&\leq h\, c\, m \, \left\| \psi \right\|_2 + M \sum_{i=1}^m \Big(\int_{\mathbb{R}} |\psi(t - x_i) - \psi(t - \tilde{x}_i)|^2 \, \mathrm{d}t \Big)^{1/2} \\
&\leq h\, c\, m \, \left\| \psi \right\|_2 + M \sum_{i=1}^m \Big(\int_{\mathbb{R}} |\psi(s) - \psi(s + (x_i - \tilde{x}_i))|^2 \, \mathrm{d}t \Big)^{1/2} \\
&\leq h\, c\, m \, \left\| \psi \right\|_2 + M \sum_{i=1}^m \sqrt{\omega_\psi^{(2)}(c\, h)} \\
&\leq h\, c\, m \Big(\left\| \psi \right\|_2 + M \sqrt{M_\psi} \Big),
\end{aligned} \tag{79}$$

hence our claim is satisfied with $C = m\, c(\left\| \psi \right\|_2 + M \sqrt{M_\psi})$.

We showed that $B(f, C\, h_n))) \supseteq \phi(B([\lambda^{\mathrm{T}}, x^{\mathrm{T}}]^{\mathrm{T}}, h_n))$, therefore

$$\begin{aligned}
&Q_X(B(f, R\, h_n))) \geq P_X \circ \phi^{-1}[\phi(B([x^{\mathrm{T}}, \lambda^{\mathrm{T}}]^{\mathrm{T}}, R/C\, h_n))] \\
&\geq P_X(B([\lambda^{\mathrm{T}}, x^{\mathrm{T}}]^{\mathrm{T}}, R/C\, h_n)).
\end{aligned} \tag{80}$$

By (Devroye, 1981, Lemma 2.2) we conclude that

$$\liminf_{n \to \infty} \frac{P_X(B([\lambda^{\mathrm{T}}, x^{\mathrm{T}}]^{\mathrm{T}}, R/C\, h_n))}{h_n^{2m}} = \frac{1}{g([\lambda^{\mathrm{T}}, x^{\mathrm{T}}]^{\mathrm{T}})} > 0 \qquad P_X - \text{a.s.} \tag{81}$$

and

$$\liminf_{n\to\infty} \frac{Q_X(B(f, R\,h_n))}{h_n^{2m}} > 0 \qquad Q_X - \text{a.s.} \tag{82}$$

By Theorem 11 the corollary follows. ■

5. Discussion

In this paper, we have introduced a new recursive estimator for the conditional kernel mean map, which is a key object to represent conditional distributions in Hilbert spaces. We have proved the weak and strong L_2 consistency of the introduced scheme in the Bochner sense, under general structural conditions. We have considered a locally compact Polish input space and assumed only the knowledge of a kernel on the (measurable) output space. Our recursive algorithm uses a smoother sequence on the metric input space and three learning sequences, each of these having a particular role in the iterative settings. Learning rate (a_n) adjusts the weights of the current estimate term and the update term in each step. Sequences (h_n) and (r_n) are used to adaptively shrink the smoother functions w.r.t. the intrinsic measuretic structure of the metric space. Our main assumption concerns the probabilities of small balls, i.e., we assume that the measure of small balls with radius h_n are comparable to r_n when n goes to infinity. We have generalized a result of Györfi et al. (2002, Theorem 25.1) for locally compact Polish input spaces and Hilbert space valued regressions. We have considered three important application domains of our recursive estimation scheme. Our consistency results are universal in case of Euclidean input spaces and complete Riemannian manifolds, because the small ball probability assumption can be satisfied by (h_n) and (r_n) for all probability distributions. A Hilbert space valued input set (containing functions) was also considered to illustrate the generality of the framework.

Acknowledgements

The authors warmly thank László Gerencsér for his many comments and suggestions regarding the paper. His advises helped us to shape the final manuscript in many ways.

The research was supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. This work has also been supported by the TKP2021-NKTA-01 grant of the National Research, Development and Innovation Office (NRDIO), Hungary. The authors acknowledge the professional support of the Doctoral Student Scholarship Program of the Cooperative Doctoral Program of the Ministry of Innovation and Technology, as well, financed from an NRDIO Fund.

Appendix A.

The Bochner Integral

Bochner integral is an extended integral notion for Banach space valued functions. Let $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ be a Banach space with the Borel σ-field of $\mathcal{G}$ and $(\Omega, \mathcal{A}, \mathbb{P})$ be a measure space with a finite measure $\mathbb{P}$. Similarly to the Lebesgue integral, the Bochner integral is introduced first for simple $\Omega \to \mathcal{G}$ type functions of the form $f(\omega) = \sum_{j=1}^{n} \mathbb{I}(\omega \in A_j)\, g_j$, where $A_j \in \mathcal{A}$

and $g_j \in \mathcal{G}$ for all $j \in [n]$, by

$$\int_\Omega f \,\mathrm{d}\mathbb{P} \doteq \sum_{j=1}^{n} \mathbb{P}(A_j)\, g_j. \tag{83}$$

Then, the integral is extended to the abstract completion of these simple functions with respect to the norm defined by the following Lebesgue integral

$$\|f\|_{L_1(\mathbb{P};\mathcal{G})} \doteq \int_\Omega \|f\|_{\mathcal{G}} \,\mathrm{d}\mathbb{P}. \tag{84}$$

Definition 16 *A function $f : \Omega \to \mathcal{G}$ is called strongly measurable (Bochner-measurable) if there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of simple functions converging to f almost everywhere.*

By the Pettis theorem (Hytönen et al., 2016, Theorem 1.1.20), this is equivalent to requiring the weak measurability of f and that $f(\Omega) \subseteq \mathcal{G}$ is separable.

Definition 17 *Let $1 \le p < \infty$. Then, the vector valued $\mathcal{L}_p(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$-space is defined as*

$$\mathcal{L}_p(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G}) = \mathcal{L}(\mathbb{P}; \mathcal{G}) \doteq \Big\{ f : \Omega \to \mathcal{G} \,\Big|\, f \text{ is strongly measurable and } \int \|f\|_{\mathcal{G}}^p \,\mathrm{d}\mathbb{P} < \infty \Big\}. \tag{85}$$

A semi-norm on $\mathcal{L}_p(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$ is defined by

$$\|f\|_{L_p(\mathbb{P};\mathcal{G})} \doteq \Big(\int \|f\|_{\mathcal{G}}^p \,\mathrm{d}\mathbb{P} \Big)^{\frac{1}{p}}, \quad \text{for } f \in \mathcal{L}_p(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G}). \tag{86}$$

Taking equivalence classes on $\mathcal{L}_p(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$ w.r.t. semi-norm $\|\cdot\|_{L_p}$ yields the vector valued L_p space (Bochner space) denoted by $L_p(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$, or simply by $L_p(\Omega, \mathbb{P}; \mathcal{G})$.

Proposition 18 *Space $(L_p(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G}), \|\cdot\|_{L_p(\mathbb{P};\mathcal{G})})$ is a Banach space for $1 \le p < \infty$.*

Several of the well-known theorems about Lebesgue integrals can be extended to Bochner integrals, see (Dinculeanu, 2000; Hytönen et al., 2016; Pisier, 2016).

Proposition 19 (Bochner's inequality) *For all $f \in L_1(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$, we have*

$$\Big\| \int f \,\mathrm{d}\mathbb{P} \Big\|_{\mathcal{G}} \le \int \|f\|_{\mathcal{G}} \,\mathrm{d}\mathbb{P}. \tag{87}$$

Moreover, f is Bochner integrable if and only if the right hand side is finite, however, recall that f is required to be strongly measurable to be included in $L_1(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$. Furthermore, observe that $\int \|f\|_{\mathcal{G}} \,\mathrm{d}\mathbb{P} = \|f\|_{L_1(\mathbb{P};\mathcal{G})}$, hence the Bochner integral is a continuous operator from $L_1(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$ to Banach space $\mathcal{G}$. By definition, the Bochner integral is also linear.

Proposition 20 *Let E and F be Banach spaces and $T : E \to F$ be a bounded linear operator. Then, for all $f \in L_1(\mathbb{P}; E)$, we have that $T \circ f \in L_1(\mathbb{P}; F)$ and*

$$T\Big(\int f \,\mathrm{d}\mathbb{P} \Big) = \int T(f) \,\mathrm{d}\mathbb{P}. \tag{88}$$

In particular, for a bounded linear functional $x^* : \mathcal{G} \to \mathbb{R}$, we have

$$\Big\langle \int_\Omega f(\omega) \,\mathrm{d}\mathbb{P}(\omega), x^* \Big\rangle = \int_\Omega \langle f(\omega), x^* \rangle \,\mathrm{d}\mathbb{P}(\omega), \tag{89}$$

where $\langle g, x^* \rangle \doteq x^*(g)$.

Now, let $(\mathcal{G}, \langle\cdot,\cdot\rangle_{\mathcal{G}})$ be a Hilbert space. By Proposition 20 for all $g \in \mathcal{G}$, we have

$$\Big\langle \int_\Omega f(\omega)\,\mathrm{d}\mathbb{P}(\omega), g \Big\rangle_{\mathcal{G}} = \int_\Omega \langle \mu(\omega), g\rangle_{\mathcal{G}}\,\mathrm{d}\mathbb{P}(\omega). \tag{90}$$

Moreover, it can be proved that $L_2(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$ is also a Hilbert space with the following inner product

$$\langle \mu_1, \mu_2\rangle_{L_2(\Omega,\mathbb{P};\mathcal{G})} \doteq \int \langle \mu_1(\omega), \mu_2(\omega)\rangle_{\mathcal{G}}\,\mathrm{d}\mathbb{P}(\omega). \tag{91}$$

Conditional Expectation in Banach Spaces

Let $\mathbb{P}$ be a probability measure and $\mathcal{B}$ a sub-σ-algebra of $\mathcal{A}$. The conditional expectation with respect to $\mathcal{B}$ is defined for any $f \in L_1(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$ as the $\mathbb{P}$-a.s. unique element $\mathbb{E}[f \,|\, \mathcal{B}] \doteq \nu \in L_1(\Omega, \mathcal{B}, \mathbb{P}; \mathcal{G})$ satisfying

$$\int_B \nu\,\mathrm{d}\mathbb{P} = \int_B f\,\mathrm{d}\mathbb{P}, \quad \text{for all } B \in \mathcal{B}. \tag{92}$$

It can be shown that the conditional expectation is a (well-defined) bounded operator from $L_1(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$ to $L_1(\Omega, \mathcal{B}, \mathbb{P}; \mathcal{G})$ (Hytönen et al., 2016, Theorem 2.6.23), (Pisier, 2016, Proposition 1.4). The nonexpanding property also holds in the $L_p(\mathbb{P}; \mathcal{G})$ sense for all $1 \leq p < \infty$, see (Hytönen et al., 2016, Corollary 2.6.30), that is for all $f \in L_p(\mathbb{P}; \mathcal{G})$ we have

$$\| \mathbb{E}[\, f \,|\, \mathcal{B}\,] \|_{\mathcal{G}}^p \leq \mathbb{E}[\, \|f\|_{\mathcal{G}}^p \,|\, \mathcal{B}\,], \tag{93}$$

and by Bochner's inequality $\big\|\mathbb{E}[f \,|\, \mathcal{B}]\big\|_{L_p(\Omega;\mathcal{G})} \leq \|f\|_{L_p(\Omega;\mathcal{G})}$. Furthermore, by (Hytönen et al., 2016, Prop. 2.6.31.) we can take out the $\mathcal{B}$-measurable term from the conditional expectation. In particular, if $\mathcal{G}$ is a Hilbert space, then for all $g \in \mathcal{G}$ and $f \in L_2(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$ we (a.s.), have that

$$\langle\, \mathbb{E}[\, \mu \,|\, \mathcal{B}\,], g\,\rangle_{\mathcal{G}} = \mathbb{E}[\, \langle \mu, g\rangle_{\mathcal{G}} \,|\, \mathcal{B}\,]. \tag{94}$$

Appendix B.

In the proof of the generalized Stone's theorem a continuous approximation of the conditional kernel mean map is needed. The existence of such approximation is ensured by the following theorem.

Theorem 21 *Let $\mathbb{X}$ be a locally compact Hausdorff space equipped with a Radon measure*[3] *Q_X and let $\mathcal{G}$ be a separable Hilbert space. The space of compactly supported continuous functions, $\mathcal{C}_c(\mathbb{X}; \mathcal{G})$, is dense in $L_2(\mathbb{X}, Q_X; \mathcal{G})$, that is for all $\mu \in L_2(\mathbb{X}, Q_X; \mathcal{G})$ and $\varepsilon > 0$, there exists a $\tilde{\mu} \in \mathcal{C}_c(\mathbb{X}; \mathcal{G})$ such that*

$$\int_{\mathbb{X}} \|\mu(x) - \tilde{\mu}(x)\|_{\mathcal{G}}^2\,\mathrm{d}Q_X(x) < \varepsilon. \tag{95}$$

A similar result is proved for $\mathbb{R}^d$ in (Hytönen et al., 2016, Lemma 1.2.31). They mention in a remark that their lemma can be generalized for locally compact spaces, but the complete

3. A Radon measure is a regular Borel measure.

proof of their claim were not found by us in the literature. Nevertheless, the presented argument has a similar flavor as the ones for scalar valued L_p spaces.

Proof First, we prove that we can approximate μ with simple functions in $L_2(\mathbb{X}, Q_X; \mathcal{G})$. Let $\mu \in L_2(\mathbb{X}, Q_X; \mathcal{G})$ and $\varepsilon > 0$. There exists a sequence of simple functions $(\mu_n)_{n\in\mathbb{N}}$ such that $\mu_n \to \mu$ a.s., i.e., $\mu_n(x) \doteq \sum_{i=1}^{k_n} g_i \, \mathbb{I}(x \in A_i)$ where $g_i \in \mathcal{G}$ and A_i is measurable with $Q_X(A_i) < \infty$ for all $i \in [k_n]$. Let $B_n \doteq \{x \in \mathbb{X} : \|\mu_n(x)\|_{\mathcal{G}} \leq 2 \, \|\mu(x)\|_{\mathcal{G}}\}$ and

$$h_n(x) \doteq \mathbb{I}(x \in B_n) \, \mu_n(x). \tag{96}$$

One can see that the $(B_n)_{n\in\mathbb{N}}$ sequence consists of measurable sets and $h_n \to \mu$ almost everywhere. We show that $h_n \to \mu$ in $L_2(\mathbb{X}, Q_X; \mathcal{G})$. Because of

$$\|h_n\|^2_{L_2(\mathbb{X};\mathcal{G})} = \int_{B_n} \|\mu_n(x)\|^2_{\mathcal{G}} \, \mathrm{d}Q_X(x) \leq 2^2 \int_{\mathbb{X}} \|\mu(x)\|^2_{\mathcal{G}} \, \mathrm{d}Q_X(x), \tag{97}$$

we have $h_n \in L_2(\mathbb{X}, Q_X; \mathcal{G})$ for $n \in \mathbb{N}$. In addition, almost everywhere we have

$$\|\mu(x) - h_n(x)\|^2_{\mathcal{G}} \leq (3 \, \|\mu(x)\|_{\mathcal{G}})^2, \tag{98}$$

therefore, by the dominated convergence theorem

$$\begin{aligned} \lim_{n\to\infty} \|\mu - h_n\|^2_{L_2(\mathbb{X},Q_X;\mathcal{G})} &= \lim_{n\to\infty} \int_{\mathbb{X}} \|\mu(x) - h_n(x)\|^2_{\mathcal{G}} \, \mathrm{d}Q_X(x) \\ &= \int_{\mathbb{X}} \lim_{n\to\infty} \|\mu(x) - h_n(x)\|^2_{\mathcal{G}} \, \mathrm{d}Q_X(x) = 0. \end{aligned} \tag{99}$$

In conclusion, for any $\varepsilon > 0$ there exists a simple function $h \in L_2(\mathbb{X}, Q_X; \mathcal{G})$ such that $\|\mu - h\|_{L_2(\mathbb{X},Q_X;\mathcal{G})} < \varepsilon$. Therefore, it is sufficient to approximate h with a continuous function on a compact support. Let $h \in L_2(\mathbb{X}, Q_X; \mathcal{G})$ be

$$h(x) = \sum_{i=1}^{N} g_i \, \mathbb{I}(x \in A_i), \tag{100}$$

where $g_i \in \mathcal{G}$ and $Q_X(A_i) < \infty$ for $i \in [N]$. The main idea is to approximate the indicator functions, $\{g_i \cdot \mathbb{I}(x \in A_i)\}_{i=1}^n$, separately. Let us fix $i = 1$ and $\varepsilon > 0$. Since Q_X is Radon and A_1 is measurable there exists a compact set $K \subseteq \mathbb{X}$ such that $K \subseteq A_1$ and

$$Q_X(A_1 \setminus K) < \frac{\varepsilon^2}{2N^2 \, \|g_1\|^2_{\mathcal{G}}}, \tag{101}$$

and there exists an open set U such that $A_1 \subseteq U$ with

$$Q_X(U \setminus A_1) < \frac{\varepsilon^2}{2N^2 \, \|g_1\|^2_{\mathcal{G}}}. \tag{102}$$

The application of Lemma 25 yields that there exists an open set E such that $\bar{E}$ is compact and $K \subseteq E \subseteq \bar{E} \subseteq U$. Because of Lemma 26 there is a continuous function f_1 such

that $f_1|_K = 1$ and $f_1|_{\mathbb{X}\setminus E} = 0$, from which it follows that f_1 has a compact support. For $h_1(x) \doteq g_1 \cdot \mathbb{I}(x \in A_1)$ and $\tilde{h}_1(x) \doteq g_1 \cdot f_1(x)$ we have

$$\begin{aligned}
&\int_{\mathbb{X}} \Big\| h_1(x) - \tilde{h}_1(x) \Big\|_{\mathcal{G}}^2 \,\mathrm{d}Q_X(x) = \int_{\mathbb{X}} \| g_1\, \mathbb{I}(x \in A_1) - g_1\, f_1(x) \|_{\mathcal{G}}^2 \,\mathrm{d}Q_X(x) \\
&= \|g_1\|_{\mathcal{G}}^2 \int_{\mathbb{X}} \big(\mathbb{I}(x \in A_1) - f_1(x)\big)^2 \,\mathrm{d}Q_X(x) \leq \|g_1\|_{\mathcal{G}}^2 \int_{E\setminus K} 1 \,\mathrm{d}Q_X(x) \\
&= \|g_1\|_{\mathcal{G}}^2 \; Q_X(E \setminus K) \leq \|g_1\|_{\mathcal{G}}^2 \; Q_X(U \setminus K) < \frac{\varepsilon^2}{N^2}.
\end{aligned} \tag{103}$$

Similarly, one can construct $\tilde{h}_i$ which is close to h_i in $L_2(\mathbb{X}, Q_X; \mathcal{G})$ for all $i \in [N]$. Let $\tilde{\mu} \doteq \sum_{i=1}^N \tilde{h}_i$. Obviously, $\tilde{\mu}$ is continuous and has a compact support. Furthermore, by the triangle inequality we have

$$\begin{aligned}
&\sqrt{\int_{\mathbb{X}} \|h(x) - \tilde{\mu}(x)\|_{\mathcal{G}}^2 \,\mathrm{d}Q_X(x)} = \sqrt{\int_{\mathbb{X}} \Big\| \sum_{i=1}^N g_i(\mathbb{I}(x \in A_i) - f_i(x)) \Big\|_{\mathcal{G}}^2 \,\mathrm{d}Q_X(x)} \\
&\leq \sum_{i=1}^N \sqrt{\int_{\mathbb{X}} \|g_i(\mathbb{I}(x \in A_i) - f_i(x))\|_{\mathcal{G}}^2 \,\mathrm{d}Q_X(x)} \leq \sum_{i=1}^N \sqrt{\frac{\varepsilon^2}{N^2}} = \varepsilon,
\end{aligned} \tag{104}$$

which finishes the proof of the theorem. ■

Lemma 22 *Let $\{(\mathbb{X}_i, \mathcal{X}_i)\}_{i=1}^3$ be measurable spaces, let X_i, for $i \in \{1,2,3\}$, be $\mathbb{X}_i$-valued independent random elements on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with push-forward measure Q_i. Let $f : \mathbb{X}_1 \times \mathbb{X}_2 \times \mathbb{X}_3 \to \mathbb{R}$ and $g : \mathbb{X}_1 \times \mathbb{X}_2 \to \mathbb{R}$ be measurable functions such that $f(X_1, X_2, X_3)$ and $g(X_1, X_2)$ are in $L_1(\Omega, \mathbb{P})$. If for Q_1-almost all $x \in \mathbb{X}_1$ it holds that*

$$\mathbb{E}\big[\, f(x_1, X_2, X_3) \,|\, X_2 \,\big] \;\leq\; g(x_1, X_2), \tag{105}$$

then almost surely

$$\mathbb{E}\Big[\int_{\mathbb{X}_1} f(x_1, X_2, X_3) \,\mathrm{d}Q_1(x_1) \,\Big|\, X_2 \Big] \;\leq\; \int_{\mathbb{X}_1} g(x_1, X_2) \,\mathrm{d}Q_1(x_1). \tag{106}$$

Proof Integrating out both sides in (105) w.r.t. Q_1 yields

$$\int_{\mathbb{X}_1} \mathbb{E}\Big[f(x_1, X_2, X_3) \,|\, X_2 \Big] \,\mathrm{d}Q_1(x_1) \leq \int_{\mathbb{X}_1} g(x_1, X_2) \,\mathrm{d}Q_1(x_1) = \mathbb{E}[g(X_1, X_2)]. \tag{107}$$

In addition, the left hand side is

$$\begin{aligned}
&\mathbb{E}\Big[\int_{\mathbb{X}} f(x_1, X_2, X_3) \,\mathrm{d}Q_1(x_1) \,|\, X_2 \Big] = \int_{\mathbb{X}_3} \int_{\mathbb{X}_1} f(x_1, X_2, x_3) \,\mathrm{d}Q_1(x_1) \,\mathrm{d}Q_3(x_3) \\
&= \int_{\mathbb{X}_1} \int_{\mathbb{X}_3} f(x_1, X_2, x_3) \,\mathrm{d}Q_3(x_3) \,\mathrm{d}Q_1(x_1) = \int_{\mathbb{X}_1} \mathbb{E}\Big[f(x_1, X_2, X_3) \,|\, X_2 \Big] \,\mathrm{d}Q_1(x_1),
\end{aligned} \tag{108}$$

because of Fubini's theorem. ■

Appendix C

The details of the proof for strong consistency in Theorem 10 are presented in this section.

Proof By the weak consistency of $(\mu_n)_{n\in\mathbb{N}}$, Fubini's theorem and (93) we have

$$\int \| \mathbb{E}[\mu_n(x)] - \mu_*(x) \|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x) \leq \mathbb{E} \int \| \mu_n(x) - \mu_*(x) \|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x) \to 0. \tag{109}$$

Because of $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned} &\mathbb{E} \int \|\mu_n(x) - \mathbb{E}[\mu_n(x)]\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x) \\ &\leq 2\Big(\mathbb{E} \int \|\mu_n(x) - \mu_*(x)\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x) + \int \|\mu_*(x) - \mathbb{E}[\mu_n(x)]\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x)\Big) \to 0. \end{aligned} \tag{110}$$

Since

$$\begin{aligned} &\int \|\mu_n(x) - \mu_*(x)\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x) \\ &\leq 2\Big(\int \|\mu_n(x) - \mathbb{E}[\mu_n(x)]\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x) + \int \|\mu_*(x) - \mathbb{E}[\mu_n(x)]\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x)\Big) \end{aligned} \tag{111}$$

holds, for the strong consistency of (μ_n) it is sufficient to prove that

$$\int \|\mu_n(x) - \mathbb{E}[\mu_n(x)]\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x) \to 0 \tag{112}$$

almost surely. By (110) the integral admits a finite limit which must agree with the weak limit (which is zero). Let us consider the following expansion:

$$\begin{aligned} \mu_{n+1}(x) - \mathbb{E}[\mu_{n+1}(x)] &= \mu_n(x) - \mathbb{E}[\mu_n(x)] \\ &\quad - a_{n+1}(\mu_n(x)k_{n+1}(x, X_{n+1}) - \mathbb{E}[\mu_n(x)k_{n+1}(x, X_{n+1})]) \\ &\quad + a_{n+1}(\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1}) - \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})]). \end{aligned} \tag{113}$$

Let $\mathcal{F}_n \doteq \sigma(X_1, Y_1, \ldots, X_n, Y_n)$ for $n \in \mathbb{N}$. Taking the conditional expectation of the norm square of (113) w.r.t. $\mathcal{F}_n$ yields

$$\begin{aligned} &\mathbb{E}\Big[\|\mu_{n+1}(x) - \mathbb{E}[\mu_{n+1}(x)]\|_{\mathcal{G}}^2 \mid \mathcal{F}_n\Big] = \|\mu_n(x) - \mathbb{E}[\mu_n(x)]\|_{\mathcal{G}}^2 \\ &+ a_{n+1}^2 \mathbb{E}\Big[\big\|\mu_n(x)k_{n+1}(x, X_{n+1}) - \mathbb{E}\big[\mu_n(x)k_{n+1}(x, X_{n+1})\big]\big\|_{\mathcal{G}}^2 \mid \mathcal{F}_n\Big] \\ &+ a_{n+1}^2 \mathbb{E}\Big[\big\|\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1}) - \mathbb{E}\big[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})\big]\big\|_{\mathcal{G}}^2 \mid \mathcal{F}_n\Big] \\ &- 2a_{n+1}\mathbb{E}\Big[\big\langle \mu_n(x) - \mathbb{E}[\mu_n(x)], \mu_n(x)k_{n+1}(x, X_{n+1}) - \mathbb{E}[\mu_n(x)k_{n+1}(x, X_{n+1})]\big\rangle_{\mathcal{G}} \mid \mathcal{F}_n\Big] \\ &+ 2a_{n+1}\mathbb{E}\Big[\big\langle \mu_n(x) - \mathbb{E}[\mu_n(x)], \ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1}) - \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})]\big\rangle_{\mathcal{G}} \mid \mathcal{F}_n\Big] \\ &- 2a_{n+1}^2\mathbb{E}\Big[\big\langle \mu_n(x)k_{n+1}(x, X_{n+1}) - \mathbb{E}[\mu_n(x)k_{n+1}(x, X_{n+1})], \\ &\qquad \ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1}) - \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})]\big\rangle_{\mathcal{G}} \mid \mathcal{F}_n\Big] \end{aligned}$$

$$= I_1 + I_2 + I_3 + I_4 + I_5 + I_6, \tag{114}$$

where I_i are defined respectively for $i \in \{1, \ldots, 6\}$. We bound these terms separately using the measurability condition, the independence of the sample and condition (ii). Our main tool to prove almost sure convergence uses almost supermartingales, Theorem 30 from (Robbins and Siegmund, 1971). Therefore, our main task in this section is to bound each I_i for $i \in \{1, \ldots, 6\}$ by $(1+\alpha_n)(\,\|\mu_n(x) - \mathbb{E}[\mu_n(x)]\|^2_{\mathcal{G}}\,) + \beta_n$ where α_n and β_n are small enough, i.e.,

$$\sum_{n\in\mathbb{N}} \mathbb{E}[\alpha_n] < \infty \qquad \text{and} \qquad \sum_{n\in\mathbb{N}} \mathbb{E}[\beta_n] < \infty \tag{115}$$

holds. Clearly, I_1 is bounded by itself. For the second term we have

$$\begin{aligned}
I_2 &\doteq a^2_{n+1}\mathbb{E}\Big[\, \big\|\mu_n(x)k_{n+1}(x, X_{n+1}) - \mathbb{E}\big[\mu_n(x)k_{n+1}(x, X_{n+1})\big]\big\|^2_{\mathcal{G}} \,\big|\, \mathcal{F}_n\Big] \\
&= a^2_{n+1}\Big(\mathbb{E}\Big[\, \|\mu_n(x)k_{n+1}(x, X_{n+1})\|^2_{\mathcal{G}} \,\big|\, \mathcal{F}_n\Big] + \|\mathbb{E}[\mu_n(x)]\mathbb{E}[k_{n+1}(x, X_{n+1})]\|^2_{\mathcal{G}} \\
&\quad - 2\langle \mu_n(x)\mathbb{E}[k_{n+1}(x, X_{n+1})], \mathbb{E}[\mu_n(x)]\mathbb{E}[k_{n+1}(x, X_{n+1})]\rangle_{\mathcal{G}}\Big) \\
&= a^2_{n+1}\Big(\, \|\mu_n(x)\|^2_{\mathcal{G}}\, \big(\mathbb{E}[k^2_{n+1}(x, X_{n+1})] - \big(\mathbb{E}[k_{n+1}(x, X_{n+1})]\big)^2\big) \\
&\quad + \|\mu_n(x)\|^2_{\mathcal{G}}\, \big(\mathbb{E}[k_{n+1}(x, X_{n+1})]\big)^2 - 2\langle \mu_n(x), \mathbb{E}[\mu_n(x)]\rangle_{\mathcal{G}}\big(\mathbb{E}[k_{n+1}(x, X_{n+1})]\big)^2 \\
&\quad + \|\mathbb{E}[\mu_n(x)]\|^2_{\mathcal{G}}\, \big(\mathbb{E}[k_{n+1}(x, X_{n+1})]\big)^2\Big) \\
&= a^2_{n+1}\Big(\, \|\mu_n(x)\|^2_{\mathcal{G}}\, \big(\mathbb{E}[k^2_{n+1}(x, X_{n+1})] - \big(\mathbb{E}[k_{n+1}(x, X_{n+1})]\big)^2\big) \\
&\quad + \|\mu_n(x) - \mathbb{E}[\mu_n(x)]\|^2_{\mathcal{G}}\, \big(\mathbb{E}[k_{n+1}(x, X_{n+1})]\big)^2\Big) \\
&\leq \frac{H^2(0)a^2_{n+1}}{r^2_{n+1}}\big(\,\|\mu_n(x)\|^2_{\mathcal{G}} + \|\mu_n(x) - \mathbb{E}[\mu_n(x)]\|^2_{\mathcal{G}}\,\big)
\end{aligned} \tag{116}$$

because recall that $r_{n+1}k(x, X_{n+1}) \leq H(0)$. We also used that (X_{n+1}, Y_{n+1}) is independent of $\mathcal{F}_n$ and $\mu_n(x)$ is measurable w.r.t. $\mathcal{F}_n$. The third term can be bounded as follows

$$\begin{aligned}
I_3 &\doteq a^2_{n+1}\mathbb{E}\Big[\, \big\|\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1}) - \mathbb{E}\big[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})\big]\big\|^2_{\mathcal{G}} \,\big|\, \mathcal{F}_n\Big] \\
&= a^2_{n+1}\Big(\mathbb{E}\Big[\, \|\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})\|^2_{\mathcal{G}}\,\Big] + \big\|\mathbb{E}\big[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})\big]\big\|^2_{\mathcal{G}} \\
&\quad - 2\mathbb{E}\Big[\langle \ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1}), \mathbb{E}\big[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})\big]\rangle_{\mathcal{G}}\Big]\Big) \\
&= a^2_{n+1}\Big(\mathbb{E}\Big[\ell(Y_{n+1}, Y_{n+1})k^2_{n+1}(x, X_{n+1})\Big] - \big\|\mathbb{E}\big[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})\big]\big\|^2_{\mathcal{G}}\Big) \\
&\leq \frac{H^2(0)a^2_{n+1}}{r^2_{n+1}}\mathbb{E}[\ell(Y, Y)].
\end{aligned} \tag{117}$$

The fourth term is less than 0 because

$$I_4 = -2a_{n+1}\mathbb{E}\Big[\big\langle \mu_n(x) - \mathbb{E}[\mu_n(x)], \mu_n(x)k_{n+1}(x, X_{n+1}) - \mathbb{E}[\mu_n(x)k_{n+1}(x, X_{n+1})]\big\rangle_{\mathcal{G}} \,\big|\, \mathcal{F}_n\Big]$$

$$
\begin{aligned}
&= -2a_{n+1}\langle \mu_n(x) - \mathbb{E}[\mu_n(x)], \mu_n(x)\mathbb{E}[k_{n+1}(x, X_{n+1})] - \mathbb{E}[\mu_n(x)]\mathbb{E}[k_{n+1}(x, X_{n+1})]\rangle_{\mathcal{G}} \\
&= -2a_{n+1} \left\| \mu_n(x) - \mathbb{E}[\mu_n(x)] \right\|_{\mathcal{G}}^2 \mathbb{E}[k_{n+1}(x, X_{n+1})] \leq 0.
\end{aligned} \tag{118}
$$

It is easy to see that $I_5 = 0$, because of independence we have

$$
\begin{aligned}
&\mathbb{E}\Big[\langle \mu_n(x) - \mathbb{E}[\mu_n(x)], \ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1}) - \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})]\rangle_{\mathcal{G}} \,|\, \mathcal{F}_n\Big] \\
&\quad = \langle \mu_n(x) - \mathbb{E}[\mu_n(x)], \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})] - \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})]\rangle_{\mathcal{G}}.
\end{aligned} \tag{119}
$$

For the last term we use the Cauchy-Schwartz inequality and $2ab \leq a^2 + b^2$ to show that

$$
\begin{aligned}
I_6 &= -2a_{n+1}^2 \mathbb{E}\Big[\langle \mu_n(x)k_{n+1}(x, X_{n+1}) - \mathbb{E}[\mu_n(x)k_{n+1}(x, X_{n+1})], \\
&\qquad\qquad \ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1}) - \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})]\rangle_{\mathcal{G}} \,|\, \mathcal{F}_n\Big] \\
&= -2a_{n+1}^2 \Big(\mathbb{E}\Big[\langle \mu_n(x)k_{n+1}(x, X_{n+1}), \ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})\rangle_{\mathcal{G}} \,|\, \mathcal{F}_n\Big] \\
&\quad - \langle \mathbb{E}[\mu_n(x)k_{n+1}(x, X_{n+1})], \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1}) \,|\, \mathcal{F}_n]\rangle_{\mathcal{G}} \\
&\quad - \langle \mathbb{E}[\mu_n(x)k_{n+1}(x, X_{n+1}) \,|\, \mathcal{F}_n], \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})]\rangle_{\mathcal{G}} \\
&\quad + \langle \mathbb{E}[\mu_n(x)k_{n+1}(x, X_{n+1})], \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1}) \,|\, \mathcal{F}_n]\rangle_{\mathcal{G}} \Big) \\
&= -2a_{n+1}^2 \mathbb{E}[k_{n+1}^2(x, X_{n+1})\langle \mu_n(x), \ell(\cdot, Y_{n+1})\rangle_{\mathcal{G}} \,|\, \mathcal{F}_n] \\
&\quad + 2a_{n+1}^2 \mathbb{E}[k_{n+1}(x, X_{n+1})]\langle \mu_n(x), \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})]\rangle_{\mathcal{G}} \\
&\leq 2a_{n+1}^2 \mathbb{E}[k_{n+1}^2(x, X_{n+1}) \left\| \mu_n(x) \right\|_{\mathcal{G}} \sqrt{\ell(Y_{n+1}, Y_{n+1})} \,|\, \mathcal{F}_n] \\
&\quad + 2a_{n+1}^2 \mathbb{E}[k_{n+1}(x, X_{n+1})] \left\| \mu_n(x) \right\|_{\mathcal{G}} \left\| \mathbb{E}[\ell(\cdot, Y_{n+1})k_{n+1}(x, X_{n+1})] \right\|_{\mathcal{G}} \\
&\leq 2a_{n+1}^2 \left\| \mu_n(x) \right\|_{\mathcal{G}} \mathbb{E}\Big[\frac{H^2(0)}{r_{n+1}^2} \sqrt{\ell(Y_{n+1}, Y_{n+1})} \,|\, \mathcal{F}_n \Big] \\
&\quad + 2a_{n+1}^2 \left\| \mu_n(x) \right\|_{\mathcal{G}} \frac{H^2(0)}{r_{n+1}^2} \mathbb{E}[\left\| \ell(\cdot, Y_{n+1}) \right\|_{\mathcal{G}}] \\
&= 4\frac{H^2(0)a_{n+1}^2}{r_{n+1}^2} \left\| \mu_n(x) \right\|_{\mathcal{G}} \mathbb{E}\Big[\sqrt{\ell(Y, Y)} \Big] \leq 2\frac{H^2(0)a_{n+1}^2}{r_{n+1}^2} (\left\| \mu_n(x) \right\|_{\mathcal{G}}^2 + \mathbb{E}[\ell(Y, Y)]).
\end{aligned} \tag{120}
$$

By summarizing these upper bounds we have almost surely that

$$
\begin{aligned}
&\mathbb{E}\Big[\left\| \mu_{n+1}(x) - \mathbb{E}[\mu_{n+1}(x)] \right\|_{\mathcal{G}}^2 \,|\, \mathcal{F}_n \Big] \\
&\leq \Big(1 + \frac{H^2(0)a_{n+1}^2}{r_{n+1}^2} \Big) \left\| \mu_n(x) - \mathbb{E}[\mu_n(x)] \right\|_{\mathcal{G}}^2 + \frac{3H^2(0)a_{n+1}^2}{r_{n+1}^2} (\left\| \mu_n(x) \right\|_{\mathcal{G}}^2 + \mathbb{E}[\ell(Y, Y)])
\end{aligned} \tag{121}
$$

for all $x \in \mathbb{X}$. By Lemma 22 one can integrate w.r.t. Q_X to prove that

$$
\begin{aligned}
&\mathbb{E}\Big[\int \left\| \mu_{n+1}(x) - \mathbb{E}[\mu_{n+1}(x)] \right\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x) \,|\, \mathcal{F}_n \Big] \\
&\quad \leq \Big(1 + \frac{H^2(0)a_{n+1}^2}{r_{n+1}^2} \Big) \int \left\| \mu_n(x) - \mathbb{E}[\mu_n(x)] \right\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x) \\
&\qquad + \frac{3H^2(0)a_{n+1}^2}{r_{n+1}^2} \Big(\int \left\| \mu_n(x) \right\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x) + \mathbb{E}[\ell(Y, Y)] \Big).
\end{aligned} \tag{122}
$$

Observe that $\mathbb{E}[\ell(Y,Y)] < \infty$ and $\mathbb{E}\big[\int \|\mu_n(x)\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x)\big]$ is convergent, hence also bounded. Consequently, from

$$\sum_{n=1}^{\infty} \frac{H^2(0) a_{n+1}^2}{r_{n+1}^2} < \infty$$

it follows that $\int \|\mu_{n+1}(x) - \mathbb{E}[\mu_{n+1}(x)]\|_{\mathcal{G}}^2 \, \mathrm{d}Q_X(x)$ converges almost surely by Theorem 30. The almost surely constant limit value is zero because of our previous argument. ■

Appendix D

In this appendix, for convenience, we state the main theorems that are crucial for our proofs. An important result is the generalized SLLN for Hilbert space valued random elements. The following theorem is from the book of Taylor (1978, Theorem 3.2.4).

Theorem 23 *If $(X_n)_{n\in\mathbb{N}}$ is a sequence of independent random elements in a separable Hilbert space such that*

$$\sum_{n=1}^{\infty} \frac{\operatorname{var}(X_n)}{n^2} < \infty, \tag{123}$$

where $\operatorname{var}(X_n) \doteq \mathbb{E}\Big[\, \|X_n - \mathbb{E}X_n\|^2 \,\Big]$, *then*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \right\| \xrightarrow{a.s.} 0. \tag{124}$$

We referred to (Ljung et al., 2012, Theorem 1.9) to prove the consistency of our recursive (unconditional) kernel mean embedding estimate of a marginal distribution.

Theorem 24 *Let $\mathcal{G}$ be a real separable Hilbert space endowed with the Borel σ-algebra and $\Lambda : \mathcal{G} \to \mathcal{G}$ be measurable and $\mu \in \mathcal{G}$ (usually but not necessarily $\Lambda(\mu) = 0$). Let further μ_n, W_n, H_n, V_n be $\mathcal{G}$-valued random elements for $n \in \mathbb{N}$ with $\mu_0 = 0$,*

$$\mu_{n+1} \doteq \mu_n - a_n(\Lambda(\mu_n) - W_n) \qquad \text{and} \qquad W_n = H_n + V_n, \tag{125}$$

where $a_n \geq 0$, $\sum_n a_n^2 < \infty$ and $\sum_n a_n = \infty$. Assume that

(i) There exists $c > 0$ such that for all $g \in \mathcal{G}$ we have $\|\Lambda(g)\|_{\mathcal{G}} \leq c(1 + \|g\|_{\mathcal{G}})$.

(ii) For all $K \in [1, \infty)$ we have $\inf\{\langle \Lambda(g), g - \mu\rangle_{\mathcal{G}} \,|\, g \in \mathcal{G} \text{ with } \frac{1}{K} \leq \|g - \mu\|_{\mathcal{G}} \leq K\} > 0$.

(iii) For all $n \in \mathbb{N}$ random elements H_n and V_n are square integrable with

$$\begin{aligned} &\sum_n a_n \, \mathbb{E}\, \|H_n\|_{\mathcal{G}} < \infty, \qquad \sum_n a_n^2 \, \mathbb{E}[\|H_n\|_{\mathcal{G}}^2] < \infty, \\ &\mathbb{E}\big[\, V_n \,|\, \mu_1, H_1, V_1, \ldots, H_{n-1}, V_{n-1} \,\big] = 0, \quad \text{and} \quad \sum_n a_n^2 \, \mathbb{E}[\|V_n\|_{\mathcal{G}}^2] < \infty. \end{aligned} \tag{126}$$

Then, $\mu_n \to \mu$ $(n \to \infty)$ almost surely.

We applied the following forms of the fundamental topological results from (Nagy, 2001, Lemma 5.1) and (Nagy, 2001, Theorem 5.1) to prove Theorem 21.

Lemma 25 *Let $(\mathbb{X}, \mathcal{T})$ be a locally compact Hausdorff space. Let K be a compact subset of $\mathbb{X}$, and let U be an open set, with $K \subseteq U$. Then, there exists an open set E such that $\bar{E}$ is compact and $K \subseteq E \subseteq \bar{E} \subseteq U$.*

Lemma 26 (Urysohn) *Let $(\mathbb{X}, \mathcal{T})$ be a locally compact Hausdorff space. Let K and F be disjoint subsets of $\mathbb{X}$, where K is compact and F is closed. Then, there exists a continuous function $f : \mathbb{X} \to [0, 1]$ such that $f|_K = 1$ and $f|_F = 0$.*

We also used the following well-known facts from real analysis, see (Dudley, 2002, Corollary 2.4.6), (Rudin, 1976, Theorem 4.15) and (Ruder, 1966, Theorem 2)

Theorem 27 (Heine-Cantor) *A continuous function from a compact metric space to any metric space is uniformly continuous.*

Theorem 28 (Weierstrass) *If f is a continuous mapping of a compact metric space $\mathbb{X}$ into $\mathbb{R}$, then f is bounded.*

Lemma 29 (Toeplitz) *Let $(a_{n,i})_{n,i\in\mathbb{N}}$ be a doubly indexed array of real numbers such that*

$$\sup_{n\in\mathbb{N}} \sum_{i=1}^{\infty} |a_{n,i}| < \infty, \qquad \text{and} \qquad \lim_{n\to\infty} a_{n,i} = 0, \tag{127}$$

for any $i \in \mathbb{N}$. If $x_n \to x$ as $n \to \infty$, then $\lim_{n\to\infty} \sum_{i=1}^{\infty} a_{n,i} x_i = 0$.

We applied a simplified version of the theorem of Robbins and Siegmund (1971, Theorem 1) to prove the strong consistency of our recursive algorithm.

Theorem 30 (almost supermartingale convergence) *Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$ be a filtration, i.e., a nondecreasing sequence of σ-algebras. For every $n \in \mathbb{N}$ let V_n, α_n and β_n be non-negative $\mathcal{F}_n$ measurable random variables such that*

$$\mathbb{E}[\, V_{n+1} \,|\, \mathcal{F}_n \,] \;\leq\; (1 + \alpha_n) V_n + \beta_n \tag{128}$$

holds almost surely. If $\sum_{n=1}^{\infty} \mathbb{E}[\alpha_n] < \infty$ and $\sum_{n=1}^{\infty} \mathbb{E}[\beta_n] < \infty$ hold almost surely, then $(V_n)_{n\in\mathbb{N}}$ converges almost surely to a finite limit.

The proof of Corollary 14 used the following theorem (Lee, 2006, Theorem 6.13).

Theorem 31 (Hopf-Rinow) *A connected Riemannian manifold is geodesically complete if and only if it is complete as a metric space.*

We used the Lipschitzness of smooth maps with bounded derivatives between Riemannian manifolds w.r.t. the geodesical distance. A published version of the following proof was not found by us, however, the argument is straightforward.

Lemma 32 *Let (M, g) and (N, h) be Riemannian manifolds and f a smooth $M \to N$ map. If there exists $C > 0$ such that $\|Df(x)\| \leq C$, for all $x \in M$, where $\|\cdot\|$ denotes the operator norm, then for all $x, y \in M$ one has*

$$d_N(f(x), f(y)) \leq C\, d_M(x, y), \tag{129}$$

where d_M and d_N are the geodesical distances.

Proof Let $x, y \in M$. For all $\varepsilon > 0$ there exists a smooth curve $\gamma : [0, 1] \to M$ such that $\gamma(0) = x$ and $\gamma(1) = y$ and

$$\int_0^1 \|\gamma'(t)\|_g \leq d_M(x, y) + \varepsilon. \tag{130}$$

Observe that $f \circ \gamma$ is a smooth curve from $f(x)$ to $f(y)$ in N, hence

$$\begin{aligned} d_N(f(x), f(y)) &\leq \int_0^1 \|(f \circ \gamma)'(t)\|_h \,\mathrm{d}t = \int_0^1 \|Df(\gamma(t))(\gamma'(t))\|_h \,\mathrm{d}t \\ &\leq \int_0^1 \|Df(\gamma(t))\|\, \|\gamma'(t)\|_g \,\mathrm{d}t \leq C\,(d_M(x, y) + \varepsilon) \end{aligned} \tag{131}$$

holds. It is true for all $\varepsilon > 0$ thus $d_N(f(x), f(y)) \leq C \cdot d_M(x, y)$ also holds. ■

References

Nachman Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2004.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Denis Bosq and Michel Delecroix. Nonparametric prediction of a Hilbert space valued random variable. *Stochastic Processes and their Applications*, 19(2):271–280, 1985.

Luc Brogat-Motte, Alessandro Rudi, Céline Brouard, Juho Rousu, and Florence d'Alché Buc. Vector-valued least-squares regression under output regularity assumptions. *Journal of Machine Learning Research*, 23(344):1–50, 2022.

Balázs Csanád Csáji and Ambrus Tamás. Semi-parametric uncertainty bounds for binary classification. IEEE Conference on Decision and Control (CDC), pages 4427–4432, 2019.

Sophie Dabo-Niang and Noureddine Rhomari. Kernel regression estimation in a Banach space. *Journal of Statistical Planning and Inference*, 139(4):1421–1434, 2009.

Luc Devroye. On the almost everywhere convergence of nonparametric regression function estimates. *Annals of Statistics*, pages 1310–1319, 1981.

Nicolae Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*, volume 48. John Wiley & Sons, 2000.

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

Lawrence C Evans. *Partial Differential Equations*, volume 19. American Mathematical Soc., 2010.

Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*, volume 76. Springer, 2006.

Liliana Forzani, Ricardo Fraiman, and Pamela Llop. Consistent nonparametric regression for functional data under the Stone–Besicovitch conditions. *IEEE Transactions on Information Theory*, 58(11):6697–6708, 2012.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. Advances in Neural Information Processing Systems, pages 489–496, 2007.

Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

Arthur Gretton and László Györfi. Nonparametric independence tests: Space partitioning and kernel approaches. International Conference on Algorithmic Learning Theory, pages 183–198, 2008.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. Advances in Neural Information Processing Systems, pages 585–592, 2008.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors. International Conference on Machine Learning, pages 1803–1810, 2012.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Massi Pontil, and Arthur Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. International Conference on Machine Learning, pages 535–542, 2012.

László Györfi and Roi Weiss. Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. *Journal of Machine Learning Research*, 22(151): 1–25, 2021.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal Bayes consistency in metric spaces. Information Theory and Applications Workshop, pages 1–33, 2020.

Tuomas Hytönen, Jan Van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach Spaces*, volume 12. Springer, 2016.

Olav Kallenberg. *Foundations of Modern Probability*. Springer Science & Business Media, second edition, 2002.

Ilja Klebanov, Ingmar Schuster, and Timothy John Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

Harold J Kushner and G George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Science & Business Media, second edition, 2003.

Peter D Lax. *Functional Analysis*, volume 55. John Wiley & Sons, 2002.

John M Lee. *Riemannian Manifolds: An Introduction to Curvature*. Springer Science & Business Media, 2006.

John M Lee. *Introduction to Smooth Manifolds*. Springer Science & Business Media, 2013.

Tong Lin and Hongbin Zha. Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):796–809, 2008.

Lennart Ljung, Georg Pflug, and Harro Walk. *Stochastic Approximation and Optimization of Random Systems*, volume 17. Birkhäuser, 2012.

Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

Gabriel Nagy. Real analysis, 2001. Electronic textbook at Kansas State University.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. Advances in Neural Information Processing Systems, pages 21247–21259, 2020.

Gilles Pisier. *Martingales in Banach Spaces*, volume 155. Cambridge University Press, 2016.

James O Ramsay and Bernhard W Silverman. *Functional Data Analysis*. Springer, 2005.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. Optimizing Methods in Statistics, pages 233–257. Elsevier, 1971.

Brian Ruder. The Silverman-Toeplitz theorem. Kansas State University, 1966.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Book Company, New York, 1976.

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill Book Company, New York, 1987.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

Bernhard Schölkopf, Krikamol Muandet, Kenji Fukumizu, Stefan Harmeling, and Jonas Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 25(4):755–766, 2015.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. International Conference on Algorithmic Learning Theory, pages 13–31, 2007.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. International Conference on Machine Learning, pages 961–968, 2009.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Arthur H Stone. Paracompactness and product spaces. *Bulletin of the American Mathematical Society*, 54(10):977–982, 1948.

Charles J Stone. Consistent nonparametric regression. *Annals of Statistics*, pages 595–620, 1977.

Ambrus Tamás and Balázs Csanád Csáji. Exact distribution-free hypothesis tests for the regression function of binary classification via conditional kernel mean embeddings. *IEEE Control Systems Letters*, 6:860–865, 2022.

Robert L Taylor. *Stochastic Convergence of Weighted Sums of Random Elements in Linear Spaces*, volume 672. Springer, 1978.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Ilya Tolstikhin, Bharath K Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. Conference on Uncertainty in Artificial Intelligence, pages 804–813, 2011.