Explicit Universal and Approximate-Universal Kernels on Compact Metric Spaces

Eloi Tanguy[1]

[1]Université Paris Cité, CNRS, MAP5, F-75006 Paris, France

May 24, 2025

Abstract

Universal kernels, whose Reproducing Kernel Hilbert Space is dense in the space of continuous functions are of great practical and theoretical interest. In this paper, we introduce an explicit construction of universal kernels on compact metric spaces. We also introduce a notion of approximate universality, and construct tractable kernels that are approximately universal.

Table of Contents

1 Introduction

1.1 Kernels in Practice and Related Works

Kernels Methods at large are a ubiquitous tool in statistics, starting with Kernel Density Estimation [Ros56; Par62] and Kernel Regression [Nad64; Wat64]. For an overview of the use of kernel methods in statistics and probability, we refer to the monograph [BT11]. In Machine Learning, the first uses of kernels hinged on the "kernel trick" [Aiz64; SSM98], which allows high expressivity of models without the need of an explicit feature map into the underlying infinite-dimensional space. A cornerstone model is the Support Vector Machine [CV95], whose statistical properties have garnered extensive attention, see for example the monograph [CS08]. A useful tool is the Kernel Mean embedding (we refer to the review [Mua+17]) which maps a measure μ to a point $M(\mu)$ in a Hilbert space of features, and can be used to compare measures with the Maximum Mean Discrepancy defined as $\mathrm{MMD}(\mu,\nu) = \|M(\mu) - M(\nu)\|_H$ which fostered numerous applications [Góm+09; Zha+12; Mua+12; FSG13; Gre+12; Dor+14; Li+17]. Theoretical guarantees for the MMD depending on properties of the kernel have been reviewed in [SFL11].

From a theoretical standpoint, Reproducing Kernel Hilbert Spaces (RKHS) introduced by Aronszajn [Aro50] have been the object of several monographs [SS02; CS08; SS16]. Some questions remain open, in particular constructing suitable kernels on non-euclidean metric spaces is a challenging problem that is the subject of ongoing research. For compact metric spaces, [CS10] show the existence of universal kernels (i.e. such that the associated RKHS is dense in the space of continuous functions) when the space is continuously embedded into a separable Hilbert space, and [SZ21] relate the notions of universality and strictly proper kernel scores. On complete Riemannian manifolds, [Jay+15] observe (Theorem 6.2) that the natural Gaussian kernel $k(x,y) = \exp(-s\ d(x,y)^2)$ is indeed a kernel only in the very restricted case where the manifold is isometric to $\mathbb{R}^d$. On Hilbert and Banach spaces, [ZGD22] introduce radial kernels and show universality-adjacent properties. Regarding universality, [MXZ06] study conditions on the feature maps that ensure universality.

Our contribution first consists in an *explicit* construction of universal kernels on a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$, in some sense extending [CS10] whose construction is not explicit and relied on the existence of an embedding. The constructed kernels use known kernels known as *Taylor* and *radial* kernels, which are defined on compact subsets of separable Hilbert spaces. Noticing that our kernels are not tractable in practice, we introduce a notion of *approximate universality* and construct other explicit kernels that are approximately universal and tractable.

1.2 Elements of RKHS Theory

For a set $\mathcal{X}$, a *kernel* $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a *positive-definite symmetric* function, which is to say a function that verifies $k(x,y) = k(y,x)$ and:

$$\forall n \in \mathbb{N}^*,\ \forall (x_1, \cdots, x_n) \in \mathcal{X}^n,\ \forall a \in \mathbb{R}^n,\ \sum_{i=1}^n \sum_{j=1}^n a_i k(x_i, x_j) a_j \geq 0.$$

By the Moore-Aronszajn theorem [Aro50], there exists a unique Hilbert space $(H, \langle \cdot, \cdot \rangle_H)$ of functions $\mathcal{X} \longrightarrow \mathbb{R}$, such that H contains all basic functions $k(\cdot, x)$, and its inner product is characterised by the "reproducing property"$\langle k(\cdot,x), k(\cdot,y)\rangle_H = k(x,y)$. Denoting by $\overline{\mathrm{Span}}$ the Hilbertian completion of the linear span of a set, it follows that $H = \overline{\mathrm{Span}}\{k(\cdot,x),\ x \in \mathcal{X}\}$. The space H is referred to as the Reproducing Kernel Hilbert Space (RKHS) associated to the kernel k. The reproducing property of the kernel implies that for any $h \in H$ and $x \in \mathcal{X}$, we have $\langle h, k(\cdot, x)\rangle_H = h(x)$.

If k is continuous (w.r.t. a metric on $\mathcal{X}$), the RKHS H is contained in the space of continuous functions from $\mathcal{X}$ to $\mathbb{R}$, denoted $\mathcal{C}(\mathcal{X})$. In this work, we will always consider continuous kernels. Some continuous kernels have an additional property called *universality*:

Definition 1. *A continuous kernel k on a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$ is said to be universal if the RKHS H is dense in $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$, the space of continuous functions from $\mathcal{X}$ to $\mathbb{R}$ equipped with the supremum norm. In other words, for any $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, there exists $h \in H$ such that $\|f - h\|_\infty \leq \varepsilon$.*

Another equivalent definition of kernels uses the notion of feature map / feature space pairs: through these lens, a kernel is any map $\mathcal{X}^2 \longrightarrow \mathbb{R}$ such that there exists a Hilbert space H_0 and a map $\Phi_0 : \mathcal{X} \longrightarrow H_0$ such that Eq. (1) holds.

$$\forall x, y \in \mathcal{X},\ k(x, y) = \langle \Phi_0(x), \Phi_0(y)\rangle_{H_0}. \tag{1}$$

The pair (Φ_0, H_0) is called a *feature map / feature space pair* (or simply *feature pair*) for k, and any kernel can be written in this form ([CS08], Theorem 4.16). The associated RKHS is then defined as:

$$H = \{x \longmapsto \langle h_0, \Phi_0(x)\rangle_{H_0},\ h_0 \in H_0\}. \tag{2}$$

The RKHS H in Eq. (2) is unique ([CS08], Theorem 4.21), and equal to $\overline{\mathrm{Span}}\,\{k(\cdot, x),\ x \in \mathcal{X}\}$ as stated above. The *canonical feature map* is defined as $\Phi(x) = k(\cdot, x)$, and the pair (Φ, H) is called the *canonical feature pair* for k.

From the space viewpoint, and RKHS can equivalently be defined as a Hilbert space of functions $\mathcal{X} \longrightarrow \mathbb{R}$ in which the evaluation $\delta_x : h \longmapsto h(x)$ is continuous for all $x \in \mathcal{X}$, as is done in [CS08], Section 4.2. The kernel is then defined as $k(x, y) = \langle L\delta_x, L\delta_y\rangle_H$, where $L\delta_x \in H$ is the Riesz representation of $\delta_x \in H'$. In this paper, we stick to the (equivalent) kernel viewpoint.

For a compact metric space (E, d_E), we will denote by $\mathrm{diam}(E)$ its diameter, which is defined by $\mathrm{diam}(E) := \max_{(x,y)\in E^2} d_E(x, y)$. Throughout this work, $\mathcal{X}$ will be assumed to be a compact metric space, and we denote $D_{\mathcal{X}} := \mathrm{diam}(\mathcal{X})$.

The first type of universal kernels of interest in this work are Taylor kernels (see [CS08] Lemma 4.8 and Corollary 4.57 for their study on compact subsets of $\mathbb{R}^d$).

Definition 2. *Let $W \subset \ell^2$ be a non-empty compact set and $D_W^2 := \mathrm{diam}(W)^2 > 0$ the square of its diameter. Take a sequence $(a_n)_{n\in\mathbb{N}} \in (0, +\infty)^{\mathbb{N}}$ such that $K(t) := \sum_n a_n t^n$ converges absolutely on $[-D_W^2, D_W^2]$. The Taylor kernel associated to K is the map*

$$k_W := \left\{ \begin{array}{ccc} W^2 & \longrightarrow & \mathbb{R} \\ (u, v) & \longmapsto & K(\langle u, v\rangle_{\ell^2}) \end{array} \right. . \tag{3}$$

Taylor kernels are shown to be universal on compact subsets of ℓ^2 in [CS10] Theorem 2.1. The second type of universal kernels we will consider are radial kernels[1].

Definition 3. *Let $W \subset \ell^2$ be a non-empty compact set and $\mu \in \mathcal{M}([0, +\infty))$ a finite Borel measure on $[0, +\infty)$ with $\mathrm{supp}(\mu) \neq \{0\}$. The associated radial function K and the radial kernel k_W are defined as follows:*

$$K := \left\{ \begin{array}{ccc} \mathbb{R}_+ & \longrightarrow & \mathbb{R} \\ t & \longmapsto & \int_0^{+\infty} e^{-st}\mathrm{d}\mu(s) \end{array} \right. ,\ k_W := \left\{ \begin{array}{ccc} W^2 & \longrightarrow & \mathbb{R} \\ (u, v) & \longmapsto & K(\|u - v\|_{\ell^2}^2) \end{array} \right. . \tag{4}$$

[1]Radial kernels can be defined (and shown to be universal) on separable Hilbert spaces and more [ZGD22], but we will use compactness for other reasons, and thus restrict to compact subsets of ℓ^2 for our purposes.

The universality of radial kernels on W is a consequence of [ZGD22] Proposition 5.2 combined with [SZ21] Theorem 3.13. Note that the well-known Gaussian (or RBF) kernel $\exp(-\| \cdot - \cdot \|_{\ell^2}^2/(2\sigma^2))$ is a particular radial kernel with $\mu := \delta_{1/(2\sigma^2)}$.

1.3 Paper Outline and Contributions

The objective of this paper is to construct kernels k on a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$ that are *universal* (see Definition 1). We also introduce a notion of approximate universality (Definition 6), and introduce other (tractable) explicit kernels $\hat{k}$ and k_t that verify this property.

Construction of universal kernels in Section 2 To construct universal kernels on $\mathcal{X}$, we first introduce an explicit continuous injection $\varphi : \mathcal{X} \longrightarrow \ell^2$ in Proposition 4. Given any universal kernel k_V on $V := \varphi(\mathcal{X}) \subset \ell^2$ we show in Theorem 5 that $k(x, y) := k_V(\varphi(x), \varphi(y))$, is universal on $\mathcal{X}$.

The construction of φ in Section 2 is based on a countable basis of $\mathcal{X}$, and the associated kernel requires inner products in ℓ^2. In Section 3, we explain how we can use instead a (finite) η-covering of $\mathcal{X}$, yielding a finite-dimensional approximation of the embedding φ, with theoretical guarantees. We also investigate the natural idea of truncating the sequence $\varphi(x)$.

Approximate universal kernels in Section 3 We introduce a notion of *approximate universal kernels* on $\mathcal{X}$, which are kernels $\hat{k}$ of RKHS $\hat{H}$ such that for all $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, there exists $\hat{h} \in \hat{H}$ such that $\|f - \hat{h}\|_\infty \leq \varepsilon + \rho(f)$, where $\rho(f) > 0$ is an error term depending on $\hat{k}$ and f. We construct a simpler map $\hat{\varphi} : \mathcal{X} \longrightarrow \mathbb{R}^J$ as a surrogate for the embedding $\varphi : \mathcal{X} :\longrightarrow \ell^2$, and embed $\mathbb{R}^J$ into ℓ^2 appropriately to compare φ and $B \circ \hat{\varphi}$. This allows us to introduce the kernel $\hat{k}(x, y) := k_W(B \circ \hat{\varphi}(x), B \circ \hat{\varphi}(y))$ for a compact set $W \supset \varphi(\mathcal{X}) \cup B(\hat{\varphi}(\mathcal{X}))$ and a Taylor or radial kernel k_W on W. In Corollary 10, we provide a tractable (as in numerically computable) expression for $\hat{k}$. Finally, we show in Theorem 14 that $\hat{k}$ is an approximate universal kernel on $\mathcal{X}$ with an explicit error term ρ depending on discretisation parameters and the "complexity" of the function f. In Section 3.3, we introduce a simple truncation of φ which leads to another approximate universal kernel $\hat{k}_t$ on $\mathcal{X}$.

2 Explicit Universal Taylor and Radial Kernels on a Compact Metric Space

2.1 Injection of $\mathcal{X}$ into ℓ^2

Let $(\mathcal{X}, d_{\mathcal{X}})$ be a non-empty compact metric space, and let $D_{\mathcal{X}} > 0$ be its diameter. We take a *basis* of $\mathcal{X}$, i.e. a countable sequence $(x_n)_{n \in \mathbb{N}}$ such that for any $x \in \mathcal{X}$ and $\varepsilon > 0$, there exists $n \in \mathbb{N}$ such that $d_{\mathcal{X}}(x, x_n) \leq \varepsilon$. Using a basis, we construct an implicit continuous injection φ from $\mathcal{X}$ into ℓ^2 (Proposition 4), then use universal kernels on $V := \varphi(\mathcal{X})$ to build a universal kernel k on $\mathcal{X}$ in Theorem 5. In Fig. 1, we illustrate the injection.

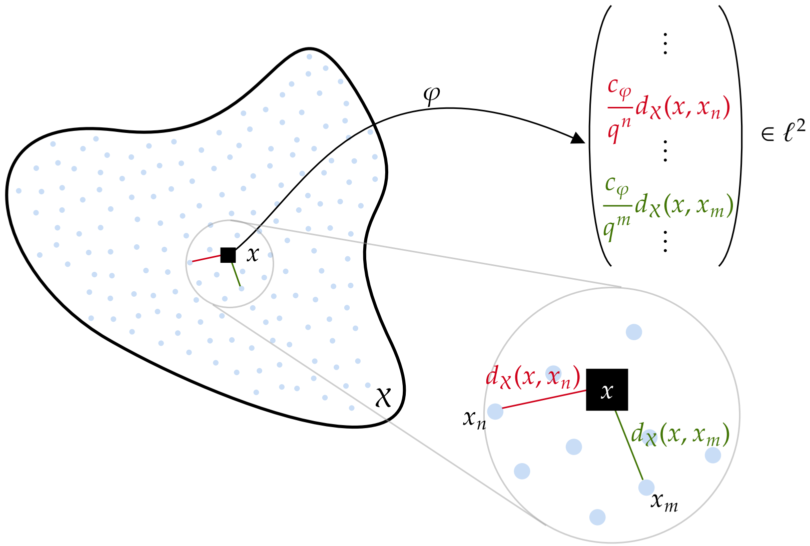

Figure 1: Given a basis $(x_n)_{n\in\mathbb{N}}$ of $\mathcal{X}$, the mapping $\varphi : \mathcal{X} \longrightarrow \ell^2$ maps a point $x \in \mathcal{X}$ to the sequence of its distances to the points of the basis.

Proposition 4. *Let (x_n) a basis of $\mathcal{X}$ and $q > 1$. The map*

$$\varphi := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & \ell^2 \\ x & \longmapsto & \left(\dfrac{c_\varphi d_{\mathcal{X}}(x, x_n)}{q^n}\right)_{n\in\mathbb{N}} \end{array} \right. , \quad c_\varphi := \frac{\sqrt{q^2-1}}{q}$$

is 1-Lipschitz and injective.

Proof. The fact that $\varphi(\mathcal{X}) \subset \ell^2$ comes from the compactness of $\mathcal{X}$. Take now $x, y \in \mathcal{X}$:

$$\|\varphi(x) - \varphi(y)\|_{\ell^2}^2 = c_\varphi^2 \sum_{n=0}^{+\infty} \frac{|d_{\mathcal{X}}(x, x_n) - d_{\mathcal{X}}(y, x_n)|^2}{q^{2n}} \leq c_\varphi^2 \sum_{n=0}^{+\infty} \frac{d_{\mathcal{X}}(x, y)^2}{q^{2n}} = \frac{c_\varphi^2 q^2}{q^2-1} d_{\mathcal{X}}(x, y)^2,$$

showing 1-Lipschitzness. As for injectivity, consider $x \neq y \in \mathcal{X}^2$ and $\varepsilon := d_{\mathcal{X}}(x, y)/3 > 0$. Since (x_n) is a basis of $\mathcal{X}$, there exists $n \in \mathbb{N}$ such that $d_{\mathcal{X}}(x, x_n) \leq \varepsilon$. The triangle inequality then shows

$$d_{\mathcal{X}}(y, x_n) \geq \underbrace{d_{\mathcal{X}}(y, x)}_{=3\varepsilon} - \underbrace{d_{\mathcal{X}}(x, x_n)}_{\in[0,\varepsilon]} \geq 2\varepsilon,$$

and thus $|\underbrace{d_{\mathcal{X}}(y, x_n)}_{\geq 2\varepsilon} - \underbrace{d_{\mathcal{X}}(x, x_n)}_{\in[0,\varepsilon]}| \geq \varepsilon$, allowing us to conclude

$$\|\varphi(y) - \varphi(x)\|_{\ell^2}^2 \geq c_\varphi^2 \frac{|d_{\mathcal{X}}(y, x_n) - d_{\mathcal{X}}(x, x_n)|^2}{q^{2n}} \geq c_\varphi^2 \frac{\varepsilon^2}{q^{2n}} > 0.$$

□

2.2 Universal Kernels on $\mathcal{X}$

We can now build universal kernels $k : \mathcal{X}^2 \longrightarrow \mathbb{R}$ using φ and a universal kernel $k_V : V^2 \longrightarrow \mathbb{R}$ (for example Taylor or radial) on $V := \varphi(\mathcal{X})$. The technique follows closely that of [CS10] Theorem 2.2. Thanks to the 1-Lipschitzness of φ, we have $\mathrm{diam}(\varphi(\mathcal{X})) \leq \mathrm{diam}(\mathcal{X}) =: D_{\mathcal{X}}$.

Theorem 5. *Let $V := \varphi(\mathcal{X}) \subset \ell^2$ and $k_V : V^2 \longrightarrow \mathbb{R}$ be a universal kernel on V (e.g. Taylor as in Definition 2 or radial as in Definition 3). The kernel*

$$k := \left\{ \begin{array}{ccc} \mathcal{X}^2 & \longrightarrow & \mathbb{R} \\ (x, y) & \longmapsto & k_V(\varphi(x), \varphi(y)) \end{array} \right.$$

is universal on $\mathcal{X}$.

Proof. Since $\mathcal{X}$ is a compact metric space and ℓ^2 is Hausdorff, the co-restriction of φ to V denoted $\varphi_V : \mathcal{X} \longrightarrow V$ is a homeomorphism. Let H_V be the unique RKHS associated to the kernel k_V on V, and $\Phi_V : V \longmapsto H_V$ its canonical feature map (i.e. $\Phi_V(u) = k_V(\cdot, u)$). Since $k(x, y) = \langle \Phi_V \circ \varphi_V(x), \Phi_V \circ \varphi_V(y)\rangle_{H_V}$, the map $\Phi_V \circ \varphi_V$ and the space H_V are a feature pair for k. In the following, given a set F of functions and g a function, we write $F \circ g := \{f \circ g,\ f \in F\}$. By uniqueness ([CS08] Theorem 4.21), it follows that the RKHS H associated to k can be written

$$H = \{x \longmapsto \langle h_V, \Phi_V \circ \varphi_V(x)\rangle_{H_V},\ h_V \in H_V\} = H_V \circ \varphi_V,$$

where the second equality comes from the reproducing property: for any $x \in \mathcal{X}$ and $h_V \in H_V$, we have $h_V \circ \varphi_V(x) = \langle h_V, \Phi_V \circ \varphi_V(x)\rangle_{H_V}$. Since φ_V is a homeomorphism, we also have $\mathcal{C}(\mathcal{X}) = \mathcal{C}(V) \circ \varphi_V$. Now for $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, take $f_V := f \circ \varphi_V^{-1} \in \mathcal{C}(V)$. By universality of k_V, there exists $h_V \in H_V$ such that $\|f_V - h_V\|_\infty \leq \varepsilon$. Taking $h := h_V \circ \varphi_V$ yields $\|f - h\|_\infty \leq \varepsilon$, and as a result k is universal. □

Note that in Section 3, we will consider instead $k(x, y) = k_W(\varphi(x), \varphi(y))$, where k_W is universal on a compact set W containing $V := \varphi(\mathcal{X})$. By [CS08] Lemma 4.55 item iii), k_W restricted to V^2 remains universal, and therefore the result still holds.

A strictly convex functional on $\mathcal{P}(\mathcal{X})$ We consider the set $\mathcal{P}(\mathcal{X})$ of probability measures on $\mathcal{X}$. As a universal kernel, k is also *characteristic* (see [SFL11] and use the compactness of $\mathcal{X}$), which is to say that the map

$$M := \left\{ \begin{array}{ccc} \mathcal{P}(\mathcal{X}) & \longrightarrow & H \\ \mu & \longmapsto & \int_{\mathcal{X}} k(\cdot, x)\mathrm{d}\mu(x) \end{array} \right. ,$$

known as the *kernel mean embedding* [Sri+10], is injective. One can show that the map

$$F := \left\{ \begin{array}{ccc} \mathcal{P}(\mathcal{X}) & \longrightarrow & \mathbb{R}_+ \\ \mu & \longmapsto & \|M(\mu)\|_H^2 \end{array} \right.$$

is continuous with respect to the weak convergence of measures (apply [HC11] Theorem A.1 using that $x \longmapsto k(\cdot, x)$ is continuous and bounded). Furthermore, by linearity of M and strict convexity of $\|\cdot\|_H^2$, the function F is strictly convex. Note that the fact that M is injective is required to prove *strict* convexity.

3 Approximate Universal Kernels

In practice, the function φ introduced in Proposition 4 is not tractable, limiting the use of the kernels proposed in Theorem 5. We will now introduce a family of tractable kernels which are approximately universal on $\mathcal{X}$. Throughout this section, the kernels k_W on a compact subset W of ℓ^2 that we will consider are Taylor or radial (see Definitions 2 and 3). Our objective is to construct another kernel $\hat{k}$ with a simpler explicit mapping $\hat{\varphi} : \mathcal{X} \longrightarrow \mathbb{R}^J$, yielding an RKHS $\hat{H}$ which we will show to be approximately universal in the sense of Definition 6.

Definition 6. *Let $\hat{k} : \mathcal{X}^2 \longrightarrow \mathbb{R}$ a kernel on $\mathcal{X}$ of RKHS $\hat{H}$ and $\rho : \mathcal{C}(\mathcal{X}) \longrightarrow \mathbb{R}_+$ an error function. We say that $\hat{k}$ is an approximate universal kernel on $\mathcal{X}$ if for all $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, there exists $\hat{h} \in \hat{H}$ such that $\|f - \hat{h}\|_\infty \leq \varepsilon + \rho(f)$.*

3.1 Constructing a Smaller RKHS $\hat{H}$

In this section, we provide a principled method to "sub-sample" the sequence $\varphi(x)$ which approximates the distance sequence $(d_\mathcal{X}(x, x_n))_{n\in\mathbb{N}}$ by a finite number of distances $(d_\mathcal{X}(x, y_j))_{j\in[\![1,J]\!]}$ for a well-chosen family $(y_j) \in \mathcal{X}^J$. We illustrate this discretisation concept in Fig. 2.

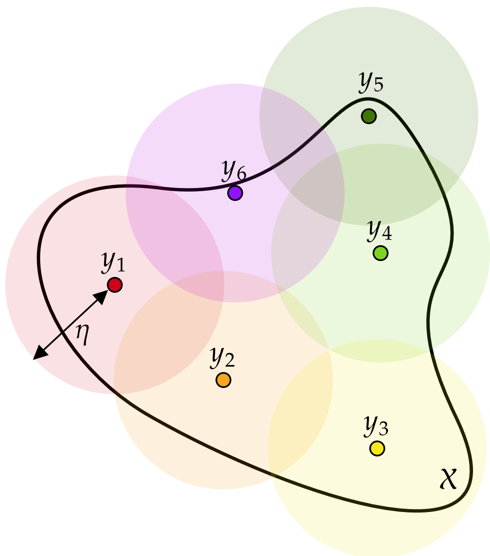

Figure 2: Discretisation of the space $\mathcal{X}$ into a cover of J balls of radius $\eta > 0$ centred at a $(y_j)_{j\in[\![1,J]\!]}$.

Instead of a basis of $\mathcal{X}$, we will now fix $\eta \in (0, D_\mathcal{X}]$ and consider $(y_j)_{j\in[\![1,J]\!]}$ a family of distinct points of $\mathcal{X}$ such that the family of balls $B_{d_\mathcal{X}}(y_j, \eta)$ covers $\mathcal{X}$. In Eq. (5), we introduce a map $\hat{\varphi} : \mathcal{X} \longrightarrow \mathbb{R}^J$ in the spirit of φ defined in Proposition 4, which we visualise in Fig. 3.

$$\hat{\varphi} := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R}^J \\ x & \longmapsto & \left(\dfrac{d_\mathcal{X}(x, y_j)}{\sqrt{J}}\right)_{j\in[\![1,J]\!]} \end{array} \right. . \tag{5}$$

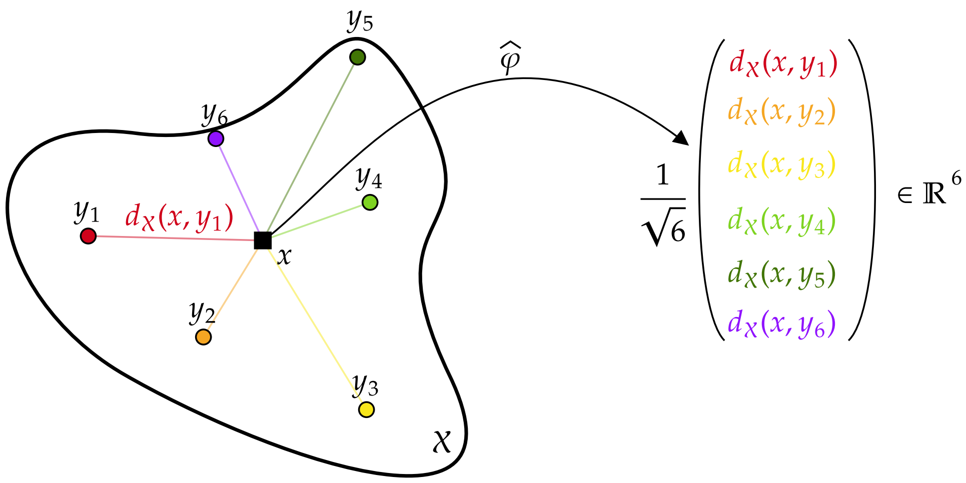

Figure 3: The mapping $\hat{\varphi} : \mathcal{X} \longrightarrow \mathbb{R}^J$ maps a point $x \in \mathcal{X}$ to the vector of distances between x and the centres y_j of the covering.

It is immediate to verify that $\hat{\varphi} : (\mathcal{X}, d_{\mathcal{X}}) \longrightarrow (\mathbb{R}^J, \|\cdot\|_2)$ is 1-Lipschitz, thanks to the $J^{-1/2}$ normalisation. In Proposition 9, we show how to embed $\mathbb{R}^J \supset \hat{\varphi}(\mathcal{X})$ into ℓ^2 with a mapping B, which will allow us to compare the RKHS induced by $\hat{\varphi}$ and a particular $\varphi : \mathcal{X} \longrightarrow \ell^2$. To construct B, we first begin with a geometric series separation lemma, which will be convenient to deal with the factor $\frac{1}{q^n}$ in φ.

Lemma 7. *Let $J \geq 2$, $q \in (1, 1 + \frac{1}{J-1})$ and coefficients $(\lambda_1, \cdots, \lambda_J) \in (0,1)^J$ such that $\sum_j \lambda_j = 1$, there exists $\alpha : \mathbb{N} \longrightarrow [\![1, J]\!]$ with for all $j \in [\![1, J]\!]$, $\alpha^{-1}(\{j\})$ infinite such that:*

$$\forall j \in [\![1, J]\!], \sum_{n \in \alpha^{-1}(\{j\})} \frac{1}{q^n} = \lambda_j \frac{q}{q-1}. \tag{6}$$

Proof. Set $S := \frac{q}{q-1}$. We will construct a sequence $(\alpha(N))_{N \in \mathbb{N}}$ by induction over N, verifying the property

$$P_N : \forall j \in [\![1, J]\!], \sum_{n \in [\![1, N]\!] : \alpha(n) = j} \frac{1}{q^n} < \lambda_j S.$$

Initialisation: set $\alpha(0)$ the first $j \in [\![1, J]\!]$ such that $1 < \lambda_j S$. Note that such a j exists, otherwise summing over $j \in [\![1, J]\!]$ yields

$$J \geq \frac{q}{q-1} > \frac{1 + \frac{1}{J-1}}{1 + \frac{1}{J-1} - 1} = J,$$

which is a contradiction. Having chosen $j \in [\![1, J]\!]$ (e.g. minimal) such that $1 \leq \lambda_j S$, we have defined $\alpha(0) := j$ verifying P_0.

Induction step: let $N \in \mathbb{N}$, suppose P_N true. We show that there exists $j \in [\![1, J]\!]$ such that

$$\sum_{n \in [\![1, N]\!] : \alpha(n) = j} \frac{1}{q^n} + \frac{1}{q^{N+1}} < \lambda_j S \tag{7}$$

by contradiction. If that were not the case, we would have by summing Eq. (7) over $j \in [\![1, J]\!]$:

$$\sum_{n=0}^{N+1} \frac{1}{q^n} + \frac{J-1}{q^{N+1}} \geq S,$$

which by computation is equivalent to $q \geq 1 + \frac{1}{J-1}$, obtaining a contradiction. Selecting $j \in [\![1, J]\!]$ such that Eq. (7) holds, we can set $\alpha(N+1) := j$ which satisfies P_{N+1}.

Now that $\alpha : \mathbb{N} \longrightarrow [\![1, J]\!]$ verifying (P_N) is constructed, we introduce the convergent series

$$\forall j \in [\![1, J]\!],\ \forall N \in \mathbb{N},\ S_N^{(j)} := \sum_{n \in [\![1,N]\!]:\alpha(n)=j} \frac{1}{q^n},\ S_\infty^{(j)} := \lim_{N \longrightarrow +\infty} S_N^{(j)}.$$

Thanks to (P_N) have for all $j \in [\![1, J]\!]$ taking the limit yields $S_\infty^{(j)} \leq \lambda_j S$, and summing over $j \in [\![1, J]\!]$ gives $\sum_j S_\infty^{(j)} = S$, hence necessarily for all $j \in [\![1, J]\!]$, $S_\infty^{(j)} = \lambda_j S$.

Finally, observing the strict inequality in P_N at each $N \in \mathbb{N}$ shows that $\alpha^{-1}(\{j\})$ has to be infinite, concluding the proof. □

We now turn to constructing an embedding $B : \mathbb{R}^J \longrightarrow \ell^2$, which will allow us to compare $\hat{\varphi}$ and φ. An important property of B will be the correspondence between the inner products in $\mathbb{R}^J$ and ℓ^2 (i.e. B will be an isometry). The construction of this embedding revolves around the construction of an adapted basis $(x_n)_{n\in\mathbb{N}}$ of $\mathcal{X}$ which is balanced with respect to the covering by the balls $B(y_j, \eta)$, as illustrated in Fig. 4.

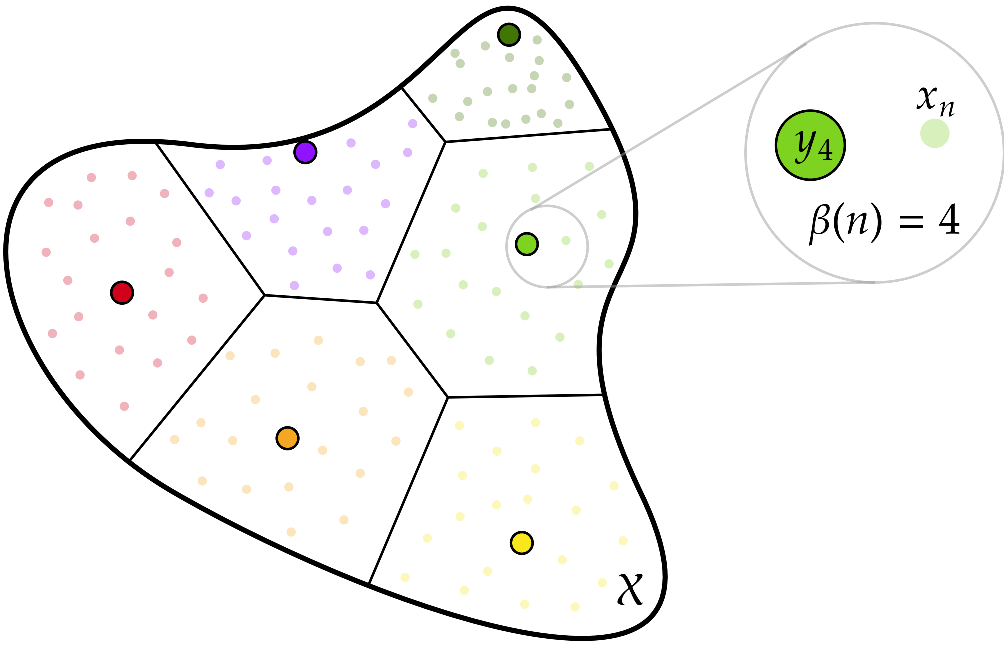

Figure 4: The basis $(x_n)_{n\in\mathbb{N}}$ is such that there equally as many (x_n) in each region $\mathcal{X}_j$ of points closest to y_j. In the figure, we observe a zoom on the region $\mathcal{X}_4$, where the example point x_n is closest to y_4. In mathematical terms, we write this property as $\beta(n) = y_j$, and in Proposition 8 we will construct (x_n) such that the sum $\sum_n q^{-2n}$ is split evenly between the sets $\beta^{-1}(\{j\})$.

Proposition 8. *Take* $q \in (1, \sqrt{1 + \frac{1}{J-1}})$. *There exists a basis* $(x_n)_{n\in\mathbb{N}}$ *of* $\mathcal{X}$ *and a mapping* $\beta : \mathbb{N} \longrightarrow [\![1, J]\!]$ *with infinite pre-images which verifies* $\forall n \in \mathbb{N}$, $d_{\mathcal{X}}(x_n, y_{\beta(n)}) = \min_j d_X(x_n, y_j)$, *and with the following property:*

$$\forall j \in [\![1, J]\!], \quad \sum_{n\in\beta^{-1}(\{j\})} \frac{1}{q^{2n}} = \frac{1}{J}\frac{q^2}{q^2-1}. \tag{8}$$

Proof. Consider for $j \in [\![1, J]\!]$ the set $\mathcal{X}_j := \{x \in \mathcal{X} : \operatorname{argmin}_m \ d_{\mathcal{X}}(x, y_m) = j\}$ (with disambiguation by taking the smallest minimiser if multiple exist). By definition, the sets $\mathcal{X}_j$ are disjoint and cover $\mathcal{X}$. Since $(\mathcal{X}, d_{\mathcal{X}})$ is a compact metric space, each subset $\mathcal{X}_j$ is separable, allowing us to choose a basis $(z_n^{(j)})_{n\in\mathbb{N}}$ of $\mathcal{X}_j$ for each $j \in [\![1, J]\!]$. By Lemma 7, we can choose $\beta : \mathbb{N} \longrightarrow [\![1, J]\!]$ with infinite pre-images which verifies Eq. (6). Since for each $j \in [\![1, J]\!]$, the set $\beta^{-1}(\{j\}) \subset \mathbb{N}$ is infinite, we can choose $\omega_j : \beta^{-1}(\{j\}) \longrightarrow \mathbb{N}$ a bijection. We can now define $\forall n \in \mathbb{N}, \ x_n := z^{(\beta(n))}_{\omega_{\beta(n)}(n)}$, which is a basis of $\mathcal{X}$ since $\cup_j \mathcal{X}_j = \mathcal{X}$ and

$$\{x_n\}_{n\in\mathbb{N}} = \bigcup_j \{z^{(j)}_{\omega_j(m)}\}_{m\in\beta^{-1}(\{j\})} = \bigcup_j \{z_n^{(j)}\}_{n\in\mathbb{N}},$$

by construction. Furthermore, by definition, we have $\forall n \in \mathbb{N}, \ \operatorname{argmin}_j \ d_{\mathcal{X}}(x_n, y_j) = \beta(n)$, which shows that the mapping β satisfies the desired properties. $\square$

Using the adapted basis from Proposition 8, we can finally construct an isometry $B : \mathbb{R}^J \longrightarrow \ell^2$:

Proposition 9. *Take a basis (x_n) of $\mathcal{X}$ and $\beta : \mathbb{N} \longrightarrow [\![1, J]\!]$ as in Proposition 8. The mapping B defined below is an isometry:*

$$B := \left\{ \begin{array}{ccc} \mathbb{R}^J & \longrightarrow & \ell^2 \\ (u_j)_{j=1}^J & \longmapsto & \left(c_B \dfrac{u_{\beta(n)}}{q^n}\right)_{n\in\mathbb{N}} \end{array} \right. , \ c_B := \frac{\sqrt{J(q^2-1)}}{q}. \tag{9}$$

Proof. The mapping B is clearly linear, and for $u, v \in \mathbb{R}^J$ we compute using Eq. (8):

$$\langle B(u), B(v)\rangle_{\ell^2} = \sum_{n=0}^{+\infty} \frac{c_B^2}{q^{2n}} u_{\beta(n)} v_{\beta(n)} = c_B^2 \sum_{j=1}^J u_j v_j \sum_{n\in\beta^{-1}(\{j\})} \frac{1}{q^{2n}} = c_B^2 \frac{1}{J}\frac{q^2}{q^2-1} \langle u, v\rangle_{\mathbb{R}^J} = \langle u, v\rangle_{\mathbb{R}^J},$$

which shows that B is an isometry. $\square$

In the following, we draw a correspondence between a RKHS $\hat{H}$ built with $\hat{\varphi}$ from Eq. (5) and another RKHS H built using φ from Proposition 4. Let $U := \hat{\varphi}(\mathcal{X})$, which is a compact subset of $\mathbb{R}^J$, then let $\hat{V} := B(U)$, it is a compact subset of ℓ^2. Consider the injection φ introduced in Proposition 4 with basis (x_n) and scale q as in Proposition 9. Define $V := \varphi(\mathcal{X})$, $W := V \cup \hat{V}$, which are also compact subsets of ℓ^2. We now summarise our objects in the following diagram:

$$\begin{tikzcd} \mathcal{X} \arrow[r, "\varphi"] \arrow[d, "\hat{\varphi}"] & V \subset W \subset \ell^2 \\ U \subset \mathbb{R}^J \arrow[r, "B"'] & \hat{V} \subset W \subset \ell^2 \end{tikzcd} \tag{10}$$

We fix a kernel $k_W : W^2 \longrightarrow \mathbb{R}$ which is of Taylor type or radial (see Definitions 2 and 3) and thus in particular universal on W, and introduce its canonical feature map:

$$\Phi_W := \left\{ \begin{array}{ccc} W & \longrightarrow & H_W \\ u & \longmapsto & k_W(\cdot, u) \end{array} \right. , \tag{11}$$

where $H_W = \overline{\mathrm{Span}}\,\{k_W(\cdot, u),\ u \in W\} \subset \mathcal{C}(W)$ is the unique RKHS associated to the kernel k_W ([CS08] Theorem 4.21). Consider the kernels $k, \hat{k}$ on $\mathcal{X}$ defined respectively as:

$$k := \left\{ \begin{array}{ccc} \mathcal{X}^2 & \longrightarrow & \mathbb{R} \\ (x,y) & \longmapsto & k_W(\varphi(x), \varphi(y)) \end{array} \right. , \quad \hat{k} := \left\{ \begin{array}{ccc} \mathcal{X}^2 & \longrightarrow & \mathbb{R} \\ (x,y) & \longmapsto & k_W(B \circ \hat{\varphi}(x), B \circ \hat{\varphi}(y)) \end{array} \right. . \tag{12}$$

By definition of the feature pair (H_W, Φ_W) for k_W, we observe that for $x, y \in \mathcal{X}$:

$$k(x,y) = \langle \Phi_W \circ \varphi(x), \Phi_W \circ \varphi(y) \rangle_{H_W}, \ \hat{k}(x,y) = \langle \Phi_W \circ B \circ \hat{\varphi}(x), \Phi_W \circ B \circ \hat{\varphi}(y) \rangle_{H_W}. \tag{13}$$

The RKHS spaces $H, \hat{H}$ associated to $k, \hat{k}$ are both subspaces of $\mathcal{C}(\mathcal{X})$ and can be written with the following respective feature pairs $(H_W, \Phi), (H_W, \hat{\Phi})$ (use Eq. (13) with [CS08] Theorem 4.21):

$$H = \{x \longmapsto \langle h_W, \Phi_W \circ \varphi(x) \rangle_{H_W},\ h_W \in H_W\}, \ \Phi := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & H_W \\ x & \longmapsto & \Phi_W \circ \varphi(x) \end{array} \right. \tag{14}$$

$$\hat{H} = \{x \longmapsto \langle h_W, \Phi_W \circ B \circ \hat{\varphi}(x) \rangle_{H_W},\ h_W \in H_W\}, \ \hat{\Phi} := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & H_W \\ x & \longmapsto & \Phi_W \circ B \circ \hat{\varphi}(x) \end{array} \right. . \tag{15}$$

Notice that the feature space H_W is shared. To finish the diagram, we introduce the "feature-to-map" functionals:

$$\Psi := \left\{ \begin{array}{ccc} H_W & \longrightarrow & H \\ h_W & \longmapsto & x \mapsto \langle h_W, \Phi(x) \rangle_{H_W} \end{array} \right. , \ \hat{\Psi} := \left\{ \begin{array}{ccc} H_W & \longrightarrow & \hat{H} \\ h_W & \longmapsto & x \mapsto \langle h_W, \hat{\Phi}(x) \rangle_{H_W} \end{array} \right. . \tag{16}$$

Extending the diagram in Eq. (10), we obtain:

$$\begin{tikzcd}
 & & V \subset W \subset \ell^2 \arrow[dr, "\Phi_W"] & & H \\
\mathcal{X} \arrow[urr, "\varphi"] \arrow[dr, "\hat{\varphi}"'] \arrow[rrr, bend left=60, "\Phi"] \arrow[rrr, bend right=60, "\hat{\Phi}"'] & & & H_W \arrow[ur, "\Psi"] \arrow[dr, "\hat{\Psi}"'] & \\
 & U \subset \mathbb{R}^J \arrow[r, "B"'] & \hat{V} \subset W \subset \ell^2 \arrow[ur, "\Phi_W"'] & & \hat{H}
\end{tikzcd} \tag{17}$$

Using the inner product correspondence induced by the isometry B from Proposition 9, a tractable formula for $\hat{k}$ is obtained immediately for Taylor and radial kernels.

Corollary 10. *The kernel $\hat{k}$ on $\mathcal{X}$ is given by, for $\forall x, y \in \mathcal{X}$:*

- *if k_W is a Taylor kernel (Definition 2):*

$$\hat{k}(x,y) = K(\langle \hat{\varphi}(x), \hat{\varphi}(y) \rangle_{\mathbb{R}^J}) = \sum_{n=0}^{+\infty} a_n \left(\frac{1}{J} \sum_{j=1}^{J} d_{\mathcal{X}}(x, y_j) d_{\mathcal{X}}(y, y_j) \right)^n ; \tag{18}$$

- *if k_W is a radial kernel (Definition 3):*

$$\hat{k}(x,y) = K(\|\hat{\varphi}(x) - \hat{\varphi}(y)\|^2_{\mathbb{R}^J}) = \int_0^{+\infty} \exp\left(-\frac{s}{J}\sum_{j=1}^J (d_{\mathcal{X}}(x,y_j) - d_{\mathcal{X}}(y,y_j))^2\right) \mathrm{d}\mu(s). \tag{19}$$

Proof. Let $x, y \in \mathcal{X}$, we remind that from Eq. (12) that $\hat{k}(x,y) := k_W(B \circ \hat{\varphi}(x), B \circ \hat{\varphi}(y))$. Now by Proposition 9, B is an isometry, yielding:

$$\langle B \circ \hat{\varphi}(x), B \circ \hat{\varphi}(y)\rangle_{\ell^2} = \langle \hat{\varphi}(x), \hat{\varphi}(y)\rangle_{\mathbb{R}^J}; \quad \|B \circ \hat{\varphi}(x) - B \circ \hat{\varphi}(y)\|^2_{\ell^2} = \|\hat{\varphi}(x) - \hat{\varphi}(y)\|^2_{\mathbb{R}^J}.$$

Eqs. (18) and (19) are then obtained by replacing k_W and K by their definitions in the Taylor and radial cases. □

We refer to the expressions in Eqs. (18) and (19) as "tractable" since they can be computed explicitly on a computer or approximated efficiently to numerical precision (note that the measure μ in the radial kernel can be discrete with finite support).

3.2 Showing that $\hat{H}$ is Approximately Universal

In this section, we show that the RKHS $\hat{H}$ introduced in Section 3.1 is approximately universal on $\mathcal{X}$. We use the notation and objects introduced in Section 3.1 extensively. The first approximation result we will show concerns a comparison in ℓ^2 between $\varphi(x)$ and $B \circ \hat{\varphi}(x)$:

Proposition 11. *For $x \in \mathcal{X}$, we have $\|\varphi(x) - B \circ \hat{\varphi}(x)\|_{\ell^2} \leq \eta$. The diameter of W verifies $D_W := \mathrm{diam}(W) \leq D_{\mathcal{X}}$.*

Proof. Let $n \in \mathbb{N}$, we look at the terms of the sequences $\varphi(x), B \circ \hat{\varphi}(x) \in W \subset \ell^2$:

$$|[\varphi(x)]_n - [B \circ \hat{\varphi}(x)]_n| = \left|\frac{c_\varphi d_{\mathcal{X}}(x,x_n)}{q^n} - \frac{c_B d_{\mathcal{X}}(x, y_{\beta(n)})}{\sqrt{J} q^n}\right| \leq \frac{1}{q^n} c_\varphi d_{\mathcal{X}}(x_n, y_{\beta(n)}).$$

By construction of the covering $(B_{d_{\mathcal{X}}}(y_j, \eta))_j$ and of β (see Proposition 8), $d_{\mathcal{X}}(x_n, y_{\beta(n)}) \leq \eta$. Summing the squares over $n \in \mathbb{N}$ and replacing c_φ with its definition yields:

$$\|\varphi(x) - B \circ \hat{\varphi}(x)\|^2_{\ell^2} \leq \eta^2 \sum_{n=0}^{+\infty} \frac{c_\varphi^2}{q^{2n}} = \eta^2.$$

For the diameter of $W := \varphi(\mathcal{X}) \cup B \circ \hat{\varphi}(\mathcal{X})$, we have by 1-Lipschitzness of $\varphi, \hat{\varphi}$ and B: $\mathrm{diam}(\varphi(\mathcal{X})) \leq D_{\mathcal{X}}$ and $\mathrm{diam}(B \circ \hat{\varphi}(\mathcal{X})) \leq D_{\mathcal{X}}$. Using the inequality in the above display and the fact that $\eta \leq D_{\mathcal{X}}$, we conclude:

$$D_W = \max(\mathrm{diam}(\varphi(\mathcal{X})), \sup_{x \in \mathcal{X}} \|\varphi(x) - B \circ \hat{\varphi}(x)\|_{\ell^2}) = D_{\mathcal{X}}.$$

□

Using regularity properties of Taylor and radial kernels, we will show that the kernel $\hat{k}$ is approximately universal on $\mathcal{X}$ by relating it to k which is universal by Theorem 5. First, we see in Lemma 12 that the canonical feature map Φ_W is Hölder-continuous for Taylor kernels, and Lipschitz for radial kernels. We introduce the radius of W: $R_W := \max_{w \in W} \|w\|_{\ell^2}$. Using the definition of W and of $\varphi, \hat{\varphi}$ and B with their well-chosen normalisations, it is easy to see that $R_W \leq D_{\mathcal{X}}$.

Lemma 12. *The feature map $\Phi_W : (W, \|\cdot\|_{\ell^2}) \longrightarrow (H_W, \|\cdot\|_{H_W})$ has the following regularity:*

- *If k_W is a Taylor kernel, then Φ_W is $\frac{1}{2}$-Hölder continuous:*

$$\forall u, v \in W,\ \|\Phi_W(u) - \Phi_W(v)\|_{H_W} \leq \sqrt{2D_{\mathcal{X}}C_{K'}}\|u - v\|_{\ell^2}^{\frac{1}{2}},$$

where $C_{K'} := \max_{t\in[-D_{\mathcal{X}}^2, D_{\mathcal{X}}^2]} |K'(t)|$.

- *If k_W is a radial kernel, then Φ_W is $\sqrt{2C_{K'}}$-Lipschitz:*

$$\forall u, v \in W,\ \|\Phi_W(u) - \Phi_W(v)\|_{H_W} \leq \sqrt{2C_{K'}}\|u - v\|_{\ell^2},$$

where $C_{K'} := \max_{t\in[0, D_{\mathcal{X}}^2]} |K'(t)|$.

Proof. First, we remind that by Proposition 11, we have $\mathrm{diam}(W) \leq D_{\mathcal{X}}$. For the proof, we take inspiration from [Fie23] Section 4.2. Using the reproducing property, we begin computations for both kernel types, letting $u, v \in W$:

$$\begin{aligned}\|\Phi_W(u) - \Phi_W(v)\|_{H_W}^2 &= k_W(u, u) - 2k_W(u, v) + k_W(v, v) \\ &\leq |k_W(u, u) - k_W(u, v)| + |k_W(v, v) - k_W(u, v)|.\end{aligned}$$

For Taylor kernels, we use the fact that K is $C_{K'}$-Lipschitz on $[-D_{\mathcal{X}}^2, D_{\mathcal{X}}^2]$ and the Cauchy-Schwarz inequality for $\langle\cdot,\cdot\rangle_{\ell^2}$:

$$\begin{aligned}\|\Phi_W(u) - \Phi_W(v)\|_{H_W}^2 &\leq C_{K'}\left(|\langle u, u\rangle_{\ell^2} - \langle u, v\rangle_{\ell^2}| + |\langle v, v\rangle_{\ell^2} - \langle u, v\rangle_{\ell^2}|\right) \\ &\leq C_{K'}(\|u\|_{\ell^2} + \|v\|_{\ell^2})\|u - v\|_{\ell^2} \\ &\leq 2R_W C_{K'}\|u - v\|_{\ell^2},\end{aligned}$$

and we conclude using $R_W \leq D_{\mathcal{X}}$. For radial kernels, we use the fact that K is $C_{K'}$-Lipschitz on $[0, D_W^2]$ (we remind that K is non-increasing on $[0, +\infty)$):

$$\|\Phi_W(u) - \Phi_W(v)\|_{H_W}^2 = 2(K(0) - K(\|u - v\|_{\ell^2}^2)) \leq 2C_{K'}\|u - v\|_{\ell^2}^2.$$

□

We now use Lemma 12 to approximate any $h \in H$ with a $\hat{h} \in \hat{H}$ with a certain error, which we approach by comparing the feature-to-map functionals Ψ and $\hat{\Psi}$ from Eqs. (14) and (15).

Proposition 13. *For $h \in H$, take $h_W \in H_W$ such that $h = x \longmapsto \langle h_W, \Phi(x)\rangle_{H_W} = \Psi(h_W)$. Then let $\hat{h} := x \longmapsto \langle h_W, \hat{\Phi}(x)\rangle_{H_W} = \hat{\Psi}(h_W)$. Denoting $\|\cdot\|_\infty$ the supremum norm on $\mathcal{X}$, we have:*

$$\|h - \hat{h}\|_\infty \leq \rho_0\|h_W\|_{H_W}, \tag{20}$$

where $\rho_0 = \eta^{\frac{1}{2}}\sqrt{2D_{\mathcal{X}}C_{K'}}$ for a Taylor kernel and $\rho_0 = \eta\sqrt{2C_{K'}}$ for a radial kernel..

Proof. First, we use the regularity of Φ_W from Lemma 12: we have for $x \in X$,

$$\begin{aligned}|h(x) - \hat{h}(x)| = \langle h_W, \Phi(x) - \hat{\Phi}(x)\rangle_{H_W} &\leq \|h_W\|_{H_W}\|\Phi(x) - \hat{\Phi}(x)\|_{H_W} \\ &= \|h_W\|_{H_W}\|\Phi_W \circ \varphi(x) - \Phi_W \circ B \circ \hat{\varphi}(x)\|_{H_W} \\ &\leq \tilde{c}\|h_W\|_{H_W}\|\varphi(x) - B \circ \hat{\varphi}(x)\|_{\ell^2}^s,\end{aligned}$$

where $(\tilde{c}, s) = (\sqrt{2D_{\mathcal{X}}C_{K'}}, \frac{1}{2})$ for a Taylor kernel and $(\tilde{c}, s) = (\sqrt{2C_{K'}}, 1)$ for a radial kernel. Combining with Proposition 11, we obtain Eq. (20). □

Using the universality of the kernel k (thanks to Theorem 5), we can frame the result of Proposition 13 as an approximate universality property of $\hat{k}$. Again, the approximation error functions depend on the type of kernel k_W.

Theorem 14. *Let $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, the element $h[\varepsilon, f] \in H$ defined by*

$$h[\varepsilon, f] := \underset{h \in H : \|h - f\|_\infty \leq \varepsilon}{\operatorname{argmin}} \|h\|_H^2 \tag{21}$$

is well-defined, and there exists $\hat{h} \in \hat{H}$ such that:

$$\|f - \hat{h}\|_\infty \leq \varepsilon + \rho_0 \|h[\varepsilon, f]\|_H, \tag{22}$$

where $\rho_0 = \eta^{\frac{1}{2}} \sqrt{2D_{\mathcal{X}} C_{K'}}$ for a Taylor kernel and $\rho_0 = \eta\sqrt{2C_{K'}}$ for a radial kernel..

Proof. First, we introduce:

$$h_W[\varepsilon, f] := \underset{h_W \in H_W : \|\Psi(h_W) - f\|_\infty \leq \varepsilon}{\operatorname{argmin}} \|h_W\|_{H_W}^2$$

We show that $h_W[\varepsilon, f]$ and $h[\varepsilon, f]$ are well-defined. The triangle inequality ensures that the sets $\mathcal{B}_{H_W} := \{h_W \in H_W : \|\Psi(h_W) - f\|_\infty \leq \varepsilon\}$ and $\mathcal{B}_H := \{h \in H : \|h - f\|_\infty \leq \varepsilon\}$ are convex. We now show the continuity of Ψ as a mapping $(H_W, \|\cdot\|_{H_W}) \longrightarrow (\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$. First, the continuity of $\Psi : (H_W, \|\cdot\|_{H_W}) \longrightarrow (H, \|\cdot\|_H)$ is a consequence of [CS08] Theorem 4.21. Next, by [CS08] Lemma 4.23, since the kernel k_W is bounded on W (we remind that W is a compact subset of ℓ^2), the inclusion $\iota : (H_W, \|\cdot\|_{H_W}) \longrightarrow (\mathcal{C}(W), \|\cdot\|_\infty)$ is continuous. We obtain the desired continuity by composition, showing that $\mathcal{B}_{H_W}$ is closed in H_W. By continuity of ι, we also obtain the closedness of $\mathcal{B}_H$ in H. Finally, the sets $\mathcal{B}_{H_W}$ and $\mathcal{B}_H$ are non-empty since $H = \Psi(H_W)$ is dense in $(\mathcal{C}(X), \|\cdot\|_\infty)$.

We conclude that $\mathcal{B}_{H_W}$, resp. $\mathcal{B}_H$ is a non-empty closed convex set in the Hilbert space $(H_W, \|\cdot\|_{H_W})$, resp. $(H, \|\cdot\|_H)$, and the Hilbert projection theorem (or directly Theorem 4.10 in [Rud87]) ensures that $h_W[\varepsilon, f]$, resp. $h[\varepsilon, f]$ is uniquely defined.

Now, we show that $\|h[\varepsilon, f]\|_H = \|h_W[\varepsilon, f]\|_{H_W}$. By [CS08] Theorem 4.21, we have for all $h \in H$:

$$\|h\|_H = \inf\{\|h_W\|_{H_W},\ h = \Psi(h_W)\}.$$

By the same argument as before (using Theorem 4.10 in [Rud87]), we show that the infimum is attained. The equality between norms is then straightforward by separating both inequalities and using $H = \Psi(H_W)$.

To obtain Eq. (22), we take $h := \Psi(h_W[\varepsilon, f])$ in Eq. (20) and $\hat{h} := \hat{\Psi}(h_W[\varepsilon, f]) \in \hat{H}$, and apply the triangle inequality for $\|\cdot\|_\infty$, using $\|h - f\|_\infty \leq \varepsilon$ and $\|h[\varepsilon, f]\|_H = \|h_W[\varepsilon, f]\|_{H_W}$. □

The approximation result in Eq. (22) shows that $\hat{k}$ is ρ-approximately universal (Definition 6) for $\rho(f) := \rho_0 \|h[\varepsilon, f]\|_H$. In the case where $\mathcal{X}$ is of dimension d (or has intrinsic dimension d), the number of covering balls scales as $J = \mathcal{O}(\eta^{-d})$, which does not impact the approximation rate, as is commonly the case in kernel methods which do not suffer from the curse of dimensionality (see for example [Gre+12] Section 4.1). However, as is typically the case for discretisation methods, the rate $J = \mathcal{O}(\eta^{-d})$ is computationally prohibitive for small discretisation step η in high dimension d.

From a functional standpoint, a larger oscillation (a large value for $C_{K'} = \max_{t \in [-D_{\mathcal{X}}^2, D_{\mathcal{X}}^2]} |K'(t)|$ e.g. for the Taylor case), of the function K worsens the error, which could be understood as excessive locality or over-fitting. Finally, the error term $\rho(f)$ is relative in the sense that it depends on $\|h[\varepsilon, f]\|_H$, which is the smallest possible norm of an ε-approximation of f within H, and can be seen as a measure of complexity of f (in loose terms). This term depends on q, and while the exact dependence is unclear, we expect it to grow as q increases.

3.3 An Approximate Universal Truncated Kernel

In this section, we consider another approximate universal kernel which is obtained by truncation of φ. Unless stated otherwise, we differ from the objects introduced in Section 3.1, and follow very similar steps.

A natural idea is to simply consider a truncation of the mapping φ from Proposition 4: fixing a basis $(y_j)_{j\in\mathbb{N}}$ of $\mathcal{X}$, a discretisation size $J \geq 2$ and scale $q > 1$, consider the mapping:

$$\varphi_t := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R}^J \\ x & \longmapsto & \left(\dfrac{c_\varphi d_{\mathcal{X}}(x, y_j)}{q^j} \right)_{j\in[\![0,J-1]\!]} \end{array} \right. .$$

Straightforward computation shows that $\varphi_t : (\mathcal{X}, d_{\mathcal{X}}) \longrightarrow (\ell^2, \|\cdot\|_{\ell^2})$ is $\sqrt{1-q^{-2J}}$-Lipschitz. We introduce the "padding" isometry:

$$B_t := \left\{ \begin{array}{ccc} \mathbb{R}^J & \longrightarrow & \ell^2 \\ (u_j)_{j=1}^J & \longmapsto & (u_0, \cdots, u_{J-1}, 0, \cdots) \end{array} \right. ,$$

Similarly to Section 3.2, we take $V := \varphi(\mathcal{X})$, $U_t := \varphi_t(\mathcal{X})$ and $V_t := B_t(U_t)$, allowing us to introduce the compact set $W := V \cup V_t \subset \ell^2$ (we use the same notation as in Section 3.1 to alleviate notation). Take k_W a Taylor or radial kernel on W, and introduce the kernel:

$$k_t := \left\{ \begin{array}{ccc} \mathcal{X}^2 & \longrightarrow & \mathbb{R} \\ (x, y) & \longmapsto & k_W(B_t \circ \varphi_t(x), B_t \circ \varphi_t(y)) \end{array} \right. .$$

We continue with the feature pair (H_W, Φ_t) for the RKHS H_t associated to k_t, where:

$$\Phi_t := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & H_W \\ x & \longmapsto & \Phi_W \circ B_t \circ \varphi_t(x) \end{array} \right. .$$

As in Eq. (16) we introduce the "feature-to-map" functionals to close the diagram:

$$\begin{tikzcd}
 & & V \subset W \subset \ell^2 \arrow[dr, "\Phi_W"] & & H \\
\mathcal{X} \arrow[urr, "\varphi"] \arrow[dr, "\varphi_t"'] \arrow[rrr, bend left=60, "\Phi"] \arrow[rrr, bend right=60, "\Phi_t"'] & & & H_W \arrow[ur, "\Psi"] \arrow[dr, "\Psi_t"'] & \\
 & U_t \subset \mathbb{R}^J \arrow[r, "B_t"'] & V_t \subset W \subset \ell^2 \arrow[ur, "\Phi_W"] & & H_t
\end{tikzcd}$$

The computation in the proof Corollary 10 stands, but the coefficients in the expression of φ_t lead to a different expression for k_t, which is a truncated version of k: if k_W is a Taylor kernel, we have:

$$k_t(x, y) = K\left(\langle \varphi_t(x), \varphi_t(y) \rangle_{\mathbb{R}^J}\right) = \sum_{n=0}^{+\infty} a_n \left(\sum_{j=0}^{J-1} \frac{c_\varphi^2 d_{\mathcal{X}}(x, y_j) d_{\mathcal{X}}(y, y_j)}{q^{2j}} \right)^n ,$$

and likewise for radial kernels:

$$k_t(x,y) = K\left(\|\varphi_t(x) - \varphi_t(y)\|_{\mathbb{R}^J}^2\right) = \int_0^{+\infty} \exp\left(-s\sum_{j=0}^{J-1} \frac{c_\varphi^2(d_\mathcal{X}(x,y_j) - d_\mathcal{X}(y,y_j))^2}{q^{2j}}\right) \mathrm{d}\mu(s).$$

We now adapt Proposition 11 to k_t:

Proposition 15. *For $x \in \mathcal{X}$, we have $\|\varphi(x) - B_t \circ \varphi_t(x)\|_{\ell^2} \leq \frac{D_\mathcal{X}}{q^J}$. The diameter of W verifies $D_W \leq D_\mathcal{X}$.*

Proof. For $n \in [\![0, J-1]\!]$, by construction $[\varphi(x)]_n = [B_t \circ \varphi_t(x)]_n$. For $n \geq J$, we have:

$$|[\varphi(x)]_n - [B_t \circ \varphi_t(x)]_n| = \frac{c_\varphi d_\mathcal{X}(x,y_n)}{q^n},$$

and by bounding the distance term by $D_\mathcal{X}$, and summing the squares, we obtain:

$$\|\varphi(x) - B_t \circ \varphi_t(x)\|_{\ell^2}^2 \leq \sum_{n=J}^{+\infty} \frac{c_\varphi^2 D_\mathcal{X}^2}{q^{2n}} = \frac{c_\varphi^2 D_\mathcal{X}^2 q^2}{q^{2J}(q^2-1)} = \frac{D_\mathcal{X}^2}{q^{2J}}.$$

As for the result on D_W, it follows from 1-Lipschitzness as done in Proposition 11. □

As in Section 3.2, it is easy to verify that $R_W := \max_{w\in \mathrm{W}} \|w\|_{\ell^2} \leq D_\mathcal{X}$. Following the same steps as in Theorem 14, we show a similar result for k_t, replacing η with $D_\mathcal{X} q^{-J}$:

Theorem 16. *Let $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, there exists $h_t \in H_t$ such that:*

$$\|f - h_t\|_\infty \leq \varepsilon + \rho_t \|h[\varepsilon, f]\|_H, \tag{23}$$

where $\rho_t = q^{-J/2} D_\mathcal{X}\sqrt{2C_{K'}}$ for a Taylor kernel and $\rho_t = q^{-J} D_\mathcal{X}\sqrt{2C_{K'}}$ for a radial kernel, with the constants $C_{K'}$ as in Lemma 12.

To compare with the rate from Theorem 14, we see that the term η is replaced by $D_\mathcal{X} q^{-J}$. While q^{-J} becomes exponentially smaller as q increases, we suspect the term $\|h[\varepsilon, f]\|_H$ to grow quickly as q increases, which would favour the kernel $\hat{k}$ from Section 3.1.

Acknowledgements

We thank Joan Glaunès for carefully proofreading this work and for his valuable insight.

This research was funded in part by the Agence nationale de la recherche (ANR), Grant ANR-23-CE40-0017 and by the France 2030 program, with the reference ANR-23-PEIA-0004.

References

[Aiz64] A Aizerman. “Theoretical foundations of the potential function method in pattern recognition learning”. In: *Automation and remote control* 25 (1964), pp. 821–837.

[Aro50] Nachman Aronszajn. “Theory of reproducing kernels”. In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[BT11] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[CS08] Andreas Christmann and Ingo Steinwart. *Support vector machines*. Springer, 2008.

[CS10] Andreas Christmann and Ingo Steinwart. “Universal kernels on non-standard input spaces”. In: *Advances in neural information processing systems* 23 (2010).

[CV95] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20 (1995), pp. 273–297.

[Dor+14] G Doran, K Muandet, K Zhang, and B Schölkopf. “A Permutation-Based Kernel Conditional Independence Test”. In: *30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*. AUAI Press. 2014, pp. 132–141.

[Fie23] Christian Fiedler. “Lipschitz and Hölder Continuity in Reproducing Kernel Hilbert Spaces”. In: *arXiv preprint arXiv:2310.18078* (2023).

[FSG13] Kenji Fukumizu, Le Song, and Arthur Gretton. “Kernel Bayes’ rule: Bayesian inference with positive definite kernels”. In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 3753–3783.

[Góm+09] Luis Gómez-Chova, Gustavo Camps-Valls, Lorenzo Bruzzone, and Javier Calpe-Maravilla. “Mean map kernel methods for semisupervised cloud classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 48.1 (2009), pp. 207–220.

[Gre+12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. “A kernel two-sample test”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[HC11] Robert Hable and Andreas Christmann. “On qualitative robustness of support vector machines”. In: *Journal of Multivariate Analysis* 102.6 (2011), pp. 993–1007.

[Jay+15] Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. “Kernel methods on Riemannian manifolds with Gaussian RBF kernels”. In: *IEEE transactions on pattern analysis and machine intelligence* 37.12 (2015), pp. 2464–2477.

[Li+17] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. “Mmd gan: Towards deeper understanding of moment matching network”. In: *Advances in neural information processing systems* 30 (2017).

[MXZ06] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. “Universal Kernels.” In: *Journal of Machine Learning Research* 7.12 (2006).

[Mua+12] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. “Learning from distributions via support measure machines”. In: *Advances in neural information processing systems* 25 (2012).

[Mua+17] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. “Kernel mean embedding of distributions: A review and beyond”. In: *Foundations and Trends® in Machine Learning* 10.1-2 (2017), pp. 1–141.

[Nad64] E. A. Nadaraya. “On Estimating Regression”. In: *Theory of Probability & Its Applications* 9.1 (1964), pp. 141–142. eprint: https://doi.org/10.1137/1109020.

[Par62] Emanuel Parzen. “On estimation of a probability density function and mode”. In: *The annals of mathematical statistics* 33.3 (1962), pp. 1065–1076.

[Ros56] Murray Rosenblatt. “Remarks on some nonparametric estimates of a density function”. In: *Ann. Math. Stat* 27 (1956), pp. 832–837.

[Rud87] Walter Rudin. *Real and complex analysis*. McGraw-Hill, Inc., 1987.

[SS16] Saburou Saitoh and Yoshihiro Sawano. *Theory of reproducing kernels and applications*. Vol. 44. Springer, 2016.

[SSM98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. “Nonlinear component analysis as a kernel eigenvalue problem”. In: *Neural computation* 10.5 (1998), pp. 1299–1319.

[SS02] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[SFL11] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. "Universality, Characteristic Kernels and RKHS Embedding of Measures." In: *Journal of Machine Learning Research* 12.7 (2011).

[Sri+10] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. "Hilbert space embeddings and metrics on probability measures". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1517–1561.

[SZ21] Ingo Steinwart and Johanna F. Ziegel. "Strictly proper kernel scores and characteristic kernels on compact spaces". In: *Applied and Computational Harmonic Analysis* 51 (2021), pp. 510–542.

[Wat64] Geoffrey S. Watson. "Smooth Regression Analysis". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 26.4 (1964), pp. 359–372.

[Zha+12] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. "Kernel-based conditional independence test and application in causal discovery". In: *arXiv preprint arXiv:1202.3775* (2012).

[ZGD22] Johanna Ziegel, David Ginsbourger, and Lutz Dümbgen. "Characteristic kernels on Hilbert spaces, Banach spaces, and on sets of measures". In: *arXiv preprint arXiv:2206.07588* (2022).