Submitted 6/21; Revised 3/23; Published 4/23

Metrizing Weak Convergence with Maximum Mean Discrepancies

Carl-Johann Simon-Gabriel cjsg@ethz.ch
Institute for Machine Learning
ETH Zürich, Switzerland

Alessandro Barp ab2286@cam.ac.uk
Department of Engineering
University of Cambridge, Alan Turing Institute, United Kingdom

Bernhard Schölkopf bs@tue.mpg.de
Empirical Inference Department
MPI for Intelligent Systems, Tübingen, Germany

Lester Mackey lmackey@microsoft.com
Microsoft Research
Cambridge, MA, USA

Editor: Ingo Steinwart

Abstract

This paper characterizes the maximum mean discrepancies (MMD) that metrize the weak convergence of probability measures for a wide class of kernels. More precisely, we prove that, on a locally compact, non-compact, Hausdorff space, the MMD of a bounded continuous Borel measurable kernel k, whose RKHS-functions vanish at infinity (i.e., $\mathscr{H}_k \subset \mathscr{C}_0$), metrizes the weak convergence of probability measures if and only if k is continuous and integrally strictly positive definite ($\int$s.p.d.) over all signed, finite, regular Borel measures. We also correct a prior result of Simon-Gabriel and Schölkopf (JMLR 2018, Thm. 12) by showing that there exist both bounded continuous $\int$s.p.d. kernels that do not metrize weak convergence and bounded continuous non-$\int$s.p.d. kernels that do metrize it.

Keywords: Maximum Mean Discrepancy, Metrization of weak convergence, Kernel mean embeddings, Characteristic kernels, Integrally strictly positive definite kernels

1. Introduction

Although the mathematical and statistical literature has studied kernel mean embeddings (KMEs) and maximum mean discrepancies (MMDs) at least since the 1970s (Guilbart, 1978), the machine learning community rediscovered and applied them only since the late 2000s (Smola et al., 2007). A KME with reproducing kernel k is a map from measures μ – in particular probability distributions – to functions f_μ in the reproducing kernel Hilbert space (RKHS) $\mathscr{H}_k$ of k. The RKHS distance between two embeddings then yields a semi-metric d_k on measures, called the maximum mean discrepancy (MMD), which can be used to compare two measures or distributions μ and ν: $d_k(\mu, \nu) := \|f_\mu - f_\nu\|_k$.

Their theoretical tractability and computational flexibility has allowed MMDs to flourish in many areas of machine learning that require comparing probability distributions, such as two-sample testing (compare two discrete distributions Gretton et al. (2012)), sample quality measurement and goodness-of-fit testing (compare a discrete distribution to a reference distribution, as in Chwialkowski et al. 2016; Liu et al. 2016; Gorham and Mackey 2017; Jitkrittum et al. 2017; Huggins and Mackey 2018), generative model fitting (compare distributions of fake and real data; see Dziugaite et al. 2015; Sutherland et al. 2017; Feng et al. 2017; Pu et al. 2017; Briol et al. 2019), de novo sampling and quadrature (Chen et al., 2010; Huszár and Duvenaud, 2012; Liu and Wang, 2016; Chen et al., 2018; Futami et al., 2019; Chen et al., 2019), importance sampling (Liu and Lee, 2017; Hodgkinson et al., 2020), and thinning (Riabiz et al., 2022).

For most applications, one seeks a kernel k whose MMD can separate all probability distributions P, Q, meaning that, $d_k(P, Q) = 0$ (if and) only if $Q = P$. Such kernels are said to be *characteristic* (to the set of probability distributions $\mathscr{P}$). If for example we optimize a parametric distribution Q to match a target P by minimizing their MMD $d_k(P, Q)$, it is rather natural to require that it be minimized only if Q perfectly matches P, i.e. $Q = P$. Another natural, but a priori stronger requirement, is that when Q gets closer to P in MMD, such as when $d_k(Q, P) \to 0$, we would like Q to "truly" converge to P, where "truly" means "for some other standard and/or more familiar notion of convergence".

Although several standard notions may come to mind – convergence in KL-divergence, in total variation or in Hellinger distance –, many are too strong for our purposes which often require handling discrete data. For example, even if $\boldsymbol{x} \to \boldsymbol{\xi}$, the Dirac masses $\delta_{\boldsymbol{x}}$ will not converge to $\delta_{\boldsymbol{\xi}}$ in total variation or KL-divergence unless $\boldsymbol{x}$ is eventually equal to $\boldsymbol{\xi}$. Said differently, a sequence of deterministic variables would not converge in total variation unless it was eventually constant. Since in practice MMDs are frequently used to compare samples or empirical (hence discrete) distributions, it comes as no surprise that MMD convergence cannot, in general, ensure these strong types of convergence. Instead we will opt for a standard, yet comparatively weak notion of convergence, known as *weak* or *narrow convergence* or *convergence in distribution*. Specifically, the central question of this paper will be

When is convergence in MMD metric equivalent to weak convergence on $\mathscr{P}$?

In that case, we will say that the kernel k *metrizes the weak convergence of probability measures*. This question lies at the heart of the learning applications described above, as the quality of these inferences depends on the metrization properties of the chosen kernel (Zhu et al., 2019, 2021; Ansari et al., 2020; Li et al., 2017). For example, Zhu et al. (2019, 2021) establish that kernel MMD tests have the "universal hypothesis testing property" introduced by Hoeffding (1965) provided that their kernels control weak convergence. Conversely, when the kernel MMD fails to reflect the convergence of distributions, the results are at best inaccurate and at worst invalid.

1.1 Previous results

The aforementioned question was studied as early as 1978 by Guilbart (1978) in his thesis. On separable metric spaces, he characterized the kernels for which weak convergence implies convergence in MMD (Thm. 1.D.I). Conversely, he showed that, in some cases, MMD convergence can also imply weak convergence, meaning that there do exist kernels that metrize weak convergence. He provided a concrete recipe to construct such kernels (Thm. 1.E.I

& Lem. 3.E.I) and used it to exhibit some examples. However, Guilbart (1978) did not characterize these kernels and left most standard kernels (Gaussian, Laplacian, etc.) aside.

These initial results went largely unnoticed by the ML community, and it is only much later, with the emergence and the new applications of MMDs in applied statistics, that the important question of weak convergence metrization re-surfaced. Sriperumbudur et al. (2010) in particular presented sufficient conditions under which the MMD metrizes weak convergence when the underlying input space is either $\mathbb{R}^d$ (Thm. 24) or a compact metric space (Thm. 23). Sriperumbudur (2016, Thm. 3.2) then considerably improved these results and showed the following theorem.

Theorem 1 (Sriperumbudur 2016) *A continuous, bounded, integrally strictly positive definite ($\int$s.p.d.) kernel over a locally compact Polish space $\mathcal{X}$ such that $\mathcal{H}_k \subset \mathscr{C}_0$ metrizes weak convergence.*

Let us explain and discuss this result, as it will provide context for the new results of this work. First, the theorem assumes that the underlying input space is locally compact *and* Polish. Either assumption taken separately is quite general: all topological manifolds (f.ex. $\mathbb{R}^d$) and all discrete spaces are locally compact, and all separable, complete, metric spaces are, by definition, Polish, which includes any separable Banach space. This generality made locally compact spaces on the one hand and Polish spaces on the other standard choices for carrying out general measure and probability theory. However, when the two assumptions are combined, the result can be quite restrictive. A Banach space, for example, is locally compact only if it has finite dimension. Therefore, combining both assumptions yields an important constraint that limits the applicability of the result: one would hope for one or the other but not both.

Second, $\mathcal{H}_k \subset \mathscr{C}_0$ means that the RKHS functions f are assumed to be continuous and vanish at infinity, i.e., for any $\epsilon > 0$, there exists a compact $\mathcal{K} \subset \mathcal{X}$ for which $\sup_{\mathcal{X}\setminus\mathcal{K}} |f| \leq \epsilon$. Many standard kernels satisfy this assumption which is typically easy to verify (see Lem. 8 below). The assumption that f be in $\mathscr{C}_0$ is also rather natural in the context of locally compact spaces $\mathcal{X}$, since, by the Riesz representation theorem, the set of finite, signed and regular measures – a.k.a. finite Radon measures – can be identified with the continuous dual of $\mathscr{C}_0$ (Villani, 2010, Def.VI-66 & Thm.VI-61). This, in turn, can be advantageously leveraged in many proofs and theorems (e.g. for the equivalence between universal and characteristic kernels). However, that same assumption $\mathcal{H}_k \subset \mathscr{C}_0$ is often inadequate on Polish spaces, because, on Polish spaces, $\mathscr{C}_0$ is typically very small: for example, on an infinite dimensional Banach space (hence not locally compact), $\mathscr{C}_0$ contains only the null function. This suggests that it might be more natural to remove the Polish assumption than the locally compact assumption, which is what we will do in this paper. However, by dropping the Polish assumption, we need to pay a bit more attention to the sets of measures that we manipulate. Specifically, on Polish spaces, all signed Borel measures happen to be regular (see definition in Section 1.4), meaning that the set of finite Borel and finite Radon measures coincide there. On locally compact Hausdorff spaces however, this need not be the case. So, when dropping the Polish assumption, we also need to decide on which measures we want to focus and in particular to which measures we would like to be characteristic. As shown in Section 2, it turns out that in cases where Borel and Radon measures do not match, no kernel can be characteristic to all Borel measures. Hence, the only sensible choice is to *focus on Radon measures.*

Third, the theorem assumes that the kernel is $\int$s.p.d., meaning that its MMD separates all finite signed measures $\mathcal{M}$: for any $\mu, \nu \in \mathcal{M}$, $d_k(\mu, \nu) = 0$ only if $\mu = \nu$. It is easy to see that an MMD that metrizes weak convergence on the set of probability measures $\mathcal{P}$, must separate $\mathcal{P}$. But by assuming that it even separates $\mathcal{M}$, which is bigger than $\mathcal{P}$, Sriperumbudur (2016)'s Thm. 2 leaves open the case of any MMD that separates $\mathcal{P}$ but not $\mathcal{M}$.

In 2018, Simon-Gabriel and Schölkopf (2018, Thm. 12) seemed to finally address all weaknesses mentioned above by characterizing the metrization of weak convergence of probability measures on locally compact spaces as follows.

Claim 2 (Simon-Gabriel and Schölkopf 2018) *On a locally compact Hausdorff space, a bounded, Borel measurable kernel metrizes the weak convergence of probability measures if and only if it is continuous and characteristic (to the set of probability measures).*

This statement weakens the sufficient condition of Thm. 1 from separation of $\mathcal{M}$ ($\int$s.p.d. kernel) to separation of $\mathcal{P}$ (characteristic kernel), which, as discussed, immediately yields the converse direction. It gets rid of the Polish assumption and, surprisingly, also drops the assumption $\mathcal{H}_k \subset \mathcal{C}_0$.

1.2 Our contributions

Unfortunately, Claim 2 turns out to be wrong when the input space $\mathcal{X}$ is not compact. Our main result, Thm. 9, provides a correction under the additional assumption that $\mathcal{H}_k \subset \mathcal{C}_0$. Crucially, we find that the compact and non-compact case are inherently different. Metrizing weak convergence on non-compact spaces requires *strictly* stronger conditions, since the MMD needs to separate, not only the probability measures – as in the compact case or in Claim 2 – but all finite signed measures. Put differently, Thm. 9 drops the Polish assumption from Thm. 1 and proves that its converse – which is too strong when $\mathcal{X}$ is compact (see Thm. 7 & Prop. 13) – does hold when $\mathcal{X}$ is non-compact. An important implication is that any $\mathcal{C}_0$ kernel that maps a probability measure to 0 *fails* to metrize weak convergence; in particular, this establishes that large classes of Stein kernels are unable to metrize convergence (see Rem. 10).

Additionally, Cor. 15 shows that Thm. 9 does not hold without the assumption $\mathcal{H}_k \not\subset \mathcal{C}_0$, while Cor. 17 provides a *sufficient* condition to metrize weak convergence when $\mathcal{H}_k \not\subset \mathcal{C}_0$. Our results also complete the findings of Chevyrev and Oberhauser (2022), who constructed a counter-example showing that Claim 2 does not hold on Polish spaces. Overall, our findings show that the old quest to characterize weak-convergence metrizing MMDs – which we close under the quite general assumption that $\mathcal{X}$ is locally compact and $\mathcal{H}_k \subset \mathcal{C}_0$ – depends in much more subtle ways on the properties of the underlying space $\mathcal{X}$ (being compact or not, Polish or not, etc.) and the kernel k ($\mathcal{H}_k$ contained in $\mathcal{C}_0$ or not) than was previously thought.

1.3 Paper structure

Section 1.4 fixes notations and makes a few important reminders and remarks. Section 3 then extends Sriperumbudur (2016)'s Thm. 1 and gives a general sufficient condition to metrize weak convergence when $\mathscr{H}_k \subset \mathscr{C}_0$. We then investigate whether this condition is also necessary, first when the input space $\mathscr{X}$ is compact (Sec. 4), where it turns out to be too strong (Thm. 7); then when $\mathscr{X}$ is not compact, but locally compact (Sec. 5), in which case the sufficient condition turns out to be necessary (Thm. 9). We finish with a few results in the general case (Sec. 6), when $\mathscr{H}_k \not\subset \mathscr{C}_0$: first a negative result (Cor. 15) showing that the assumption $\mathscr{H}_k \subset \mathscr{C}_0$ cannot be dropped without replacement; then a result that generalizes the condition $\mathscr{H}_k \subset \mathscr{C}_0$. Section 7 concludes.

1.4 Notation, definitions, and reminders

We use letter k to denote a reproducing kernel (i.e. a positive definite function) over a locally compact Hausdorff (LCH) $\mathscr{X}$ and $\mathscr{H}_k$ denotes its RKHS. $\mathscr{C}_b$ is the space of bounded, continuous and real valued [1] functions f over $\mathscr{X}$. $\mathscr{C}_0$ is its subspace of functions that vanish at infinity, i.e. such that for any $\epsilon > 0$, there exists a compact $\mathscr{K} \subset \mathscr{X}$ such that $|f| \leq \epsilon$ on $\mathscr{X} \backslash \mathscr{K}$. We say that k is a $\mathscr{C}_0$-kernel if $\mathscr{H}_k \subset \mathscr{C}_0$ and that it is $\mathscr{C}_0$-universal if its $\mathscr{H}_k$ is also dense in $\mathscr{C}_0$. We denote by $\mathcal{M}^*$ the set of finite, signed Borel measures, and by $\mathcal{M}$ the set of finite *Radon* measures, i.e., the subset of signed measures in $\mathcal{M}^*$ that are also *regular*. Recall that a positive Borel be *regular* if for any Borel measurable set $\mathscr{A}$ and any $\epsilon > 0$, there exists a compact $\mathscr{K}$ and an open set $\mathscr{O}$ in $\mathscr{X}$ such that $\mathscr{K} \subset \mathscr{A} \subset \mathscr{O}$, $|\mu(\mathscr{A}) - \mu(\mathscr{K})| \leq \epsilon$ and $|\mu(\mathscr{O}) - \mu(\mathscr{A})| \leq \epsilon$. Said differently, a measure is regular if any measurable set can be approximated (in terms of measure) from the inside by the compacts it contains and from the outside by the open sets that contain it. A signed Borel measure is regular if its positive and negative parts are. Except for Section 2 where we discuss the differences between Borel and Radon measures, this work focuses on finite Radon measures. *When used without further specification, the word "measure" designates an element in* $\mathcal{M}$. We denote by $(\mathscr{C}_0)'$ the continuous dual of $\mathscr{C}_0$ which, by the Riesz representation theorem (a.k.a. Riesz-Markov-Kakutani theorem Villani 2010, VI-61), can be identified with $\mathcal{M}$. $\mathrm{L}(\mu)$ denotes the set of μ-integrable functions (i.e. verifying $\int_{\mathscr{X}} |f| \, \mathrm{d}|\mu| < \infty$) and for any such function f we write $\mu(f) := \int_{\mathscr{X}} f \, \mathrm{d}\mu$. We denote by $\mathcal{M}_+$, $\mathscr{P}$ and $\mathcal{M}^0$ the subsets of $\mathcal{M}$ consisting of non-negative measures, of probability measures, and of signed measures μ such that $\mu(\mathscr{X}) = 0$ respectively.

Definition of KMEs and MMDs. For a continuous, bounded kernel k and any $\mu \in \mathcal{M}$, $\int_{\mathscr{X}} \|k(.,\boldsymbol{x})\|_k \, \mathrm{d}|\mu| = \int_{\mathscr{X}} \sqrt{k(\boldsymbol{x},\boldsymbol{x})} \, \mathrm{d}|\mu|(\boldsymbol{x}) < \infty$. By standard properties of the so-called *Bochner integral* (Schwabik, 2005), the (Bochner-)integral

$$f_\mu(\cdot) := \int_{\mathscr{X}} k(.,\boldsymbol{x}) \, \mathrm{d}\mu(\boldsymbol{x})$$

is a well-defined function in the RKHS $\mathscr{H}_k$ of k, and all functions $f \in \mathscr{H}_k$ are μ-integrable and verify what we call the *Pettis property*: $\mu(f) = \langle f_\mu \, , \, f \rangle_k$. In particular, for any $\mu, \nu \in \mathcal{M}$,

$$\langle \mu \, , \, \nu \rangle_k := \langle f_\mu \, , \, f_\nu \rangle_k = \mu \otimes \nu(k) \qquad \text{and} \qquad \|\mu\|_k^2 = \mu \otimes \mu(k) \ ,$$

[1] Our results extend to complex valued functions modulo some obvious slight modifications.

where $\mu \otimes \nu$ denotes the (tensor) product measure between μ and ν. The maximum mean discrepancy (MMD) $d_k(\mu, \nu)$ between μ and ν is then defined as the RKHS distance between their embeddings:

$$d_k(\mu, \nu) := \|\mu - \nu\|_k = \|f_\mu - f_\nu\|_k \ .$$

Why bounded kernels? In all our results, we will assume that the kernel k is bounded. One may wonder if those results could be generalized to unbounded kernels. To do so, one would need a definition of KMEs and MMDs that allows unbounded kernels. Such generalizations do exist (see f.ex. Def. 1 in Simon-Gabriel and Schölkopf 2018), but they all at least require that $\mathscr{H}_k \subset \mathrm{L}(\mu)$ for any embeddable measure μ. But if k is unbounded, then $\mathscr{H}_k$ contains an unbounded function f (Simon-Gabriel and Schölkopf, 2018, Cor. 3), and therefore, it is easy to construct a probability measure P such that $f \notin \mathrm{L}(P)$. So P does not embed into $\mathscr{H}_k$ and the MMD is not defined over all probability measures and cannot, a fortiori, metrize weak convergence there.

Equivalence of universal, characteristic and $\int$s.p.d. kernels. Let $\mathscr{F}$ be a normed set of functions and $\mathscr{D}$ a subset of $\mathcal{M}$. A kernel k is said to be *universal to* $\mathscr{F}$ if $\mathscr{H}_k$ is a dense subset of $\mathscr{F}$. It is *characteristic to* $\mathscr{D}$ – or just *characteristic* when $\mathscr{D} = \mathscr{P}$ – if the KME is well-defined and injective over $\mathscr{D}$. It is said to be *integrally strictly positive definite* (*$\int$s.p.d.*) *to* $\mathscr{D}$ – or just *$\int$s.p.d.* when $\mathscr{D} = \mathcal{M}$ – if its MMD separates all measures in $\mathscr{D}$. It will be useful to remember that a kernel is universal to $\mathscr{F}$ (f.ex. to $\mathscr{C}_0$) if and only if it is characteristic to its dual ($(\mathscr{C}_0)' = \mathcal{M}$) (Simon-Gabriel and Schölkopf, 2018, Thm. 6 & Tab. 1). Also, it is characteristic to a set if and only if it is $\int$s.p.d. to that same set (which is almost immediate to see). The distinction between characteristicness and $\int$s.p.d. is mostly due to historical reasons. We advice to simply think in terms of separation of $\mathscr{D}$.

2. Radon versus Borel measures

As explained in introduction, we would like to drop the Polish assumption in Thm. 1 and focus on LCH spaces. However, while on Polish spaces all finite, signed Borel measures happen to be regular, i.e. $\mathcal{M}^* = \mathcal{M}$ by Ulam's lemma (Villani, 2010, Thm. I-54), on LCH spaces, this need not be the case. So, if we drop the Polish assumption in Thm. 1, should we focus on characteristicness to Borel or to Radon measures? The following theorem answers this question. It is a direct consequence of the proof of Thm 3.13 in Steinwart and Ziegel (2021).

Theorem 3 (Steinwart and Ziegel 2021) *Let $\mathcal{X}$ be a locally compact Hausdorff space. If a $\mathscr{C}_0$-kernel is characteristic to a set of measures $\mathscr{D} \subset \mathcal{M}^*(\mathcal{X})$, then $\mathscr{D} \subset \mathcal{M}(\mathcal{X})$. In particular, if $\mathcal{M}^*(\mathcal{X}) \neq \mathcal{M}(\mathcal{X})$, then no $\mathscr{C}_0$-kernel is characteristic to $\mathcal{M}^*(\mathcal{X})$.*

Said differently, the biggest set of finite Borel measures that a $\mathscr{C}_0$-kernel can be characteristic to is the set of finite Radon measures $\mathcal{M}$.

However, how common is it that $\mathcal{M} \neq \mathcal{M}^*$? First, we note that the conclusion of Ulam's lemma ($\mathcal{M} = \mathcal{M}^*$) also holds for LCH spaces $\mathcal{X}$ if one additionally assumes that $\mathcal{X}$ is σ-compact, i.e., that it can be covered by at most countably many compact sets (Villani, 2010, Thm. I-56). However, how restrictive is the σ-compact assumption for an LCH space? Examples of non-σ-compact LCH spaces such as the *long line* (a.k.a. Alexandroff line) exist, but they may seem irrelevant to the working machine learner. In contrast, the following theorem shows that, on an LCH space, (a) considering a continuous $\mathscr{C}_0$-universal kernel (i.e., a continuous $\mathscr{C}_0$-kernel that is characteristic to $\mathcal{M}$), amounts to assuming that $\mathcal{X}$ is metrizable and (b) that in this context, assuming σ-compactness amounts to adding a separability assumption on $\mathcal{X}$. The proof is in Appendix B.

Theorem 4 *Let $\mathcal{X}$ be an LCH space.*

(a) A $\mathscr{C}_0$-universal kernel k on $\mathcal{X}$ is continuous if and only if (iff) it metrizes $\mathcal{X}$, i.e. if $d_k(x,y) := \|k(.,x) - k(.,y)\|_k$ is a metric for the topology of $\mathcal{X}$. In particular, if there exists a continuous $\mathscr{C}_0$-universal kernel on $\mathcal{X}$, then $\mathcal{X}$ is metrizable.

(b) Moreover, the following is equivalent.

(i) $\mathcal{X}$ is metrizable and separable.

(ii) $\mathcal{X}$ is σ-compact and there exists a continuous $\mathscr{C}_0$-universal kernel k on $\mathcal{X}$.

Point (a) shows that even if we drop the Polish assumption, whenever we consider a continuous $\mathscr{C}_0$-universal kernel, we are still assuming that $\mathcal{X}$ is metrizable. Point (b) adds that, if additionally $\mathcal{X}$ is assumed to be σ-compact, then the only missing assumption for $\mathcal{X}$ to be Polish is its completeness.

To finish this section, let us discuss some of the hypotheses made in Thm. 4. First, Guilbart (1978, Thm. 4.D.I) shows that, even without the LCH assumption on $\mathcal{X}$, (i) implies the existence of a kernel that is characteristic to $\mathcal{M}$ (i.e., to $\mathcal{M}^*$). Second, separability (or σ-compactness) is not a required condition for the existence of a $\mathscr{C}_0$-universal kernel on a (metrizable) LCH space. For example, the discrete kernel $k_\delta(x,y) = \mathbb{1}(x=y)$ is a $\mathscr{C}_0(\mathbb{R}_\delta)$-universal kernel over the discrete real line $\mathbb{R}_\delta$, i.e., the real line equipped with the discrete topology. (To see this, notice that the compact sets in $\mathbb{R}_\delta$ are the finite subsets of $\mathbb{R}$, and hence that $\mathscr{C}_0(\mathbb{R}_\delta)$ is the set of real functions for which only finitely many points have a value $\geq \epsilon$, whatever $\epsilon > 0$ you choose.) We do not know wether, more generally, the converse of (a) is true, i.e., whether on an LCH space, metrizability alone suffices to guarantee the existence of a $\mathscr{C}_0$-universal kernel. Finally, we note that in some publications, the continuity assumption on k is hidden in the definition of $\mathscr{C}_0$-universality (see f.ex. Thm 2 in Steinwart et al. 2006). However, this need not be the case (see Simon-Gabriel and Schölkopf, 2018, Cor 3&Def 5) and so one may wonder if a non-continuous $\mathscr{C}_0$-universal kernel exist. We do not know.

3. Sufficient conditions to metrize weak convergence

We start with a lemma that extends Thm. 1. Its main message is the same: bounded, continuous, $\int$s.p.d. kernels metrize weak convergence of probability measures. But, importantly, it drops the Polish assumption and adds a few interesting details. For one thing, it shows that weak and MMD convergence also coincide with (the a priori even weaker) vague and weak RKHS convergence. For another, it adds a form of converse: weak convergence implies MMD convergence if *and only if* the kernel is bounded and continuous. Since most usual kernels are bounded and continuous, this lemma also confirms what we mentioned earlier: convergence in MMD is often rather weak and can, at best, metrize weak convergence, but not convergence in total variation or KL divergence (since those are known to be strictly stronger than weak convergence).

Lemma 5 *Let k be an $\int$s.p.d. kernel such that $\mathscr{H}_k \subset \mathscr{C}_0$ and let (P_α) (sequence or net) and P be probability measures. If k is continuous, then the following are equivalent.*

(i) $\|P_\alpha - P\|_k \to 0$ *(convergence in strong RKHS topology)*
(ii) $P_\alpha(f) \to P(f)$ *for all* $f \in \mathscr{H}_k$ *(convergence in weak RKHS topology)*
(iii) $P_\alpha(f) \to P(f)$ *for all* $f \in \mathscr{C}_0$ *(convergence in weak-$*$ or vague topology)*
(iv) $P_\alpha(f) \to P(f)$ *for all* $f \in \mathscr{C}_b$ *(convergence in weak topology)*

Conversely, if (iv) *implies* (i) *for any probability measures (P_α) and P, then k is continuous.*

When (i) and (iv) are equivalent for all sequences of probability measures, we say that k *metrizes the weak convergence of probability measures.*

Proof Since $\mathscr{H}_k \subset \mathscr{C}_0 \subset \mathscr{C}_b$, (iv) ⇒(iii) ⇒(ii). Moreover, strong RKHS convergence implies weak RKHS convergence, that is (i) ⇒(ii), since $P(f) = \langle P\,,\, f\rangle_k$ for any $f \in \mathscr{H}_k$. Now assume k is continuous. If (iv), then the product measures $P_\alpha \otimes P$, $P \otimes P_\alpha$ and $P_\alpha \otimes P_\alpha$ converge weakly to $P \otimes P$ (Berg et al., 1984, Thm. 2.3.3). Hence

$$\|P_\alpha - P\|_k^2 = P_\alpha \otimes P_\alpha(k) + P \otimes P(k) - P_\alpha \otimes P(k) - P \otimes P_\alpha(k) \to 0\ ,$$

i.e. (iv) ⇒(i). Summing up so far: (iv) ⇒(i) ⇒(ii) and (iv) ⇒(iii) ⇒(ii).

Conversely, assume (ii). Since k is $\int$s.p.d. and $\mathscr{H}_k \subset \mathscr{C}_0$, by Cor. 3 and Thm. 8 in Simon-Gabriel and Schölkopf (2018), $\mathscr{H}_k$ is dense in $\mathscr{C}_0$. And since $\mathscr{P}$ is a bounded subset of the dual $\mathcal{M}$ of $\mathscr{C}_0$ (which is a Banach, hence barreled space), by Thm. 33.2 in Treves (1967), $\mathscr{P}$ is equicontinuous. So, by Prop. 32.5 in Treves (1967), (ii) implies vague convergence, i.e. (iii). Cor. 2.4.3 in Berg et al. (1984) then yields (iv). Hence the equivalence of (i) to (iv).

Now assume (iv) ⇒(i) on $\mathscr{P}$, and suppose that $\boldsymbol{x} \to \boldsymbol{\xi}$ and $\boldsymbol{y} \to \boldsymbol{\zeta}$ in $\mathcal{X}$. Then the Dirac point masses $\delta_{\boldsymbol{x}}$ and $\delta_{\boldsymbol{y}}$ converge weakly to $\delta_{\boldsymbol{\xi}}$ and $\delta_{\boldsymbol{\zeta}}$, which, by assumption, implies convergence in RKHS norm. Since the inner product is continuous (for the RKHS norm/topology), we get

$$k(\boldsymbol{x}, \boldsymbol{y}) = \langle \delta_{\boldsymbol{x}}\,,\, \delta_{\boldsymbol{y}}\rangle_k \to \langle \delta_{\boldsymbol{\xi}}\,,\, \delta_{\boldsymbol{\zeta}}\rangle_k = k(\boldsymbol{\xi}, \boldsymbol{\zeta})\ ,$$

so k is continuous. ■

Remark 6 *The proof shows that* (ii) *and* (iii) *are even equivalent on any bounded subset of $\mathcal{M}$ (Treves, 1967, Prop. 32.5) (even without continuity of k) and that* (i)–(iv) *are actually equivalent on any bounded subset of $\mathcal{M}_+$ whenever $P_\alpha(\mathcal{X}) \to P(\mathcal{X})$ (which is always true for probability measures).*

The previous lemma gives sufficient conditions to metrize weak convergence. We now investigate whether they are necessary. To do so, we have to distinguish the case where the input space $\mathcal{X}$ is compact and where the conditions turn out to be too strong, from the one where $\mathcal{X}$ is locally compact but not compact (and $\mathscr{H}_k \subset \mathscr{C}_0$), where they are necessary.

4. Necessary condition for compact input space $\mathcal{X}$

When the underlying space $\mathcal{X}$ is not just locally compact but compact, the equivalence given in Claim 2 actually turns out to hold: contrary to the general case, here, a continuous kernel only needs to separate the probability measures to also metrize their weak convergence. The reason for this difference is essentially that, because $\mathcal{X}$ is compact, measures cannot diffuse to 0 at infinity (see Section 5).

Theorem 7 *On a compact Hausdorff space, a bounded, measurable kernel metrizes the weak convergence of probability measures if and only if it is continuous and characteristic to $\mathscr{P}$.*

Proof If k metrizes weak convergence, then the RKHS metric needs to separate all probability measures, i.e. k is characteristic to $\mathscr{P}$. And the last sentence of Lem. 5 shows that k is continuous. Conversely, if k is characteristic to $\mathscr{P}$, then the kernel $\kappa := k + 1$ is $\int$s.p.d. (Simon-Gabriel and Schölkopf, 2018, Thm. 8). Also, since k is continuous, κ is continuous. Thus $\mathcal{H}_\kappa$ is a continuous subspace of $\mathscr{C} = \mathscr{C}_b = \mathscr{C}_0$ (Simon-Gabriel and Schölkopf 2018, Cor. 3 and compactness). By Lem. 5, κ metrizes weak convergence on $\mathscr{P}$, and by Thm. 8 of Simon-Gabriel and Schölkopf (2018), κ and k induce the same metric on $\mathscr{P}$. ■

What is surprising here is that, on a compact space and for a continuous kernel, it suffices to separate probability measures to also metrize their weak convergence, which, a priori, may have seemed a strictly stronger requirement. We will see that when $\mathcal{X}$ is not compact, this need not be the case.

5. Necessary condition when $\mathcal{X}$ is locally compact but non-compact and $\mathcal{H}_k \subset \mathscr{C}_0$

Since the condition $\mathcal{H}_k \subset \mathscr{C}_0$ is at the heart of this section, we would like to remind the reader that, by the following lemma (Simon-Gabriel and Schölkopf, 2018, Cor. 3), it is satisfied by many standard kernels: Gaussian, Laplacian, Matern, inverse multi-quadratic kernels, etc.

Lemma 8 *$\mathcal{H}_k \subset \mathscr{C}_0$ if and only if k is bounded (i.e. $\sup_{\boldsymbol{x}\in\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}) < \infty$) and for all $\boldsymbol{x} \in \mathcal{X}$, $k(\boldsymbol{x}, .) \in \mathscr{C}_0$.*

We now turn to our main theorem, which corrects Claim 2 when $\mathcal{X}$ is non-compact and $\mathcal{H}_k \subset \mathscr{C}_0$.

Theorem 9 *Suppose that the locally compact Hausdorff space $\mathcal{X}$ is not compact and that, for some kernel k on $\mathcal{X} \times \mathcal{X}$, $\mathcal{H}_k(\mathcal{X}) \subset \mathscr{C}_0(\mathcal{X})$. Then k metrizes the weak convergence of probability measures if and only if k is continuous and $\int$s.p.d. (i.e. characteristic to $\mathcal{M}(\mathcal{X})$).*

We see that, contrary to the compact case, it is not enough to separate all probability measures $\mathscr{P}$ to metrize their weak convergence: d_k must separate all finite measures $\mathcal{M}$, which strictly contains $\mathscr{P}$. Moreover, Prop. 13 below confirms that there are indeed kernels that separate $\mathscr{P}$ but not $\mathcal{M}$. Hence, Thms. 7 and 9 show that, surprisingly, the converse of Sriperumbudur's Thm. 1 is generally too restrictive when $\mathcal{X}$ is compact but does hold when it is not. Also, they confirm that the Polish assumption made in Thm. 1 is superfluous.

Remark 10 (On the significance of Thm. 9) *One advantage of dropping the Polish assumption is that our result may cover more sets, e.g. non complete ones. Besides, we believe that dropping unnecessary hypotheses helps clarifying the role of each remaining assumption. However, in our view, the main contribution of Thm. 9 is its converse part, which implies that many popular kernels* fail *to metrize weak convergence. For example, it rules out any RKHS contained in $\mathscr{C}_0$ that maps some probability measure(s) to 0. This has important implications for the Stein kernels adopted in Liu and Wang (2016); Jitkrittum et al. (2017); Gorham and Mackey (2017); Huggins and Mackey (2018); Feng et al. (2017); Pu et al. (2017); Liu and Wang (2016); Chen et al. (2018, 2019); Hodgkinson et al. (2020) which, by design, map a particular target distribution to* 0 *and which, if one is not careful, will also induce RKHSes in $\mathscr{C}_0$.*

We now turn towards the proof. While it is almost obvious that metrization of weak convergence implies separation of $\mathscr{P}$, showing that it also implies separation of $\mathcal{M}$ will require some work and, in light of Lem. 5, is essentially all that remains to be proven. To do so, we will use the following two lemmata. The first one is a straightforward extension of a basic property of locally compact sets (every *point* has a compact neighborhood) from points to compact sets (*every* compact set has a compact neighborhood). The second shows that when $\mathscr{H}_k \subset \mathscr{C}_0$ and $\mathcal{X}$ is not compact, then the RKHS metric cannot prevent some positive measures from "diffusing" to the null measure. This will imply that if k is not characteristic to all finite measures, one can construct a sequence of probability measures that converges in RKHS norm but has some of its mass diffusing to 0.

Lemma 11 *Let $\mathcal{K}$ be a compact subset of a locally compact space $\mathcal{X}$. Then there exists an open neighborhood of $\mathcal{K}$ with compact closure. Equivalently, there exist an open set $\mathcal{O}$ and a compact set $\mathcal{K}'$ in $\mathcal{X}$ such that $\mathcal{K} \subset \mathcal{O} \subset \mathcal{K}'$.*

Lemma 12 *Suppose that the locally compact Hausdorff space $\mathcal{X}$ is not compact and that k is continuous with $\mathscr{H}_k \subset \mathscr{C}_0$. Then there exists a sequence of probability measures P_n such that $\|P_n\|_k \to 0$. Moreover, for any compact $\mathcal{K} \subset \mathcal{X}$, one can additionally impose that $P_n(\mathcal{K}) = 0$ for all n.*

As a side remark, note that Lem. 12 complements Lem 3.2 of Steinwart and Ziegel (2021), which states that, if the constant-1 function $\mathbb{1}$ is in $\mathscr{H}_k$, then $\mathcal{M}_0$ is $\mathscr{H}_k$-closed. In contrast, Lem. 12 from above shows that, if $\mathscr{H}_k \subset \mathscr{C}_0$ (in which case $\mathbb{1} \notin \mathscr{H}_k$), then $\mathscr{P}$ is not $\mathscr{H}_k$-closed and neither is $\mathcal{M}_0$ (to see this, replace $(P_n)_n$ by $(P_n - P)_n$ for some arbitrary $P \in \mathscr{P}$ in Lem. 12).

Proof [Proof of Lem. 11] Since $\mathcal{X}$ is locally compact, every point has a compact neighborhood. So let us consider the set of all compact neighborhoods of the points contained in $\mathcal{K}'$. Their interiors form an open cover of $\mathcal{K}'$, and, since $\mathcal{K}'$ is compact, a finite number of them suffices to cover $\mathcal{K}'$. Let $\mathcal{O}$ be the finite union of these interiors and $\mathcal{K}'$ the union of their closures (i.e., the union of the corresponding compact supersets). Then $\mathcal{O}$ is open, $\mathcal{K}'$ is compact, and $\mathcal{K} \subset \mathcal{O} \subset \mathcal{K}'$ as advertised. We finally note that this property is equivalent to the first claim (that there exists an open neighborhood of $\mathcal{K}$ with compact closure) as $\mathcal{O}$ is contained in a compact set if and only if its closure is compact. ■

Proof [Proof of Lem. 12] First we show that for any $\epsilon > 0$ and any integer $n > 0$, we can construct a sequence of n points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in $\mathcal{X} \backslash \mathcal{K}$ such that for any $1 \leq i \neq j \leq n$, $|k(\boldsymbol{x}_i, \boldsymbol{x}_j)| \leq \epsilon$. We will construct it one point at a time. Choose a point $\boldsymbol{x}_1 \in \mathcal{X} \backslash \mathcal{K}$. By assumption on k, there exists a compact $\mathcal{K}_1 \subset \mathcal{X}$ such that for any point $\boldsymbol{x} \in \mathcal{X} \backslash \mathcal{K}_1$, $|k(\boldsymbol{x}, \boldsymbol{x}_1)| \leq \epsilon$. Choose $\boldsymbol{x}_2$ to be also outside of $\mathcal{K}$, i.e. $\boldsymbol{x}_2 \in \mathcal{X} \backslash (\mathcal{K} \cup \mathcal{K}_1)$ (non-empty, since $\mathcal{K} \cup \mathcal{K}_1$ is compact and $\mathcal{X}$ is not). There exists a compact $\mathcal{K}_2 \subset \mathcal{X}$ such that for any point $\boldsymbol{x} \in \mathcal{X} \backslash \mathcal{K}_2$, $|k(\boldsymbol{x}, \boldsymbol{x}_2)| \leq \epsilon$. Let $\boldsymbol{x}_3$ be any point in $\mathcal{X} \backslash (\mathcal{K} \cup \mathcal{K}_1 \cup \mathcal{K}_2)$ (non empty because $\mathcal{X}$ is not compact). Continue this procedure until point $\boldsymbol{x}_n$. The sequence obviously satisfies the requirement.

Now, for any integer $n > 0$, construct a finite sequence $\boldsymbol{x}_1^{(n)}, \ldots \boldsymbol{x}_n^{(n)}$ such that for any $1 \leq i \neq j \leq n$, $|k(\boldsymbol{x}_i, \boldsymbol{x}_j)| \leq 1/n$. Define the probability measures $P_n := \frac{1}{n} \sum_{i=1}^n \delta_{\boldsymbol{x}_i^{(n)}}$. Then all $P_n(\mathcal{K}) = 0$, since all $\boldsymbol{x}_i^{(n)} \in \mathcal{X} \backslash \mathcal{K}$, and:

$$\|P_n\|_k^2 = \frac{1}{n^2} \sum_{1 \leq i \leq n} k(\boldsymbol{x}_i, \boldsymbol{x}_i) + \frac{1}{n^2} \sum_{1 \leq i \neq j \leq n} k(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq \frac{n}{n^2} \|k\|_\infty + \frac{n(n-1)}{n^2} \frac{1}{n} \overset{n \to \infty}{\longrightarrow} 0. \quad \blacksquare$$

Proof [Proof of Thm. 9] Lem. 5 yields the "if" part and the continuity of the kernel in the converse. Assume now that k is not characteristic to $\mathcal{M}$. Then there exists a non-zero, finite measure μ such that $f_\mu = 0$. Let μ_+, μ_- be its positive and negative parts respectively – which are mutually singular (Hahn decomposition). By renormalizing μ if needed, we can assume without loss of generality that $\mu_-(\mathcal{X}) \leq \mu_+(\mathcal{X}) = 1$. If $\mu_-(\mathcal{X}) = \mu_+(\mathcal{X})$, then μ_- and μ_+ are two non-equal probability measures that are at RKHS distance 0, hence k does not metrize weak convergence. So, for the sequel, assume that $\mu_-(\mathcal{X}) < \mu_+(\mathcal{X})$.

Now, let $\mathcal{K}$ be a compact subset of $\mathcal{X}$ that satisfies $\mu_+(\mathcal{K}) \geq (\mu_-(\mathcal{X}) + \mu_+(\mathcal{X}))/2$, which exists because μ_+ is regular and $\mu_-(\mathcal{X}) < \mu_+(\mathcal{X})$. Select now an open set $\mathcal{O}$ and a compact set $\mathcal{K}'$ satisfying $\mathcal{K} \subset \mathcal{O} \subset \mathcal{K}'$, which exist by Lem. 11. Then, since $\mathcal{K} \subset \mathcal{O}$, $\mu_+(\mathcal{O}) \geq \mu_+(\mathcal{K})$. Let now P_n be probability measures as in Lem.12 such that $P_n(\mathcal{K}') = 0$ (and hence $P_n(\mathcal{O}) = 0$) for all n. Consider the sequence of probability measures $\mu_n := \mu_- + (1 - \mu_-(\mathcal{X}))P_n$. Then

$$\begin{aligned} \|\mu_n - \mu_+\|_k &= \|\mu_n - \mu_-\|_k \quad (\text{because } f_{\mu_-} = f_{\mu_+}) \\ &= (1 - \mu(\mathcal{X})) \|P_n\|_k \quad \longrightarrow \quad 0, \end{aligned}$$

hence μ_n converges to μ_+ in the RKHS metric. But μ_n does not converge weakly to μ_+ since

$$\mu_+(\mathcal{O}) \geq \mu_+(\mathcal{K}) \geq (\mu_-(\mathcal{X}) + \mu_+(\mathcal{X}))/2 > \mu_-(\mathcal{X}) \geq \mu_-(\mathcal{O}) = \mu_n(\mathcal{O}) \ ,$$

which contradicts the Portmanteau lemma ($\limsup_n \mu_n(\mathcal{O}) \not\geq \mu_+(\mathcal{O})$). ■

To prove that the initial claim (Claim 2) is indeed wrong when $\mathcal{X}$ is not compact, it remains to show that being characteristic to $\mathcal{M}$ is not equivalent to being characteristic to $\mathcal{P} \subset \mathcal{M}$, i.e. that there exists a kernel k with $\mathcal{H}_k \subset \mathcal{C}_0$ that is characteristic to $\mathcal{P}$ but not to $\mathcal{M}$. We show this under the assumption that there already exists a kernel of $\mathcal{X}$ that is characteristic to $\mathcal{M}$, which is in particular satisfied when $\mathcal{X}$ is metrizable and separable (Thm. 4(b) or Thm. 4.D.I in Guilbart 1978), such as when $\mathcal{X}$ is an open subset of $\mathbb{R}^d$.

Proposition 13 *If there exists a bounded continuous kernel over a locally compact Hausdorff space $\mathcal{X}$ that is characteristic to $\mathcal{M}$, then there also exists a kernel k with $\mathcal{H}_k \subset \mathscr{C}_0(\mathcal{X})$ that is characteristic to $\mathcal{P}$ but not characteristic to $\mathcal{M}$. In particular, this k does not metrize the weak convergence of probability measures.*

Proof Let κ be any bounded kernel over $\mathcal{X}$ that is $\int$s.p.d., i.e., characteristic to $\mathcal{M}$, $\boldsymbol{\xi} \in \mathcal{X}$ and $g \in \mathscr{C}_0$ such that $g(\boldsymbol{\xi}) = 0$ and $g(\boldsymbol{x}) > 0$ for any $\boldsymbol{x} \neq \boldsymbol{\xi}$. Consider $k(\boldsymbol{x}, \boldsymbol{y}) := g(\boldsymbol{x})\kappa(\boldsymbol{x}, \boldsymbol{y})g(\boldsymbol{y})$. Then k is a kernel such that $\mathcal{H}_k \subset \mathscr{C}_0$ (Lem. 8) and $f_{\delta_{\boldsymbol{\xi}}}$ is the null function, hence $\|\delta_{\boldsymbol{\xi}}\|_k = 0$, so k is not $\int$s.p.d. But we will now show that k is characteristic to $\mathcal{M}^0$, i.e. to $\mathcal{P}$. Indeed, let $\mu \in \mathcal{M}^0$ such that $\iint k(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\mu(\boldsymbol{x}) \, \mathrm{d}\mu(\boldsymbol{y}) = 0$. Since the product $g\mu$ is a finite measure and κ is $\int$s.p.d., the previous equality implies that $g\mu$ is the null measure. Since $g > 0$ on any $\boldsymbol{x} \neq \boldsymbol{\xi}$, for any open set $\mathcal{O} \subset \mathcal{X} \backslash \{\boldsymbol{\xi}\}$, $|\mu|(\mathcal{O}) = 0$. Hence the support of μ (well-defined, because μ is regular) is contained in $\{\boldsymbol{\xi}\}$, i.e. μ is proportional to the Dirac point mass in $\boldsymbol{\xi}$. Hence, if $\mu \in \mathcal{M}^0$, then μ is the null measure. ■

Prop. 13 has two implications. First, it shows that the metrization condition in the non compact case is *strictly* stronger than in the compact case: on compact spaces, some kernels do metrize weak convergence *without* separating all finite signed measures. Second, combining it with Thm. 9 shows that the alleged proof of Claim 2 must be flawed. Another confirmation will be given by point (i) in Cor. 15, with an explicit counter-example constructed in its proof. However, to strengthen our claim, we now explicitly point out the flaw in the proof of Claim 2 by Simon-Gabriel and Schölkopf (2018).

5.1 Flaw in the proof of Claim 2 of Simon-Gabriel and Schölkopf

The flaw in the proof of Theorem 12 of Simon-Gabriel and Schölkopf (2018) (our Claim 2) resides in their auxiliary Lemma 20, which is essentially our Lem. 5, but without the assumption $\mathcal{H}_k \subset \mathscr{C}_0$. Their proof essentially consists in saying that, since (P_α) (denoted (μ_α) there) is bounded, it is relatively vaguely compact, so one can extract a subnet (P_β) that converges vaguely to a measure P' (denoted μ' there). They then try to identify the vague limit P' with the MMD- (or weak RKHS-) limit P (denoted μ there) of the original net (P_α), by arguing that weak and vague convergence coincide on $\mathcal{P}$, and that weak convergence implies MMD-convergence. Unfortunately, $\mathcal{P}$ is not closed in $\mathcal{M}$ for the vague topology, so nothing guarantees a priori that $P' \in \mathcal{P}$. And if $P' \notin \mathcal{P}$, then vague convergence to P' does not imply weak convergence to P' (Berg et al., 1984, Thm. 2.4.2), which is why the proof fails – irremediably.

We can go further and exhibit a counter-example for the previous failure, i.e. a bounded, continuous, $\int$s.p.d. kernel and a sequence (P_n) that converges to $P \in \mathcal{P}$ in MMD, but converges vaguely to another measure $P' \neq P$ in $\mathcal{M}$. Indeed, consider the kernel $\kappa := k + 1$ from the proof of Cor. 15(i) below. Let $\mathcal{K}$ be a compact neighborhood of $\boldsymbol{\xi}$ (which exists because $\mathcal{X}$ is locally compact) and choose a sequence $(P_n) \subset \mathcal{P}$ as in Lem. 12, i.e. such that $\|P_n\|_k \to 0$ and $P_n(\mathcal{K}) = 0$ for all n. By using the vague compactness of $\mathcal{B}_+ := \{\mu \in \mathcal{M}_+ \,|\, \mu(\mathcal{X}) \leq 1\}$ (Berg et al., 1984, Prop. 2.4.6) and extracting a subsequence if needed, we may assume that (P_n) converges vaguely to a measure $P' \in \mathcal{B}_+$. Applying Urysohn's lemma (Villani, 2010, Thm. I-33) to the compact set $\{\boldsymbol{\xi}\}$ and an open neighborhood $\mathcal{O} \subset \mathcal{K}$ of $\boldsymbol{\xi}$, we get a continuous function f whose support is contained in $\mathcal{K}$ and such that $f(\boldsymbol{\xi}) = 1$. Since $f \in \mathscr{C}_0$ and $P_n(f) = 0 < 1 = f(\boldsymbol{\xi}) = \delta_{\boldsymbol{\xi}}(f)$, P_n does not converge vaguely to $\delta_{\boldsymbol{\xi}}$, i.e.

$P' \neq \delta_{\boldsymbol{\xi}}$. Now κ is bounded, continuous and $\int$s.p.d., and induces the same metric than k on $\mathscr{P}$. So, since the KME of k maps the Dirac measure $\delta_{\boldsymbol{\xi}}$ to the null function in $\mathscr{H}_k$ (see proof of Prop.13), we get

$$\|P_n - \delta_{\boldsymbol{\xi}}\|_\kappa = \|P_n - \delta_{\boldsymbol{\xi}}\|_k = \|P_n\|_k \to 0 \ .$$

Hence $(P_n) \to \delta_{\boldsymbol{\xi}}$ in MMD, but (P_n) converges vaguely to a different measure P'.

Remark 14 *The sequence (P_n) converges neither weakly to P' nor weakly to $\delta_{\boldsymbol{\xi}}$, since weak convergence would imply vague and MMD convergence to the same limit, i.e. would imply $P' = \delta_{\boldsymbol{\xi}}$. Hence $P'(\mathcal{X}) \neq 1$ (otherwise, vague convergence would imply weak convergence, since both coincide on $\mathscr{P}$ (Berg et al., 1984, Cor. 2.4.3)), and since $P' \in \mathscr{B}_+$, we get $P'(\mathcal{X}) < 1$. So (P_n) illustrates a phenomenon called* mass escaping at infinity*, which vague convergence, contrary to weak convergence, cannot prevent.*

6. General case: $\mathcal{X}$ locally compact but non-compact and $\mathscr{H}_k \not\subset \mathscr{C}_0$

All previous sections assumed that $\mathscr{H}_k \subset \mathscr{C}_0$ (automatically satisfied when k continuous and $\mathcal{X}$ is compact). So one may naturally wonder whether this assumption could be dropped without replacement or at least extended. Cor. 15 shows that dropping it without replacement is not possible; but Cor. 17 proposes a slight extension.

Corollary 15 *Let $\mathcal{X}$ be a locally compact Hausdorff space that is not compact and for which there exists a $\mathscr{C}_0$-universal kernel (such as when $\mathcal{X}$ is metrizable and separable). Then*

(i) there exists a bounded continuous kernel that is $\int$s.p.d., but does not metrize the weak convergence of probability measures;

(ii) there exists a bounded, continuous, characteristic (to $\mathscr{P}$) kernel that is not $\int$s.p.d. but metrizes the weak convergence of probability measures.

Remark 16 *Note, however, that* some *kernels with non-vanishing RKHS functions do satisfy the characterization of Thm. 9. For example, Thm. 9 extends to any kernel of the form $k_c = k + c$ for $c > 0$ and $\mathscr{H}_k \not\subset \mathscr{C}_0$, since k_c and k induce the same MMD.*

Proof (i) By assumption, there exists a $\mathscr{C}_0$-universal kernel. Since that kernel is continuous and characteristic to $\mathcal{M}$ (see Section 1.4), by Prop. 13, there also exists a kernel k that is characteristic to $\mathscr{P}$ but not characteristic to $\mathcal{M}$, with $\mathscr{H}_k \subset \mathscr{C}_0$. Consider the new kernel $\kappa := k + 1$. Then κ is $\int$s.p.d. (Simon-Gabriel and Schölkopf, 2018, Thm. 8), but κ induces the same metric than k on the set of probability measures $\mathscr{P}$. Hence it does not metrize their weak convergence.

(ii) Let $\boldsymbol{\xi}$ be a point in $\mathcal{X}$. Let k be a $\mathscr{C}_0$-universal on $\mathcal{X}$. k is characteristic to $\mathscr{H}_k$ (Section 1.4), so, by Thm. 9, k metrizes the weak convergence over $\mathscr{P}$. Now, consider the kernel $\kappa(\boldsymbol{x}, \boldsymbol{y}) := \langle \delta_{\boldsymbol{x}} - \delta_{\boldsymbol{\xi}} \, , \, \delta_{\boldsymbol{y}} - \delta_{\boldsymbol{\xi}} \rangle_k$. κ is not $\int$s.p.d. (since the KME of $\delta_{\boldsymbol{\xi}}$ is the null function) but it induces the same RKHS metric than k on $\mathscr{P}$, that is $\|P - Q\|_\kappa = \|P - Q\|_k$ for any $P, Q \in \mathscr{P}$. Hence κ metrizes weak convergence on $\mathscr{P}$. (Remark: this implies that $\mathscr{H}_\kappa \not\subset \mathscr{C}_0$, which is also easy to check directly.)

The existence of a $\mathscr{C}_0$-universal kernel when $\mathcal{X}$ is a metrizable and separable LCH space is given by Thm. 4(b). ■

Let us mention that, in a side remark of Guilbart (1978, p.18), Guilbart already exhibits a theoretical construction of kernels on $\mathbb{R}$ that are $\int$s.p.d. but do not metrize weak convergence. Hence, Claim 2 was actually disproved before being written.

We finish with a slight generalization of Thm. 9 that encompasses some kernels whose RKHS is not contained in $\mathscr{C}_0$. The result builds on the same idea than in the proof of Cor. 15(ii).

Corollary 17 *Suppose that $\mathcal{X}$ is not compact and that $\mathcal{H}_k \subset \mathscr{C}_0$. Fix $a \geq 0$ and $P \in \mathscr{P}$ and define*

$$k_P^a(\boldsymbol{x}, \boldsymbol{y}) := \langle \delta_{\boldsymbol{x}} - P \, , \, \delta_{\boldsymbol{y}} - P \rangle_k + a = (\delta_{\boldsymbol{x}} - P) \otimes (\delta_{\boldsymbol{y}} - P)(k) + a \ .$$

Then k_P^a metrizes weak convergence of probability measures if and only if k is continuous and $\int$s.p.d.

Proof Since $k_P^a(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{x}, \boldsymbol{y}) - f_P(\boldsymbol{x}) - f_P(\boldsymbol{y}) + \|P\|_k^2 + a$, for any probability measures $S, T \in \mathscr{P}$, we get

$$\|S - T\|_{k_P^a}^2 = (S - T) \otimes (S - T)(k_P^a) = (S - T) \otimes (S - T)(k) = \|S - T\|_k^2 \ .$$

Hence k and k_P^a define the same metric on $\mathscr{P}$ and Thm. 9 concludes. ■

7. Conclusion

MMDs are at the heart of machine learning solutions to a variety of fundamental tasks including two-sample testing, sample quality measurement and goodness-of-fit testing, learning generative models, de novo sampling and quadrature, importance sampling, and thinning. While these applications benefit from the tractability of MMDs compared to more classical probability metrics, the validity of their results depends critically on the MMD's ability to ensure weak convergence. Simon-Gabriel and Schölkopf (2018) developed their Theorem 12 to provide a complete characterization of weak-convergence metrization for MMDs with bounded continuous kernels. However, our work shows that their characterization was incorrect and provides an alternative result that fully characterizes the weak-convergence metrization of MMDs with bounded $\mathscr{C}_0$ kernels. Surprisingly, we find that the compact and non compact cases are inherently different, the latter requiring *strictly* stronger conditions for the metrization. This suggests that the question of weak-convergence metrization by MMDs is more subtle than was previously thought. Our main results can also be seen as a converse to Sriperumbudur's Thm. 1, which in particular show that many popular kernels, particularly Stein kernels, can *fail* to metrize weak convergence, if one is not careful enough. In that spirit, we hope that our work will inform the selection of appropriate kernels and MMDs in the future and launch new inquiries into the metrization properties of other classes of MMDs.

Acknowledgments

CJSG was supported by the ETH Foundations of Data Science postdoctoral fellowship and is associate fellow of the Center for Learning Systems (ETH/MPI Tübingen). AB was supported by the Department of Engineering at the University of Cambridge, and this material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and the U. S. Army Research Office, and by the U.K. Ministry of Defence and the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number [EP/R018413/2]. We declare no conflict of interests.

References

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 3 edition, 2006.

Abdul Fatir Ansari, Jonathan Scarlett, and Harold Soh. A characteristic function approach to deep implicit generative modeling. In *CVPR*, 2020.

Christian Berg, Jens P. R. Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups Theory of Positive Definite and Related Functions*. Springer, 1984.

Francois-Xavier Briol, Alessandro Barp, Andrew B Duncan, and Mark Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944*, 2019.

Wilson Y. Chen, Lester Mackey, Jackson Gorham, François-Xavier Briol, and Chris J. Oates. Stein points. In *ICML*, 2018.

Wilson Ye Chen, Alessandro Barp, François-Xavier Briol, Jackson Gorham, Mark Girolami, Lester Mackey, Chris Oates, et al. Stein point Markov chain Monte Carlo. In *ICML*, 2019.

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *UAI*, 2010.

Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *Journal of Machine Learning Research*, 23(176):1–42, 2022.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *NeurIPS*, 2016.

John B. Conway. *A Course in Functional Analysis*. Springer, New York, 2 edition, 1994.

Gintare K. Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.

Yihao Feng, Dilin Wang, and Qiang Liu. Learning to draw samples with amortized Stein variational gradient descent. In *UAI*, 2017.

Futoshi Futami, Zhenghang Cui, Issei Sato, and Masashi Sugiyama. Bayesian posterior approximation via greedy particle optimization. In *AAAI*, 2019.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *ICML*, 2017.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13:723–773, 2012.

Christian Guilbart. *Etude des Produits Scalaires sur l'Espace des Mesures: Estimation par Projections.* PhD thesis, Université des Sciences et Techniques de Lille, 1978.

Liam Hodgkinson, Robert Salomone, and Fred Roosta. The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv:2001.09266*, 2020.

Wassily Hoeffding. Asymptotically optimal tests for multinomial distributions. *The Annals of Mathematical Statistics*, pages 369–401, 1965.

Jonathan Huggins and Lester Mackey. Random feature stein discrepancies. In *NeurIPS*, 2018.

Ferenc Huszár and David Duvenaud. Optimally-weighted herding is bayesian quadrature. In *UAI*, 2012.

Wittawat Jitkrittum, Wenkai Xu, Zoltan Szabo, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *NeurIPS*, 2017.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabas Poczos. MMD GAN: Towards deeper understanding of moment matching network. In *NeurIPS*, 2017.

Qiang Liu and Jason D. Lee. Black-box importance sampling. In *AISTATS*, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *NeurIPS*, 2016.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML*, 2016.

Nate Eltredge. Stackoverflow proof. `https://math.stackexchange.com/questions/3346313/prove-that-c-0x-is-separable-given-that-x-is-locally-compact-metric-space`, 2019. Accessed: 2023-03-10.

Yuchen Pu, Zhe Gan, Ricardo Henao, Chunyuan Li, Shaobo Han, and Lawrence Carin. VAE learning via Stein variational gradient descent. In *NeurIPS*, 2017.

Marina Riabiz, Wilson Chen, Jon Cockayne, Pawel Swietach, Steven A Niederer, Lester Mackey, Chris Oates, et al. Optimal thinning of MCMC output. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(4):1059–1081, 2022.

Štefan Schwabik. *Topics in Banach Space Integration.* Number 10 in Series in Real Analysis. World Scientific, 2005.

C.-J. Simon-Gabriel and B. Schölkopf. Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions. *JMLR*, 2018.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *ALT*, 2007.

Bharath K. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

Ingo Steinwart and Johanna F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542, 2021.

Ingo Steinwart, Don Hush, and Clint Scovel. Function classes that approximate the bayes risk. In *COLT*, 2006.

Dougal J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR*, 2017.

François Treves. *Topological Vector Spaces, Distributions and Kernels*. Academic Press, 1967.

Cédric Villani. *Intégration et analyse de Fourier*. ENS de Lyon, 2010.

Shengyu Zhu, Biao Chen, Pengfei Yang, and Zhitang Chen. Universal hypothesis testing with kernels: Asymptotically optimal tests for goodness of fit. In *AISTATS*, 2019.

Shengyu Zhu, Biao Chen, Zhitang Chen, and Pengfei Yang. Asymptotically optimal one- and two-sample testing with kernels. *IEEE Transactions on Information Theory*, 67(4): 2074–2092, 2021.

Appendix A. Translation of Some Results from Non-English References

For the convenience of the reader, we translate here some of the important results from Villani 2010 that we cite, since the original manuscript is in French.

Theorem 18 (Urysohn's Lemma. Translation of Thm-I.33 in Villani 2010) *Let $\mathcal{X}$ be a locally compact Hausdorff space, $\mathcal{O}$ an open and $\mathcal{K}$ a compact subset of $\mathcal{X}$, $\mathcal{K} \subset \mathcal{O}$. Then there exists a continuous function f with values in $[0,1]$, that is equal to the constant 1 on a neighborhood of $\mathcal{K}$, and whose support is compact and included in $\mathcal{O}$. In particular,*

$$\mathbb{1}_{\mathcal{K}} \leq f \leq \mathbb{1}_{\mathcal{O}},$$

where $\mathbb{1}_{\mathcal{K}}$ and $\mathbb{1}_{\mathcal{O}}$ designate the functions that are equal to 1 on $\mathcal{K}$ and $\mathcal{O}$ respectively, and 0 otherwise.

Theorem 19 (Ulam's Lemma. Thm. I-54 in Villani 2010) *Let $\mathcal{X}$ be a Polish space equipped with a σ-finite non-negative Borel measure μ (i.e., $\mathcal{X}$ is the countable union of sets A_k that satisfy $\mu(A_k) < \infty$). Then μ is regular (and concentrated on a σ-compact set).*

Theorem 20 (Ulam's Lemma for LCH spaces. Thm. I-56 in Villani 2010) *Let $\mathcal{X}$ be an LCH space where every open set is σ-compact, equipped with a non-negative Borel measure μ that is finite on compact sets. Then μ is regular.*

Theorem 21 (Riesz-Markov-Kakutani Representation. Thm.VI-61 in Villani 2010) *Let $\mathcal{X}$ be an LCH space. Then one can identify (i.e., find an isometric bigection) between*

- ▷ *the continuous linear forms Λ on the space $\mathscr{C}_0(\mathcal{X})$ of continuous functions on $\mathcal{X}$ that converge to 0 at infinity, equipped with the supremum norm convergence;*
- ▷ *the set of signed, regular, finite Borel measures on $\mathcal{X}$; i.e., measures that can be written as $\mu_+ - \mu_-$, where μ_+ and μ_- are non-negative, regular, finite Borel measures which are orthogonal to each other;*

via the following formula:

$$\Lambda f = \int f \, \mathrm{d}\mu := \int f \, \mathrm{d}\mu_+ - \int f \, \mathrm{d}\mu_- \ .$$

In short:

$$(\mathscr{C}_0)' = \mathcal{M}(X) \ .$$

Definition 22 (Radon Measures. Def. VI-66 in Villani 2010) *Let $\mathcal{X}$ be an LCH space equipped with its Borel σ-algebra, and let Ω be an open set in $\mathcal{X}$. A Radon measure on Ω if it is signed, locally finite (i.e., finite on any compact in Ω) and regular.*

Appendix B. Proof of Thm. 4

B.1 Proof of Point (a)

By Lem. 5 below, if k is continuous, then it metrizes the weak-$*$ topology on the set $\mathscr{P}$ of Radon probability measures. But, by Theorem V.5.1 in Conway (1994), $\mathcal{X}$ is homeomorphic to the subset of Dirac measures in $\mathscr{P}$, i.e. to $\{\delta_x \,|\, x \in \mathcal{X}\}$, when equipped with the weak-$*$ topology. Hence $d_k(x,y) = \|\delta_x - \delta_y\|_k$ metrizes $\mathcal{X}$. Conversely, if k metrizes $\mathcal{X}$, then Lemma 4.29 (point $iv \Rightarrow i$) in Steinwart and Christmann (2008) shows that k is continuous.

B.2 Proof of Point (b)

To prove (b), we are going to prove the following, more complete set of equivalences.

Theorem 23 *On a LCH space $\mathcal{X}$ the following is equivalent.*

(i) $\mathcal{X}$ is metrizable and separable.
(ii) $\mathcal{X}$ is σ-compact and there exists a continuous $\mathscr{C}_0$-universal kernel k on $\mathcal{X}$.
(iii) $\mathcal{X}$ is second countable.
(iv) $\mathscr{C}_0(\mathcal{X})$ is separable.

Proof **(iii)⇔(i).** Since $\mathcal{X}$ is LCH, $\mathcal{X}$ is completely regular (Aliprantis and Border, 2006, Cor 2.74) and hence regular. Urysohn's metrization theorem concludes (Aliprantis and Border, 2006, Thm. 3.40).

(i) ⇔ (iv). An LCH space is completely regular (Aliprantis and Border, 2006, Cor 2.74). So, if $\mathcal{X}$ is compact, then Theorem V.6.6 by Conway (1994) concludes. Otherwise, let $\mathcal{X}_\infty$ be the one-point compactification of $\mathcal{X}$ (Aliprantis and Border, 2006, Thm. 2.72). We saw above that $\mathcal{X}$ is metrizable and separable iff $\mathcal{X}$ is second countable, and, by Thm 3.44 of the same reference, the latter holds iff $\mathcal{X}_\infty$ is metrizable. Since $\mathcal{X}_\infty$ is compact and Hausdorff (hence completely regular), this is equivalent to $\mathscr{C}_b(\mathcal{X}_\infty)$ (equipped with its canonical supremum norm topology) being separable (Conway, 1994, Thm. V.6.6), which in turn happens iff its hyperplane $H := \{f \in \mathscr{C}_b(\mathcal{X}_\infty) \,:\, f(\infty) = 0\}$ is separable as well. Conclude by noting that H is homeomorphic to $\mathscr{C}_0(\mathcal{X})$.

(iv) ⇒(ii). We have already shown that, since $\mathscr{C}_0$ is separable, $\mathcal{X}$ is second countable, which, in turn, implies that $\mathcal{X}$ is σ-compact (Aliprantis and Border, 2006, Lem 2.76). To show that there exists a universal kernel, we now follow the proof of Thm 2 in Steinwart et al. 2006. Let $\{f_n\}_n$ be an at most countable dense subset of $\mathscr{C}_0$. For any integer $n \geq 0$, define $\Phi_n(x) := 2^{-n} f_n / \|f_n\|_\infty$ if $f_n \neq 0$ and $\Phi_n = 0$ otherwise. Then, clearly, $\Phi(x) := (\Phi_n(x))_n$ satisfies $\Phi(x) \in \ell_2$ for all $x \in \mathcal{X}$, hence $k(x, y) := \langle \Phi(x) \,, \Phi(y) \rangle_{\ell_2}$, where $x, y \in \mathcal{X}$, defines a kernel on $\mathcal{X}$ with feature map $\Phi : \ \mathcal{X} \ \longrightarrow \ \ell_2$. Fix $f \in \mathscr{C}_0$ and $\epsilon > 0$. There exists an integer n such that $\|f_n - f\|_\infty \leq \epsilon$. Define the function $w := 2^n \|f_n\|_\infty e_n$ where $(e_n)_n$ is the canonical orthonormal basis of ℓ_2. This gives $\langle w \,, \Phi(x) \rangle = f_n(x)$ for all $x \in \mathcal{X}$, and since $\{\langle v \,, \Phi(x) \rangle \,:\, v \in \ell_2\}$ is the RKHS of k (Steinwart and Christmann, 2008, Thm. 4.21), we obtain the universality of k. It remains to be shown that k is continuous. To do so, we show that that Φ is continuous. Indeed, let $(x_\alpha)_\alpha$ be a net that converges to x in $\mathcal{X}$. Fix $\epsilon > 0$.

$$\|\Phi(x_\alpha) - \Phi(x)\|_2^2 \leq \sum_{n \geq 0} |\Phi_n(x_\alpha) - \Phi_n(x)|^2 = \sum_{n \geq 0} \frac{1}{2^{2n}} |g_n(x_\alpha) - g_n(x)|^2 \ ,$$

where we defined $g_n(x) := f_n(x) / \|f_n\|_\infty$ (or 0 if $f_n = 0$). Since $|g_n| \leq 1$, the summands verify $|g_n(x_\alpha) - g_n(x)|^2 / 2^{2n} \leq 12^{2(n-1)}$ for $n \geq 1$. So let N be an integer such that $\sum_{n > N} \frac{1}{2^{2(n-1)}} \leq \epsilon/2$ and let A be an index such that $\sum_{n=0}^{N} |g_n(x_\alpha) - g(x)|^2 / 2^{2n} \leq \epsilon/2$ whenever $\alpha > A$ (which exists, since we are considering a finite sum of continuous functions). The continuity then follows from the following.

$$\|\Phi(x_\alpha) - \Phi(x)\|_2^2 \leq \sum_{n=0}^{N} \frac{1}{2^{2n}} |g_n(x_\alpha) - g_n(x)|^2 + \sum_{n > N} \frac{1}{2^{2(n-1)}} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \leq \epsilon \ .$$

(ii) $\Rightarrow$(iv). We adapt the proof of Thm 2, point $(i) \Rightarrow (ii)$, given by Steinwart et al. (2006). Let k be a continuous universal kernel on $\mathcal{X}$ and let $\Phi : \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathcal{H}_k \\ x & \longmapsto & k(.,x) \end{array}$ be its canonical feature map. Then Φ is continuous (Steinwart and Christmann, 2008, Lem 4.29). Since $\mathcal{X}$ is σ-compact, let $(K_i)_i$ be an at most countable compact cover of $\mathcal{X}$. For each i, $\Phi(K_i)$ is compact, and, since $\mathcal{H}_k$ is a metric space, $\Phi(K_i)$ is separable. Hence $\Phi(X) = \cup_i \Phi(K_i)$ is separable, and consequently, so is $\mathcal{H}_k = \mathrm{cl}(\mathrm{span}\, \Phi(X))$, the closed span of $\Phi(\mathcal{X})$ in $\mathcal{H}_k$. Since $\mathcal{H}_k$ is dense in $\mathcal{C}_0$, we then obtain that $\mathcal{C}_0$ is separable.

Alternative proof of (ii) $\Rightarrow$(iv). Point (a) shows that $\mathcal{X}$ is metrizable. Conclude by noting that a σ-compact metrizable spaces is separable, since it can be covered by countably many compacts and, being a metrizable space, any compact is separable. ■

Proof [Alternative proof of (i) $\Leftrightarrow$ (iv)] **(i) $\Rightarrow$(iv).** We will adapt the proof given by Conway (1994, Thm. V.6.6) for compact spaces. Let d be a metric that metrizes the topology of $\mathcal{X}$. Since $\mathcal{X}$ is separable, let $(x_k)_k$ be a dense sequence in $\mathcal{X}$. For any positive integer n, let B_k^n be the open ball of radius $1/n$ centered on x_k. For any n, $(B_k^n)_k$ is an open cover of $\mathcal{X}$. Since $\mathcal{X}$ is a metric space, apply Theorems 3.22 and 2.90 in Aliprantis and Border (2006) to construct a continuous locally finite partition of unity $(f_k^n)_k$ subordinated to $(B_k^n)_k$ (see Def 2.89 therein). Let $\mathcal{Y}$ be the rational linear span of $(f_k^n)_{k,n}$, i.e., the finite linear combinations of functions f_k^n with rational coefficients. $\mathcal{Y}$ is countable. We will show that $\mathcal{Y}$ is dense in $\mathcal{C}_0(\mathcal{X})$.

Fix $f \in \mathcal{C}_0$ and $\epsilon > 0$. Since f vanishes at infinity, it is uniformly continuous. So there is a $\delta > 0$ such that $|f(x) - f(y)| \leq \epsilon/2$ whenever $d(x,y) < \delta$. Choose $n > 1/\delta$. Consider the cover $(B_k^n)_k$. If $x \in B_k^n$, $d(x, x_k) \leq 1/n \leq \delta$; hence $|f(x) - f(x_k)| \leq \epsilon/2$. Let α_k be a rational number such that $|\alpha_k - f(x_k)| \leq \epsilon/2$. Let $g := \sum_k \alpha_k f_k^n$; so $g \in \mathcal{Y}$. For every $x \in \mathcal{X}$,

$$\begin{aligned} |f(x) - g(x)| &\leq |\sum_k f(x) f_k^n(x) - \alpha_k f_k^n(x)| \\ &\leq \sum_k |f(x) - \alpha_k| f_k^n(x). \end{aligned}$$

Examine each of these summands. If $x \in B_k^n$, then $|f(x) - \alpha_h| \leq |f(x) - f(x_k)| + |f(x_k) - \alpha_h| \leq \epsilon$; otherwise $f_k^n(x) = 0$. In both cases $|f(x) - \alpha_k| f_k^n(x) \leq \epsilon f_k^n(x)$, hence $|f(x) - g(x)| \leq \epsilon \sum_k f_k^n(x) = \epsilon$. Thus $\|f - g\|_\infty \leq \epsilon$ and $\mathcal{Y}$ is dense in $\mathcal{C}_0$. Hence $\mathcal{C}_0$ is separable.

(iv) $\Rightarrow$(i). We will prove that, if $\mathcal{C}_0$ is separable, then $\mathcal{X}$ is second countable, which concludes, since we already showed that (iii) and (i) are equivalent. The proof follows Nate Eltredge (2019). Let $\{f_n\}_n$ be a countable dense subset of $\mathcal{C}_0$, and for each n let $U_n = \{x \in \mathcal{X} : f_n(x) > 1/2\}$, which is an open subset of $\mathcal{X}$. We claim that $\{U_n\}_n$ is a countable base for the topology of $\mathcal{X}$. For let $x \in \mathcal{X}$ and let V be an open neighborhood of x. Then by Urysohn's lemma for locally compact Hausdorff spaces, there exists a function f compactly supported inside V with $f(x) = 1$. In particular $f \in \mathcal{C}_c(\mathcal{X}) \subset \mathcal{C}_0$, so by density, we can find some f_n with $\|f - f_n\|_\infty < 1/2$. Then we have $f_n(x) > 1/2$ so $x \in U_n$. Moreover, if $y \in U_n$ then $f_n(y) > 1/2$ and so $f(y) > 0$, which implies $y \in V$. Therefore $U_n \subset V$. This proves that $\{U_n\}_n$ is a base. ■