Journal of Machine Learning Research 25 (2024) 1-50 Submitted 10/22; Revised 9/24; Published 12/24

Targeted Separation and Convergence with Kernel Discrepancies

Alessandro Barp† ALESSANDRO.BARP@UCL.AC.UK
University College London & The Alan Turing Institute, GB

Carl-Johann Simon-Gabriel†⋆ CJSG@MIRELO.AI
Mirelo AI

Mark Girolami MAG92@CAM.AC.UK
University of Cambridge & The Alan Turing Institute, GB

Lester Mackey† LMACKEY@MICROSOFT.COM
Microsoft Research, New England, US

Editor: Bharath Sriperumbudur

Abstract

Maximum mean discrepancies (MMDs) like the kernel Stein discrepancy (KSD) have grown central to a wide range of applications, including hypothesis testing, sampler selection, distribution approximation, and variational inference. In each setting, these kernel-based discrepancy measures are required to (*i*) separate a target P from other probability measures or even (*ii*) control weak convergence to P. In this article we derive new sufficient and necessary conditions to ensure (*i*) and (*ii*). For MMDs on separable metric spaces, we characterize those kernels that separate Bochner embeddable measures and introduce simple conditions for separating all measures with unbounded kernels and for controlling convergence with bounded kernels. We use these results on $\mathbb{R}^d$ to substantially broaden the known conditions for KSD separation and convergence control and to develop the first KSDs known to exactly metrize weak convergence to P. Along the way, we highlight the implications of our results for hypothesis testing, measuring and improving sample quality, and sampling with Stein variational gradient descent.

Keywords: Maximum mean discrepancy, kernel Stein discrepancy, targeted separation, targeted weak convergence control, enforcing tightness

1. Introduction

Maximum mean discrepancies (MMDs) like the Langevin kernel Stein discrepancy (KSD) are kernel-based discrepancy measures widely used for hypothesis testing (Gretton et al., 2012; Liu et al., 2016; Chwialkowski et al., 2016), sampler selection and tuning (Gorham and Mackey, 2017), parameter estimation (Briol et al., 2019; Barp et al., 2019; Dziugaite et al., 2015), generalized Bayesian inference (Chérief-Abdellatif and Alquier, 2020; Matsubara et al., 2021, 2022; Dellaporta et al., 2022), discrete approximation and numerical integration (Chen et al., 2019, 2018; Barp et al., 2022b), control variate design (Oates et al.,

†. These authors contributed equally.
⋆. Work done while at AWS Lablets.

2014, 2019; Sun et al., 2023), compression (Riabiz et al., 2022), and bias correction (Liu and Lee, 2017; Hodgkinson et al., 2020; Riabiz et al., 2022).

Each MMD uses a kernel function to measure the integration error between a pair of probability measures Q and P, and, in each setting above, their successful application relies on either P-*separation*, that is $\mathrm{MMD}(\mathrm{Q},\mathrm{P}) > 0$ whenever $\mathrm{Q} \neq \mathrm{P}$, or P-*convergence control*, namely $\mathrm{MMD}(\mathrm{Q}_n,\mathrm{P}) \to 0$ implies $\mathrm{Q}_n \to \mathrm{P}$ weakly. Unfortunately, these properties have so far only been established under overly restrictive assumptions, e.g., for Q with continuously differentiable log densities (Chwialkowski et al., 2016; Liu et al., 2016; Barp et al., 2019), for P with strongly log concave tails and Lipschitz log density gradients $\boldsymbol{s}_\mathrm{p} = \partial \log \mathrm{p}$ (Gorham and Mackey, 2017), or for bounded MMD kernels (Sriperumbudur et al., 2010; Sriperumbudur, 2016; Simon-Gabriel and Schölkopf, 2018; Simon-Gabriel et al., 2023). In this work, by fixing P as the target measure and allowing Q to vary, we establish new broadly applicable conditions for P-separation and P-convergence control. Our main results include

- **Bochner P-separation with MMDs:** Theorem 2 exactly characterizes those MMDs that separate P from Bochner embeddable measures on general Radon spaces. For MMDs with bounded kernels, this result exposes an important relationship between separation and convergence: separating P from all probability measures is equivalent to controlling P-convergence for *tight* sequences $(\mathrm{Q}_n)_n$.

- **Score P-separation with KSDs:** Theorem 3 shows that KSDs with standard characteristic kernels separate P from all measures Q that finitely integrate the score $\boldsymbol{s}_\mathrm{p}$. This strengthens past work that only established separation from Q with continuously differentiable log densities (Chwialkowski et al., 2016; Barp et al., 2019).

- **L^2 P-separation with KSDs:** Theorems 4 and 5 show that KSDs with standard translation-invariant kernels separate P from all measures with densities q and finitely square-integrable $\mathrm{q}\boldsymbol{s}_\mathrm{q}$ and $\mathrm{q}\boldsymbol{s}_\mathrm{p}$. This strengthens past work that provided no examples of L^2-separating kernels (Liu et al., 2016).

- **General P-separation with MMDs:** Theorem 6 provides a simple sufficient condition for general P-separation: any MMD—even one with an unbounded kernel—separates P from **all** probability measures and controls tight convergence to P if the bounded functions in its associated reproducing kernel Hilbert space (RKHS) are P-separating. All of our remaining results explicitly check this new convenient condition.

- **General P-separation with KSDs:** Theorem 9 shows that KSDs with standard translation-invariant kernels separate P from **all** probability measures and control tight P-convergence whenever $\boldsymbol{s}_\mathrm{p}$ is continuous and grows at most root-exponentially. Prior P-separation results applied only to a small subset of these targets, those with strongly log concave tails and Lipschitz $\boldsymbol{s}_\mathrm{p}$ (Gorham and Mackey, 2017; Huggins and Mackey, 2018; Chen et al., 2018).

- **Enforcing tightness with MMDs:** Theorem 10 provides a new sufficient condition for *enforcing tightness*, i.e., for ensuring that $(\mathrm{Q}_n)_n$ is tight whenever $\mathrm{MMD}(\mathrm{Q}_n,\mathrm{P}) \to 0$: an MMD enforces tightness if elements of its RKHS suitably bound the indicators of compact sets. Prior tightness-enforcing guarantees relied on a much stronger

condition: the presence of a coercive (and hence unbounded) function in the RKHS (Gorham and Mackey, 2017; Huggins and Mackey, 2018; Chen et al., 2018; Hodgkinson et al., 2020).

- **Metrizing P-convergence with KSDs:** Building on Theorem 10, Theorem 12 develops the first KSDs known to *metrize* weak convergence to P (i.e., $\mathrm{KSD}(\mathrm{Q}_n, \mathrm{P}) \to 0 \Leftrightarrow \mathrm{Q}_n \to \mathrm{P}$ weakly) by constructing **bounded** convergence-controlling Stein kernels. Since all prior convergence-controlling KSDs featured unbounded Stein kernels, these are also the first KSDs known to satisfy the Stein variational gradient descent convergence assumptions of Liu (2017) (see Application 4).

- **Failing to control P-convergence:** Finally, Theorem 13 provides new necessary conditions for an MMD to control P-convergence which notably fail to be satisfied when standard KSDs are paired with heavy-tailed targets.

As we highlight in the sections to follow, these results have immediate implications for a variety of inferential tasks in machine learning and statistics including goodness-of-fit testing (Applications 1 and 2), measuring and improving sample quality (Application 3), and variational inference (Application 4).

Notation For a given separable metric space $\mathcal{X}$, we let $\mathscr{C}(\mathbb{R}^d)$ denote the space of continuous $\mathbb{R}^d$-valued functions on $\mathcal{X}$. When $\mathcal{X} = \mathbb{R}^d$, we say that the derivative of a set of $\mathbb{R}^\ell$-valued functions exists, if the functions in that set are differentiable, and we additionally denote by $\mathscr{C}^\ell(\mathbb{R}^d)$ the space of ℓ-times continuously differentiable $\mathbb{R}^d$-valued functions on $\mathcal{X}$ (i.e., $f \in \mathscr{C}^\ell(\mathbb{R}^d)$ if the partial derivatives of order ℓ of f^i exist and are continuous for $i \in [d] \equiv \{1, \dots, d\}$). We let ∂f denote the vector of partial derivatives of a function f, and, for each multi-index p, let $\partial^p f$ denote the p-th partial derivatives of f. When $d = 1$ or $\ell = 0$ we will use the abbreviations $\mathscr{C}^\ell \equiv \mathscr{C}^\ell(\mathbb{R}^1)$ or $\mathscr{C}(\mathbb{R}^d) \equiv \mathscr{C}^0(\mathbb{R}^d)$. Decay requirements will appear as subscripts: $\mathscr{C}_b(\mathbb{R}^d)$, $\mathscr{C}_c(\mathbb{R}^d)$, and $\mathscr{C}_0(\mathbb{R}^d)$ will respectively denote the spaces of $\mathbb{R}^d$-valued continuous functions that are bounded, compactly supported, and vanishing at infinity. Analogously, for each function $h : \mathcal{X} \to [0, \infty)$, $\mathscr{C}_h(\mathbb{R}^d)$ and $\mathscr{C}_{0,h}(\mathbb{R}^d)$ respectively denote the spaces of $\mathbb{R}^d$-valued continuous functions f with $f/(1+h)$ bounded or vanishing at infinity. Recall a function $f : \mathcal{X} \to \mathbb{R}^a$ vanishes at infinity if $\forall \epsilon > 0$ there exists a compact set C s.t., $\sup_{x \in C^c} \|f(x)\| \leq \epsilon$, where C^c is the set complement of C, and $\|\cdot\|$ the Euclidean norm. For any function of two arguments $K(y, x)$, we write $K_x \equiv K(\cdot, x)$, and $K \in \mathscr{C}_b^{(1,1)}(\mathbb{R}^d)$ if $\partial_y^{p_y} \partial_x^{p_x} K(y, x)$ exists, is bounded, and is separately continuous for multi-indices satisfying $\|p_x\|_1, \|p_y\|_1 \leq 1$, where $\|\cdot\|_1$ is the Euclidean 1-norm (i.e., the multi-index absolute value). Given a map $T : \mathcal{S}_1 \to \mathcal{S}_2$ between sets, we denote the image of T by $T(\mathcal{S}_1) \equiv \{T(s) : s \in \mathcal{S}_1\}$. Given a measure μ and a μ-integrable function h, we denote integration by $\mu h \equiv \int h(x) \mu(\mathrm{d}x)$, and we shall omit the domain of integration, which is always $\mathcal{X}$. Some additional notation for the appendices is presented in Appendix A.

2. Maximum Mean Discrepancies and Kernel Stein Discrepancies

We begin by extending the usual notions of maximum mean discrepancy and kernel Stein discrepancy to accommodate both arbitrary probability measures Q and unbounded kernels.

Throughout, we let $\mathcal{P}$ the denote set of (Borel) probability measures on a separable metric space $\mathcal{X}$. Moreover, for any function $f : \mathcal{X} \to \mathbb{R}^\ell$, we let $\mathcal{P}_f \equiv \{\mathrm{Q} \in \mathcal{P} : \|f\| \in \mathrm{L}^1(\mathrm{Q})\}$ denote the set of probability measures that finitely integrate $\|f\| \equiv \|\cdot\| \circ f$.

2.1 Maximum mean discrepancies

Consider a (reproducing) kernel k on $\mathcal{X}$ with reproducing kernel Hilbert space $\mathcal{H}_k$ (Aronszajn, 1950; Schwartz, 1964). Traditionally, the associated kernel MMD is defined as the worst-case integration error across test functions in the RKHS unit norm ball $\mathcal{B}_k$ (Gretton et al., 2012):

$$\mathrm{MMD}_k(\mathrm{Q}, \mathrm{P}) \equiv \sup_{h \in \mathcal{B}_k} |\mathrm{Q}h - \mathrm{P}h| \,. \tag{1}$$

However, the expression $\mathrm{Q}h - \mathrm{P}h$ is not well defined when either (i) both $\mathrm{Q}h$ and $\mathrm{P}h$ are infinite or (ii) h is not integrable under Q. Unfortunately, both of these cases can occur when k is unbounded as $\mathcal{B}_k$ then necessarily contains an unbounded test function (see Lemma 3).

Since we are interested in a fixed target measure P, we address the first issue by focusing on kernels with finitely P-integrable test functions, i.e., with $\mathcal{B}_k \subseteq \mathrm{L}^1(\mathrm{P})$. To address the second issue, we extend the MMD definition (1) to all probability measures Q by taking the supremum only over the Q-integrable elements of $\mathcal{B}_k$, that is, h with either $h_+ \equiv \max(h, 0) \in \mathrm{L}^1(\mathrm{Q})$ or $h_- \equiv \max(-h, 0) \in \mathrm{L}^1(\mathrm{Q})$. In fact, since $\mathcal{B}_k$ is a symmetric set, considering only h with $h_+ \in \mathrm{L}^1(\mathrm{Q})$ suffices to ensure $|\mathrm{Q}h - \mathrm{P}h|$ is well defined and belongs to $[0, \infty]$.

Definition 1 (Maximum mean discrepancy (MMD))**.** *For a given kernel k, define the set of* embeddable probability measures $\mathcal{P}_{\mathcal{H}_k} \equiv \{\mathrm{Q} \in \mathcal{P} : \mathcal{H}_k \subseteq \mathrm{L}^1(\mathrm{Q})\}$*. For any target measure $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_k}$, we define the* maximum mean discrepancy $\mathrm{MMD}_k(\cdot, \mathrm{P}) : \mathcal{P} \to [0, \infty]$ *by*

$$\mathrm{MMD}_k(\mathrm{Q}, \mathrm{P}) \equiv \sup_{h \in \mathcal{B}_k : h_+ \in \mathrm{L}^1(\mathrm{Q})} |\mathrm{Q}h - \mathrm{P}h| \,. \tag{2}$$

Remark 1 (Embeddability)**.** *We show in Appendix C that (i) the embeddability condition $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_k}$ holds if and only if $x \mapsto k(\cdot, x)$ is* Pettis integrable by P *and (ii) Pettis integrability in turn implies that the* kernel mean $\int k(\cdot, x)\, \mathrm{dP}(x)$ *belongs to the RKHS $\mathcal{H}_k$. See Definition 6 for the definition of Pettis integrability.*

As we show in Appendix C.4, one user-friendly sufficient condition for $\mathrm{Q} \in \mathcal{P}_{\mathcal{H}_k}$ is *Bochner-embeddability*, that is, $\mathrm{Q} \in \mathcal{P}_{\sqrt{k}}$ where $\sqrt{k}$ represents the function $x \mapsto \sqrt{k(x, x)}$. When $\mathcal{H}_k$ is separable, Carmeli et al. (2006) proved that one can alternatively check the weaker condition $\iint |k(x, y)| \mathrm{dQ}(x) \mathrm{dQ}(y) < \infty$. The next proposition summarizes these convenient embeddability conditions.

Proposition 1 (Embeddability conditions)**.** *The following claims hold true.*

(a) $\mathcal{P}_{\sqrt{k}} \subsetneq \mathcal{P}_{\mathcal{H}_k}$.

(b) *If $\mathcal{H}_k$ is separable, $\iint |k(x, y)| \mathrm{dQ}(x) \mathrm{dQ}(y) < \infty$ implies $\mathrm{Q} \in \mathcal{P}_{\mathcal{H}_k}$ (Carmeli et al., 2006, Cor. 4.3).*

Remark 2 (Sufficient condition for separability). *Note that when $\mathcal{X}$ is a locally compact topological space, for $\mathcal{H}_k$ to be separable, it is sufficient that $\mathcal{H}_k \subseteq \mathscr{C}$ (Carmeli et al., 2006, Cor. 5.2). Moreover, $\mathcal{H}_k \subseteq \mathscr{C} \Leftrightarrow k$ is locally bounded[1] and $k_x \in \mathscr{C}$ for each x (Carmeli et al., 2006, Prop. 5.1).*

Moreover, when both Q and P are embeddable, the MMD can be re-expressed as a convenient double-integral (Simon-Gabriel and Schölkopf, 2018, Prop. 13).

Proposition 2 (MMD as a double integral). *If $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_k}$ and $\mathrm{Q} \in \mathcal{P}_{\mathcal{H}_k}$, then*

$$\mathrm{MMD}_k^2(\mathrm{Q}, \mathrm{P}) = \iint k(x, y)\mathrm{d}(\mathrm{Q} - \mathrm{P})(x)\mathrm{d}(\mathrm{Q} - \mathrm{P})(y).$$

2.2 Kernel Stein discrepancies

Building on the Stein discrepancy formalism of Gorham and Mackey (2015) and the zero-mean reproducing kernel theory of Oates et al. (2014), Chwialkowski et al. (2016); Liu et al. (2016); Gorham and Mackey (2017) concurrently developed special MMDs that can be computed without any explicit integration under the target P. When discussing these *Langevin KSDs* we will restrict our focus to $\mathcal{X} = \mathbb{R}^d$ and assume the target P has a strictly positive density p with respect to Lebesgue measure. We will also make use of a matrix-valued kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$ which generates an RKHS $\mathcal{H}_K$ of vector-valued functions; for an introduction to vector-valued RKHSes, please see Appendix B.

The Langevin KSD is defined in terms of a matrix-valued *base kernel* K and the differential operator

$$\mathcal{S}_{\mathrm{p}}(v) \equiv \frac{1}{\mathrm{p}} \nabla \cdot (\mathrm{p}v) \equiv \frac{1}{\mathrm{p}} \sum_j \partial_{x^j}(\mathrm{p}v^j),$$

known as the *Langevin Stein operator* in the machine learning and statistics communities (Gorham and Mackey, 2015; Anastasiou et al., 2023), which, under mild conditions, maps $\mathbb{R}^d$-valued functions $v = (v^1, \ldots, v^d) : \mathbb{R}^d \to \mathbb{R}^d$ to $\mathbb{R}$-valued functions with mean zero under the target, $\mathrm{P}\mathcal{S}_{\mathrm{p}}(v) = 0$. Specifically, for K chosen so that $\mathcal{S}_{\mathrm{p}}(v)$ has expectation zero under P for each $v \in \mathcal{H}_K$, Chwialkowski et al. (2016); Gorham and Mackey (2017); Barp et al. (2019) defined[2] the Langevin KSD as an integral probability metric (Müller, 1997) over $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K)$:

$$\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) \equiv \sup_{v \in \mathcal{B}_K} |\mathrm{Q}\mathcal{S}_{\mathrm{p}}(v)| = \sup_{v \in \mathcal{B}_K} |\mathrm{Q}\mathcal{S}_{\mathrm{p}}(v) - \mathrm{P}\mathcal{S}_{\mathrm{p}}(v)|. \tag{3}$$

However, $\mathcal{S}_{\mathrm{p}}(v)$ is often unbounded so that, for the same reasons described in Section 2.1, the expression (3) need not be well defined for all $\mathrm{Q} \in \mathcal{P}$. To enable meaningful KSD evaluation for all probability measures, we follow the recipe of Definition 1 to extend the definition of KSD to all $\mathrm{Q} \in \mathcal{P}$.

1. Recall that a function f from a topological space to a normed space is *locally bounded* if every point in its domain has a neighbourhood U for which the restriction of f to U is bounded.
2. The distinct definition of Liu et al. (2016) coincides with (3) under the assumptions of their Thm. 3.8.

Definition 2 (Kernel Stein discrepancy (KSD)). *Consider a target* $\mathrm{P} \in \mathcal{P}$ *with density* $\mathrm{p} > 0$ *and matrix-valued base kernel* K *for which the set* $\mathrm{p}\mathcal{H}_K = \{\mathrm{p}h : h \in \mathcal{H}_K\}$ *consists of partially differentiable functions. When* $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K) \subseteq \mathrm{L}^1(\mathrm{P})$ *and* $\mathrm{P}(\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K)) = \{0\}$*, we define the* kernel Stein discrepancy $\mathrm{KSD}_{K,\mathrm{P}} : \mathcal{P} \to [0, \infty]$ *by*

$$\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) \equiv \sup_{v \in \mathcal{B}_K : \mathcal{S}_{\mathrm{p}}(v)_+ \in \mathrm{L}^1(\mathrm{Q})} |\mathrm{Q}\mathcal{S}_{\mathrm{p}}(v)| \,. \tag{4}$$

Remark 3 (Relation to prior definitions of KSD). *For scalar kernels, Definition 2 is identical to the definition of KSD given in two of the papers that originally defined KSDs, Chwialkowski et al. (2016, Sec. 2.1) and Gorham and Mackey (2017, Sec. 3.1 with* $\|\cdot\| = \|\cdot\|_2$*), except for the extra constraint* $\mathcal{S}_{\mathrm{p}}(v)_+ \in \mathrm{L}^1(\mathrm{Q})$ *that we include simply to ensure that the* $\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q})$ *is well defined for all probability measures* Q*. Moreover, for probability measures* Q *satisfying the constraint* $\mathcal{S}_{\mathrm{p}}(v)_+ \in \mathrm{L}^1(\mathrm{Q})$ *for all* $v \in \mathcal{B}_K$ *our definition exactly recovers those given by Chwialkowski et al. and Gorham and Mackey. However, unlike Definition 2, the prior definitions of KSD from Chwialkowski et al. and Gorham and Mackey are not well defined for probability measures* Q *failing to satisfy the extra constraint, even though this restriction is not discussed explicitly in either work.*

Under additional assumptions, like Bochner embeddability of P and Q and continuous differentiability of K and p, prior work showed that the KSD (4) is equivalent to an MMD with a scalar *Stein kernel* k_{p} and that $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K)$ defines a *Stein RKHS* $\mathcal{H}_{k_{\mathrm{p}}}$ of scalar-valued functions (Oates et al., 2014; Chwialkowski et al., 2016; Liu et al., 2016; Gorham and Mackey, 2017; Barp et al., 2019). Our next result, proved in Appendix C.5, shows that **no** additional assumptions are necessary: $\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = \mathrm{MMD}_{k_{\mathrm{p}}}(\mathrm{Q}, \mathrm{P})$ and $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K) = \mathcal{H}_{k_{\mathrm{p}}}$ whenever the left-hand side quantities are well defined.

Theorem 1 (KSD as MMD). *Consider a target* $\mathrm{P} \in \mathcal{P}$ *with density* $\mathrm{p} > 0$ *and matrix-valued base kernel* K *for which* $\mathrm{p}\mathcal{H}_K$ *consists of partially differentiable functions. Then* $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K)$ *is the* Stein RKHS $\mathcal{H}_{k_{\mathrm{p}}}$ *induced by the* Stein kernel[3]

$$k_{\mathrm{p}}(x, y) \equiv \frac{1}{\mathrm{p}(x)\mathrm{p}(y)} \nabla_y \cdot \nabla_x \cdot (\mathrm{p}(x) K(x, y) \mathrm{p}(y)) \,. \tag{5}$$

Moreover, for target measures with zero-mean Stein RKHSes, i.e., for P *in*

$$\mathcal{P}_{K,0} \equiv \{\mathrm{Q} \in \mathcal{P} \text{ with density } \mathrm{q} > 0 \colon \mathrm{q}\mathcal{H}_K \text{ are partially differentiable functions}, \\ \mathcal{S}_{\mathrm{q}}(\mathcal{H}_K) \subseteq \mathrm{L}^1(\mathrm{Q}), \text{and } \mathrm{Q}(\mathcal{S}_{\mathrm{q}}(\mathcal{H}_K)) = \{0\}\},$$

the KSD matches the Stein kernel MMD:

$$\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = \mathrm{MMD}_{k_{\mathrm{p}}}(\mathrm{Q}, \mathrm{P}) \quad \text{for all} \quad \mathrm{Q} \in \mathcal{P}.$$

Remark 4 (Scalar kernel KSD). *When* $K = k \,\mathrm{Id}$ *for a scalar kernel* k*, we will say that* k_{p} *is induced by* k *and write* $\mathcal{P}_{k,0} \equiv \mathcal{P}_{k\mathrm{Id},0}$*. In this case,*

$$k_{\mathrm{p}}(x, y) = \sum\nolimits_{i=1}^{d} \frac{1}{\mathrm{p}(x)\mathrm{p}(y)} \, \partial_{x^i} \partial_{y^i} (\mathrm{p}(x) k(x, y) \mathrm{p}(y)).$$

3. Note we have $k_{\mathrm{p}}(x, y) = \sum_{i,j=1}^{d} \frac{1}{\mathrm{p}(x)\mathrm{p}(y)} \, \partial_{y^j} \partial_{x^i} (\mathrm{p}(x) K_{ij}(x, y) \mathrm{p}(y))$.

The zero-mean condition $\mathrm{P} \in \mathcal{P}_{K,0}$ ensures that all functions in the Stein RKHS integrate to zero under the target measure so that the KSD can be evaluated without any explicit integration under P. Moreover, by Proposition 2 and Theorem 1, when Q embeds into the Stein RKHS, the KSD takes on its more familiar double integral form.

Corollary 1 (KSD as a double integral)**.** *If* $\mathrm{P} \in \mathcal{P}_{K,0}$ *and* $\mathrm{Q} \in \mathcal{P}_{\mathcal{H}_{k_\mathrm{p}}}$*, then*

$$\mathrm{KSD}^2_{K,\mathrm{P}}(\mathrm{Q}) = \iint k_\mathrm{p}(x,y)\mathrm{dQ}(x)\mathrm{dQ}(y).$$

Finally, the following result proved in Appendix C.6 provides user-friendly sufficient conditions for verifying that $\mathrm{P} \in \mathcal{P}_{K,0}$, which requires verifying that $\mathcal{H}_{k_\mathrm{p}} = \mathcal{S}_\mathrm{p}(\mathcal{H}_K) \subseteq \mathrm{L}^1(\mathrm{P})$, and $\mathrm{P}(\mathcal{H}_{k_\mathrm{p}}) = 0$. Hereafter, we let $\boldsymbol{s}_\mathrm{p} \equiv \partial \log \mathrm{p}$ denote the "score" function of P whenever $\log \mathrm{p}$ is partially differentiable.

Proposition 3 (Stein embeddability conditions)**.** *Consider a target* $\mathrm{P} \in \mathcal{P}$ *with density* $\mathrm{p} > 0$ *and matrix-valued base kernel* K *for which* $p\mathcal{H}_K$ *consists of partially differentiable functions. The following claims hold true.*

(a) $\mathcal{S}_\mathrm{p}(\mathcal{H}_K) \subseteq \mathrm{L}^1(\mathrm{P}) \Leftrightarrow \mathrm{P} \in \mathcal{P}_{\mathcal{H}_{k_\mathrm{p}}}$.

(b) *If* $\mathrm{P} \in \mathcal{P}_{\sqrt{k_\mathrm{p}}}$*, then* $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_{k_\mathrm{p}}}$.

(c) *If* $\mathrm{P} \in \mathcal{P}_{\boldsymbol{s}_\mathrm{p}}$ *and all* v *in* $\mathcal{H}_K$ *are bounded with bounded partial derivatives, then* $\mathrm{P} \in \mathcal{P}_{\sqrt{k_\mathrm{p}}}$.

(d) *If* $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_{k_\mathrm{p}}}$*, then* $\iint k_\mathrm{p}(x,y)\mathrm{dP}(x)\mathrm{dP}(y) = 0 \Leftrightarrow \mathrm{P} \in \mathcal{P}_{K,0}$.

(e) *If* $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_{k_\mathrm{p}}}$ *and* $\mathcal{H}_K \subseteq \mathrm{L}^1(\mathrm{P}) \cap \mathscr{C}^1(\mathbb{R}^d)$*, then* $\mathrm{P} \in \mathcal{P}_{K,0}$.

Remark 5 (User-friendly conditions on $\mathcal{H}_K$)**.** *The requirements on* $\mathcal{H}_K$ *in Proposition 3 (c) and (e) can often be verified by examining simple properties of the base kernel* K*. For example, by Lemma 3, all* v *in* $\mathcal{H}_K$ *are bounded iff* $x \mapsto \|K(x,x)\|$ *is bounded, and all* v *in* $\mathcal{H}_K$ *have bounded* x^i*-partial derivatives if* $(x,y) \mapsto \left\|\partial_{x^i}\partial_{y^i}K(x,y)\right\|$ *exists and is bounded for any matrix norm* $\|\cdot\|$*. In particular, if* $K \in \mathscr{C}_b^{(1,1)}(\mathbb{R}^d)$*, then* $\mathcal{H}_K \subseteq \mathscr{C}_b^1(\mathbb{R}^d)$*. Moreover, by Micheli and Glaunes (2013, Thm. 2.11),* $\mathcal{H}_K \subseteq \mathscr{C}^1(\mathbb{R}^d)$ *iff* $(x,y) \mapsto \partial_{x^i}\partial_{y^i}K(x,y)$ *is separately continuous and locally bounded.*

3. Conditions for Separating Measures

Our first goal is to identify when an MMD distinguishes P from other measures. Given a set of probability measures $\mathcal{M} \subseteq \mathcal{P}$, we will say that k *separates* P *from* $\mathcal{M}$ if for any $\mathrm{Q} \in \mathcal{M}$, $\mathrm{MMD}_k(\mathrm{Q},\mathrm{P}) = 0$ implies that $\mathrm{Q} = \mathrm{P}$. When k separates P from all probability measures $\mathcal{P}$ we say simply that k is P*-separating*. We will first discuss restricted P-separation—that is, separation from a distinguished subset of measures $\mathcal{M} \neq \mathcal{P}$—in Sections 3.1 and 3.2 and then turn to general P-separation—separation from all probability measures $\mathcal{P}$—in Section 3.4.

3.1 Bochner P-separation with MMDs

Our first result, proved in Appendix D, exactly characterizes the kernels that separate P from Bochner embeddable measures on Radon spaces (Ambrosio et al., 2005, Def. 5.1.4).

Recall that a set of probability measures $\mathcal{M} \subseteq \mathcal{P}$ is *tight* when for each $\epsilon > 0$ there exists a compact set $S \subseteq \mathcal{X}$ such that $\mathrm{Q}(S^c) \leq \epsilon$ for all $\mathrm{Q} \in \mathcal{M}$. We also say that a measurable function $\varphi : \mathcal{X} \to \mathbb{R}$ is *uniformly integrable* by $\mathcal{M} \subseteq \mathcal{P}$ if for each $\epsilon > 0$ there exists $r > 0$ such that $\sup_{\mu \in \mathcal{M}} \int_{\{x : |\varphi|(x) > r\}} |\varphi| \, \mathrm{d}\mu < \epsilon$.

Theorem 2 (Bochner P-separation with MMDs)**.** *Let k be a continuous kernel over a Radon space $\mathcal{X}$ (for example, a Polish space). Then k separates $\mathrm{P} \in \mathcal{P}_{\sqrt{k}}$ from $\mathcal{P}_{\sqrt{k}}$ iff, for any sequence $(\mathrm{Q}_n)_n \subseteq \mathcal{P}_{\sqrt{k}}$,*

$$\mathrm{Q}_n h \to \mathrm{P} h \quad \forall h \in \mathscr{C}_{\sqrt{k}} \qquad \Longleftrightarrow \qquad \begin{cases} \text{(a)} & \mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0 \\ \text{(b)} & (\mathrm{Q}_n)_n \text{ is tight} \\ \text{(c)} & (\mathrm{Q}_n)_n \text{ uniformly integrates } \sqrt{k}. \end{cases} \tag{6}$$

Theorem 2 exposes an important relationship between our two goals of separation and convergence control. In particular, when k is bounded, the uniform integrability condition (c) always holds, $\mathcal{P}_{\sqrt{k}}$ is the set of all probability measures $\mathcal{P}$, $\mathscr{C}_{\sqrt{k}}$ is the set of all bounded continuous functions, and the convergence on the left-hand side of (6) is the usual weak convergence in $\mathcal{P}$. Hence for bounded kernels we obtain Corollary 2: separating P from all probability measures is equivalent to *controlling tight* P*-convergence*, i.e., having $\mathrm{Q}_n \to \mathrm{P}$ weakly whenever $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0$ and $(\mathrm{Q}_n)_n$ is tight.

Corollary 2 (P-separation with bounded kernels)**.** *Let k be a continuous bounded kernel over a Radon space $\mathcal{X}$ (for example, a Polish space). Then k separates $\mathrm{P} \in \mathcal{P}$ from $\mathcal{P}$ iff, for any sequence $(\mathrm{Q}_n)_n \subseteq \mathcal{P}$,*

$$\mathrm{Q}_n h \to \mathrm{P} h \quad \forall h \in \mathscr{C}_b \qquad \Longleftrightarrow \qquad \begin{cases} \text{(a)} & \mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0 \\ \text{(b)} & (\mathrm{Q}_n)_n \text{ is tight.} \end{cases}$$

Remark 6 (Comparison with Simon-Gabriel et al. (2023))**.** *When $\mathcal{X}$ is also locally compact and Hausdorff, for instance when $\mathcal{X} = \mathbb{R}^d$, Theorem 9 in Simon-Gabriel et al. (2023) implies that, if $\mathcal{H}_k \subset \mathscr{C}_0$ and k separates every finite measure μ from the set of finite measures, then k metrizes the weak convergence of probability measures (i.e., for every probability measure* P, $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0 \Leftrightarrow \mathrm{Q}_n \to \mathrm{P}$ *weakly). Comparing with Corollary 2, we observe no explicit tightness requirement appears. This is because the assumption of separation for* ***every finite measure*** *μ (instead of separation of a single finite measure* P *from $\mathcal{P}$) implicitly does the work of enforcing tightness. In the proof of Theorem 2 we can see the role of tightness is to ensure relative compactness, which in turn allows us to use the existence of convergent subsequences to promote the separation assumption into a convergence control. But $\mathcal{P}$ is a bounded and thus relatively compact subset of the space of finite measures (Treves, 1967, Thm. 33.2). Hence, by Treves (1967, Prop. 32.5), the assumption of k separating all finite measures is enough to guarantee the equivalence between $\mathrm{Q}_n h \to \mathrm{P} h$ for all $h \in \mathcal{H}_k$ and $\mathrm{Q}_n h \to \mathrm{P} h$ for all $h \in \mathscr{C}_0$. The latter is further equivalent to $\mathrm{Q}_n h \to \mathrm{P} h$ for all $h \in \mathscr{C}_b$ by Berg et al. (1984, Cor. 2.4.3).*

3.2 Score P-separation with KSDs

The standard practice in the KSD literature is to identify easily-verified properties of the base kernel K, target P, and alternative measure Q that ensure separation. One class of KSD separation conditions—introduced by Chwialkowski et al. (2016, Thm. 2.2) and generalized by Barp et al. (2019, Prop. 1)—applies to measures that finitely integrate the score $\boldsymbol{s}_{\mathrm{p}}$ but additionally requires Q to have a continuously differentiable log-density. The first main result of this work, proved in Appendix E, removes the extraneous continuity conditions and extends P-separation to *all* measures $\mathrm{Q} \in \mathcal{P}_{\boldsymbol{s}_{\mathrm{p}}}$ under a standard separating assumption on the base kernel, $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$-*characteristicness*.

Theorem 3 (Score P-separation with KSDs)**.** *Suppose a matrix-valued kernel K with $\mathcal{H}_K \subseteq \mathscr{C}^1_b(\mathbb{R}^d)$ is $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$-characteristic. If $\mathrm{P} \in \mathcal{P}_{K,0}$, then k_{p} separates P from $\mathcal{P}_{\boldsymbol{s}_{\mathrm{p}}}$.*

We provide formal definitions of $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$ and characteristicness in Definition 8 and Definition 7 respectively. In brief, $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$ is the d-dimensional product of the space $\mathscr{D}^1_{L^1}$ of finite measures and their distributional derivatives[4], and a $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$-characteristic kernel is one that can separate any pair of $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$ elements. Our proof of Theorem 3 builds on the kernel Schwartz distribution theory of Simon-Gabriel and Schölkopf (2018), wherein the space $\mathscr{D}^1_{L^1}$ naturally arises from the construction of the Stein RKHS via the Langevin Stein operator $\mathcal{S}_{\mathrm{p}}$. Specifically, we show in Appendix P that the Stein kernel k_{p} separates P from $\mathrm{Q} \in \mathcal{P}_{\mathcal{H}_{k_{\mathrm{p}}}}$ if and only if the base kernel K separates the Schwartz distribution $\boldsymbol{s}_{\mathrm{p}}\mathrm{Q} - \partial_{x^j}\mathrm{Q}$ from the zero measure. Moreover, $\boldsymbol{s}_{\mathrm{p}}\mathrm{Q} - \partial_{x^j}\mathrm{Q} \in \mathscr{D}^1_{L^1}(\mathbb{R}^d)$ when $\mathrm{Q} \in \mathcal{P}_{\boldsymbol{s}_{\mathrm{p}}}$, which yields Theorem 3.

Application 1: Goodness-of-fit Testing

In goodness-of-fit (GOF) testing, one uses a sequence of datapoints $X_1, \ldots, X_n$ generated from a Markov chain to test whether the chain's stationary distribution Q coincides with a target distribution P. KSDs with $\mathscr{D}^1_{L^1}$-characteristic translation-invariant base kernels are commonly used as GOF test statistics, and such tests are known to consistently reject $\mathrm{Q} = \mathrm{P}$ whenever $\mathrm{KSD}(\mathrm{Q}, \mathrm{P}) > 0$ (Chwialkowski et al., 2016; Liu et al., 2016; Gorham and Mackey, 2017). However, prior to this work, the separating condition $\mathrm{KSD}(\mathrm{Q}, \mathrm{P}) > 0$ had only been established for a restricted class of alternatives (continuous $\mathrm{Q} \in \mathcal{P}_{\sqrt{k_{\mathrm{p}}}}$ with differentiable log densities satisfying $\mathrm{Q}(\|\boldsymbol{s}_{\mathrm{p}} - \boldsymbol{s}_{\mathrm{q}}\|) < \infty$, Barp et al. 2019, Prop. 1) or a restricted class of targets (P with Lipschitz $\boldsymbol{s}_{\mathrm{p}}$ and strongly log concave tails, Gorham and Mackey 2017, Thm. 7). The former restriction excludes discrete and discontinuous Q, as well as Q with tails heavier than P or non-differentiable densities. Meanwhile, the latter restriction excludes P with tails heavier than or lighter than a Gaussian. Theorem 3 in the present work ensures that $\mathrm{KSD}(\mathrm{Q}, \mathrm{P}) > 0$ for *any* $\mathrm{P} \in \mathcal{P}_{K,0}$ and $\mathrm{Q} \in \mathcal{P}_{\boldsymbol{s}_{\mathrm{p}}}$. In particular, this accommodates discontinuous or non-smooth Q and all targets P for which the KSD (4) is defined. Moreover, Theorem 3 holds for all $\mathscr{D}^1_{L^1}$-characteristic kernels, a strict superset of the $\mathscr{C}^1_0$-universal kernels (Carmeli et al., 2010, Def. 4.1) assumed in prior work.

4. Distributional derivatives extend the usual notion of derivative to objects that are not smooth, in particular to non-smooth distributions Q. When Q has a differentiable Lebesgue density q then we recover the usual derivative, $\partial_{x^j}\mathrm{Q} = \partial_{x^j} q \, \mathrm{d}x$, while in general $\partial_{x^j}\mathrm{Q}$ will be a Schwartz distribution (Schwartz, 1978).

Indeed, Simon-Gabriel and Schölkopf (2018, Thm. 12, Tab. 1, and Cor. 38) showed that any $\mathscr{C}_0^1$-universal k and any $\mathscr{C}^{(1,1)}$ translation-invariant k with fully supported spectral measure is $\mathscr{D}_{L^1}^1$-characteristic. These results already cover all of the translation-invariant base kernels commonly used with KSDs including Gaussian, inverse multiquadric (IMQ), log inverse, sech, Matérn, B-spline, and Wendland's compactly supported kernels. Moreover, as we prove in Appendix F.1, characteristicness to $\mathscr{D}_{L^1}^1$ is preserved under the following operations, which allows one to construct even more flexible base kernels.

Proposition 4 (Preserving characteristicness)**.** *Suppose a matrix-valued kernel K with $\mathcal{H}_K \subseteq \mathscr{C}_b^1(\mathbb{R}^d)$ is $\mathscr{D}_{L^1}^1(\mathbb{R}^d)$-characteristic. Then the following claims hold true.*

(a) *If $a \in \mathscr{C}_b^1$ is strictly positive, then $a(x)K(x,y)a(y)$ is $\mathscr{D}_{L^1}^1(\mathbb{R}^d)$-characteristic.*

(b) *If $b : \mathbb{R}^d \to \mathbb{R}^d$ is a Lipschitz $\mathscr{C}^1(\mathbb{R}^d)$-diffeomorphism, then the composition kernel $K(b(x), b(y))$ is $\mathscr{D}_{L^1}^1(\mathbb{R}^d)$-characteristic.*

(c) *If k_j is $\mathscr{D}_{L^1}^1$-characteristic for $j \in [d]$, then $\operatorname{diag}(k_1, \dots, k_d)$ is $\mathscr{D}_{L^1}^1(\mathbb{R}^d)$-characteristic.*

As a final remark on Theorem 3, we note that the score embedding measures $\mathcal{P}_{\boldsymbol{s}_\mathrm{p}}$ and the Bochner embeddable measures $\mathcal{P}_{\sqrt{k_\mathrm{p}}}$ exactly coincide under mild conditions satisfied by every $\mathscr{C}^{(1,1)}$ translation-invariant base kernel K. See Appendix C.3 for the proof of this result.

Proposition 5 (Score vs. Bochner embeddability)**.** *Under the assumptions of Theorem 3, $\mathcal{P}_{\boldsymbol{s}_\mathrm{p}} \subseteq \mathcal{P}_{\sqrt{k_\mathrm{p}}}$. If, in addition, $x \mapsto \sqrt{\langle \boldsymbol{s}_\mathrm{p}(x), K(x,x)\boldsymbol{s}_\mathrm{p}(x)\rangle}/\|\boldsymbol{s}_\mathrm{p}(x)\|$ is bounded away from zero, then $\mathcal{P}_{\boldsymbol{s}_\mathrm{p}} = \mathcal{P}_{\sqrt{k_\mathrm{p}}}$.*

3.3 L^2 P-separation with KSDs

Liu et al. (2016) introduced a second class of KSD separation conditions based on an L^2 separating property of the base kernel. We say that a matrix-valued kernel K is $\mathrm{L}^2(\mathbb{R}^d)$-*integrally strictly positive definite (ISPD)* if $\mathcal{H}_K \subseteq \mathrm{L}^2(\mathbb{R}^d)$ and

$$g \in \mathrm{L}^2(\mathbb{R}^d) \quad \text{and} \quad g \neq 0 \quad \Rightarrow \quad \iint g(x)^T K(x,y) g(y) \mathrm{d}x\mathrm{d}y > 0.$$

Unfortunately, the L^2 requirement on $\mathcal{H}_K$ excludes certain popular base kernels like slowly decaying IMQ and log inverse kernels, and Liu et al. (2016) did not provide any examples of kernels satisfying the L^2-ISPD conditions. Our next result fills this gap by showing that many standard kernels are L^2-ISPD, including Gaussian, Matérn, sech, B-spline, faster decaying IMQ, and Wendland's compactly supported kernels, along with their tilted variants. The proof can be found in Appendix G.

Theorem 4 (L^2-ISPD conditions)**.** *The following claims hold true for a matrix-valued kernel K.*

(a) *Suppose $(k_j)_{j=1}^d$ are translation-invariant continuous kernels with $\mathcal{H}_{k_j} \subseteq \mathrm{L}^2$. If the spectral measure of each k_j is fully supported, then $K = \operatorname{diag}(k_j)$ is $\mathrm{L}^2(\mathbb{R}^d)$-ISPD.*

(b) *If K is $\mathrm{L}^2(\mathbb{R}^d)$-ISPD and $A : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is bounded measurable with $A(x)$ invertible for each x, then the tilted kernel $A(x)K(x,y)A(y)^T$ is also $\mathrm{L}^2(\mathbb{R}^d)$-ISPD.*

(c) *If* $\mathcal{H}_K$ *is separable,* $\sup_x \|K_x u\|_{\mathrm{L}^1} < \infty$*, and* $K_x u \in \mathrm{L}^2(\mathbb{R}^d)$ *for each* x *and* $u \in \mathbb{R}^d$*, then* $\mathcal{H}_K \subseteq \mathrm{L}^2(\mathbb{R}^d)$.

(d) *Suppose* $K_x u \in \mathrm{L}^1(\mathbb{R}^d)$ *for some* $u \in \mathbb{R}^d$*. If* K *is translation-invariant or, more generally, if* $K_x u$ *is bounded, then* $K_x u \in \mathrm{L}^2(\mathbb{R}^d)$.

Previous results in the literature have focused on properties similar to but distinct from the $\mathrm{L}^2(\mathbb{R}^d)$-ISPD condition. These include conditions under which kernels are (i) $\mathrm{L}^p(\mu)$-ISPD for $p \in [1, \infty)$ with respect to a probability measure μ in place of Lebesgue measure (Carmeli et al., 2010); (ii) ISPD, meaning that $\iint k(x, y)\mathrm{d}\mu(x)\mu(y) > 0$ for all non-zero finite measures μ (Sriperumbudur et al., 2011); (iii) L^1 integrally *non-strictly* positive definite (INPD), meaning $g \in \mathrm{L}^1 \Rightarrow \iint g(x)^T K(x, y)g(y)\mathrm{d}x\mathrm{d}y \geq 0$ (Bochner, 1932; Stewart, 1976); (iv) L^2_c-INPD where L^2_c is the space of compactly supported L^2 functions (Cooper, 1960); or (v) L^2-INPD for continuous $k \in \mathrm{L}^2$ when $d = 1$ (Buescu et al., 2004, Rem. 2.10) or translation-invariant k with $k_x \in \mathrm{L}^1$ (Phillips, 2018, Thm. 2.5.1).

Liu et al. (2016, Prop. 3.3 & Thm. 3.8) showed that KSDs with L^2-ISPD base kernels separate certain measures with continuously differentiable densities q. Theorem 5, proved in Appendix H, generalizes this finding to matrix-valued K and partially differentiable q and provides user-friendly L^2 conditions for ensuring that Q can be separated.

Theorem 5 (L^2 P-separation with KSDs)**.** *Suppose* $\mathrm{P} \in \mathcal{P}_{K,0}$ *for a matrix-valued kernel* K*. The following claims hold true.*

(a) *If* K *is* $\mathrm{L}^2(\mathbb{R}^d)$*-ISPD, then* k_{p} *separates* P *from* $\{\mathrm{Q} \in \mathcal{P}_{\mathcal{H}_{k_{\mathrm{p}}}} \cap \mathcal{P}_{K,0} : (\boldsymbol{s}_{\mathrm{p}} - \boldsymbol{s}_{\mathrm{q}})\mathrm{q} \in \mathrm{L}^2(\mathbb{R}^d)\}$.

(b) *If* $\mathrm{Q} \in \mathcal{P}_{\mathcal{H}_{k_{\mathrm{p}}}}$*,* $(\boldsymbol{s}_{\mathrm{p}} - \boldsymbol{s}_{\mathrm{q}})\mathrm{q} \in \mathrm{L}^2(\mathbb{R}^d)$ *and* $\mathcal{H}_K \subseteq \mathrm{L}^2(\mathbb{R}^d) \cap \mathrm{L}^\infty(\mathbb{R}^d)$*, then* $\mathrm{Q} \in \mathcal{P}_{K,0}$.

(c) *If* $\mathcal{H}_K \subseteq \mathrm{L}^2(\mathbb{R}^d)$*,* $\partial\mathcal{H}_K \subseteq \mathrm{L}^\infty(\mathbb{R}^d)$*, and* $\boldsymbol{s}_{\mathrm{p}}\mathrm{q} \in \mathrm{L}^2(\mathbb{R}^d)$*, then* $\mathrm{Q} \in \mathcal{P}_{\mathcal{H}_{k_{\mathrm{p}}}}$.

While Theorem 5 only applies to continuous Q, it does cover certain measures excluded by Theorem 3. For example, Theorem 5 implies that Cauchy alternatives Q are separated from Gaussian targets P since $\mathrm{q}\,\|\boldsymbol{s}_{\mathrm{p}}\|^2$ and $\mathrm{q}\,\|\boldsymbol{s}_{\mathrm{q}}\|^2$ are bounded and hence $\mathrm{q}\boldsymbol{s}_{\mathrm{p}}, \mathrm{q}\boldsymbol{s}_{\mathrm{q}} \in \mathrm{L}^2(\mathbb{R}^d)$. Meanwhile, Theorem 3 cannot be applied to these (Q, P) pairings as the heavy tails of a Cauchy cannot finitely integrate a Gaussian score $\boldsymbol{s}_{\mathrm{p}}$.

3.4 General P-separation

The results in the preceding sections only yield general P-separation when applied to bounded kernels, and indeed this has been the standard in much of the MMD literature (Sriperumbudur et al., 2010; Sriperumbudur, 2016; Simon-Gabriel and Schölkopf, 2018; Simon-Gabriel et al., 2023). To accommodate the unbounded Stein kernels that often arise in KSDs, our next definition and result (proved in Appendix J) provide a new, convenient means to check that *unbounded* kernels separate P from $\mathcal{P}$.

Definition 3 (Bounded P-separating property)**.** *We say a set of functions* $\mathcal{F}$ *is* bounded P-separating *if* $\mathrm{L}^\infty \cap \mathcal{F}$ *is* P*-separating, i.e., if* $\mathrm{Q} \in \mathcal{P}$ *and* $\mathrm{Q}h = \mathrm{P}h$ *for all* $h \in \mathrm{L}^\infty \cap \mathcal{F}$ *then* $\mathrm{Q} = \mathrm{P}$.

Theorem 6 (Controlling tight convergence with bounded separation)**.** *If $\mathcal{H}_k$ is bounded* P*-separating, then k is* P*-separating and controls tight* P*-convergence.*

According to Theorem 6, to establish general P-separation, it suffices to restrict focus to the bounded functions in an RKHS. Moreover, Theorem 6 suggests a convenient strategy for proving P-separation with unbounded kernels k: (i) identify a sub-RKHS of bounded functions that belongs to $\mathcal{H}_k$ and (ii) appeal to a broadly applicable bounded-kernel result to establish the P-separation of the bounded sub-RKHS.

To apply this strategy to KSDs, we first show in Appendix K that any suitably tilted $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$-characteristic base kernel yields a bounded and P-separating Stein kernel:

Theorem 7 (Controlling tight convergence with bounded Stein kernels)**.** *Suppose a matrix-valued kernel K with $\mathcal{H}_K \subseteq \mathscr{C}^1_b(\mathbb{R}^d)$ is $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$-characteristic. If $\|\boldsymbol{s}_{\mathrm{P}}(x)\| \leq \theta(x)$ for $\theta \in \mathscr{C}^1$ with $\frac{1}{\theta} \in \mathscr{C}^1_b$, then the Stein kernel induced by the tilted base kernel $\frac{K(x,y)}{\theta(x)\theta(y)}$ is bounded and* P*-separating and controls tight* P*-convergence.*

Next we show that standard translation-invariant base kernels have sub-RKHSes of precisely the form needed by Theorem 7:

Theorem 8 (Translation-invariant kernels have rapidly decreasing sub-RKHSes)**.** *Suppose a kernel k with $\mathcal{H}_k \subseteq \mathscr{C}^1$ is translation invariant with a spectral density bounded away from zero on compact sets. Then there exist a translation-invariant, $\mathscr{D}^1_{L^1}$-characteristic kernel $k_s \in \mathscr{C}^{(1,1)}$ and, for each $c > 0$, a positive-definite function f with $\frac{1}{f} \in \mathscr{C}^1$, $\max(|f(x)|, \|\partial f(x)\|) = O(e^{-c\sum_{i=1}^d \sqrt{|x^i|}})$, and*

$$\mathcal{H}_{k_f} \subseteq \mathcal{H}_{k_{fs}} \subseteq \mathcal{H}_k \quad \text{for} \quad k_f(x,y) \equiv f(x)k_s(x,y)f(y) \quad \text{and} \quad k_{fs}(x,y) \equiv k_s(x,y)f(x-y).$$

Theorem 8 applies to all of the translation-invariant base kernels commonly used with KSDs including Gaussian, IMQ, log inverse, sech, Matérn, B-spline, and Wendland's compactly supported kernels. Moreover, our proof in Appendix L explicitly constructs the $\mathscr{D}^1_{L^1}$-characteristic kernel k_s and the rapidly decreasing tilt function f and may be of independent interest.

We now apply our Stein operator to the base kernels of Theorem 8 and invoke Theorem 7 to deduce the second main result of this work: KSDs based on standard translation invariant kernels achieve general P-separation, even when their Stein kernels are unbounded. The proof of this result can be found in Appendix M.

Theorem 9 (Controlling tight convergence with KSDs)**.** *For k as in Theorem 8, define the tilted kernel $k_a(x,y) = a(x)k(x,y)a(y)$ for each strictly positive $a \in \mathscr{C}^1$.*

(a) *If $\mathrm{P} \in \mathcal{P}_{k,0}$ and $\|\boldsymbol{s}_{\mathrm{P}}\|$ has at most root exponential growth,*[5] *then the Stein kernel induced by k is bounded* P*-separating and controls tight* P*-convergence.*

(b) *Moreover, if $\mathrm{P} \in \mathcal{P}_{k_a,0}$ and a, ∂a, and $a\|\boldsymbol{s}_{\mathrm{P}}\|$ have at most root exponential growth, then the Stein kernel induced by k_a is bounded* P*-separating and controls tight* P*-convergence.*

5. A function a has *at most root exponential growth* if $a(x) = O(\exp(c\sum_{i=1}^d \sqrt{|x^i|}))$ for some $c > 0$.

Application 2: Goodness-of-fit Testing, continued

In the testing setting of Application 1, Theorem 9 extends the reach of KSD GOF testing by guaranteeing $\mathrm{KSD}(\mathrm{Q}, \mathrm{P}) > 0$ for *all* alternatives Q whenever $\|\boldsymbol{s}_{\mathrm{p}}\|$ has at most root exponential growth. Since the Stein kernels of Theorem 9 are also bounded P-separating, the same consistency guarantees immediately extend to the computationally efficient stochastic KSDs of Gorham et al. (2020, Thm. 4).

4. Conditions for Convergence Control

Having derived sufficient conditions on the RKHS to separate measures and control tight convergence, we now present both sufficient and necessary conditions to ensure that an MMD controls weak convergence to P. Hereafter, we will say that *k controls weak convergence to* P or *controls* P*-convergence* whenever $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0$ implies $\mathrm{Q}_n \to \mathrm{P}$ weakly. Moreover, we will say that *k enforces tightness* whenever $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0$ implies that $(\mathrm{Q}_n)_n$ is tight. Enforcing tightness is central to our developments as, if k controls tight weak convergence to P and enforces tightness, then it also controls weak convergence to P.

4.1 Sufficient conditions

We begin by introducing a new sufficient condition to ensure that MMDs and integral probability metrics more generally enforce tightness.

Definition 4 (P-dominating indicators)**.** *Consider a set of functions $\mathcal{F} \subseteq \mathrm{L}^1(\mathrm{P})$. We say that $\mathcal{F}$* P-dominates indicators *if, for each $\epsilon > 0$, there exists a compact set $S \subseteq \mathcal{X}$ and a function $h \in \mathcal{F}$ that satisfy*

$$h - \mathrm{P}h \geq \mathbb{I}\left[S^c\right] - \epsilon. \tag{7}$$

Definition 4 ensures that a sequence $(\mathrm{Q}_n)_n$ can only approximate P well if it places uniformly little mass outside of a compact set S. As we show in Appendix N, this is sufficient to ensure that integral probability metrics like the MMD enforce tightness.

Theorem 10 (Controlling P-convergence by dominating indicators)**.** *If $\mathcal{F} \subseteq \mathrm{L}^1(\mathrm{P})$* P*-dominates indicators then $(\mathrm{Q}_n)_n$ is tight whenever the integral probability metric*

$$d_{\mathcal{F}}(\mathrm{Q}_n, \mathrm{P}) \equiv \sup_{h \in \mathcal{F}:\ h_+ \in \mathrm{L}^1(\mathrm{Q}_n) \text{ or } h_- \in \mathrm{L}^1(\mathrm{Q}_n)} |\mathrm{Q}_n h - \mathrm{P}h| \to 0.$$

Hence, if $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_k}$ and $\mathcal{H}_k$ P*-dominates indicators then $(\mathrm{Q}_n)_n$ is tight whenever $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0$. If, in addition, k controls tight* P*-convergence, then k also controls* P*-convergence.*

We can now combine Theorem 10 with any of our KSD tight convergence results to immediately obtain P-convergence control for KSDs.

Corollary 3 (Controlling P-convergence with KSDs)**.** *Under the conditions of Theorem 3, 7, or 9, if $\mathcal{H}_{k_{\mathrm{p}}}$* P*-dominates indicators, then k_{p} controls* P*-convergence.*

Before we discuss applications of these results, let us compare them to existing results in the literature. Prior work relied on a stronger, coercive function condition to establish that KSDs enforce tightness with generalized multiquadric (Gorham and Mackey, 2017, Lem. 16), IMQ score (Chen et al. 2018, Thm. 4; Hodgkinson et al. 2020, Ex. 6), log inverse (Chen et al., 2018, Thm. 3), or unbounded tilted translation invariant (Huggins and Mackey, 2018, Thm. 3.2) base kernels. Hodgkinson et al. (2020) used the following general definition of coercivity.

Definition 5 (Coercive function (Hodgkinson et al., 2020, Assump. 1))**.** *We say a function* $h : \mathcal{X} \to \mathbb{R}$ *is* coercive *if, for any* $M > 0$*, there exists a compact set* $S \subseteq \mathcal{X}$ *such that* $\inf_{x \in S^c} h(x) > M$*.*

Remark 7 (Bounded coercive functions)**.** *Any continuous coercive function is also bounded below as continuous functions are bounded on compact sets.*

Our next result, proven in Appendix O, shows that this coercive function condition is stronger than our P-dominating indicator condition.

Lemma 1 (Coercive functions dominate indicators)**.** *If* $h \in \mathcal{H}_k$ *is coercive and bounded below and* $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_k}$*, then* $\mathcal{H}_k$ P*-dominates indicators.*

As a first application of Corollary 3, we show that KSDs with IMQ base kernels enforce tightness and control convergence whenever the dissipativity rate of the target dominates the decay rate of the kernel. Generalizing the argument in Gorham and Mackey (2017, Lem. 16), our proof in Appendix Q explicitly constructs a coercive function in the associated Stein RKHS.

Theorem 11 (IMQ KSDs control P-convergence)**.** *Consider a target measure* $\mathrm{P} \in \mathcal{P}$ *with score* $\boldsymbol{s}_{\mathrm{p}} \in \mathscr{C}(\mathbb{R}^d) \cap \mathrm{L}^1(\mathrm{P})$*. If, for some dissipativity rate* $u > 1/2$ *and* $r_0, r_1, r_2 > 0$*,* P *satisfies the* generalized dissipativity *condition*

$$-\langle \boldsymbol{s}_{\mathrm{p}}(x), x\rangle - r_0 \left\|\boldsymbol{s}_{\mathrm{p}}(x)\right\|_1 \geq r_1 \left\|x\right\|^{2u} - r_2 \quad \textit{for all} \quad x \in \mathbb{R}^d. \tag{8}$$

If $k(x,y) = (c^2 + \|x-y\|^2)^{-\gamma}$ *for* $c > 0$ *and* $\gamma \in (0, 2u-1)$*, then* $\mathcal{H}_{k_{\mathrm{p}}}$ P*-dominates indicators and enforces tightness. If, in addition,* $\|\boldsymbol{s}_{\mathrm{p}}\|$ *has at most root exponential growth, then* k_{p} *controls* P*-convergence.*

Application 3: Measuring and Improving Sample Quality

Because the KSD provides a computable quality measure that requires no explicit integration under P, KSDs are now commonly used to select and tune MCMC sampling algorithms (Gorham and Mackey, 2017), generate accurate discrete approximations to P (Liu and Wang, 2016; Chen et al., 2018, 2019; Futami et al., 2019), compress Markov chain output (Riabiz et al., 2022), and correct for biased or off-target sampling (Liu and Lee, 2017; Hodgkinson et al., 2020; Riabiz et al., 2022). Each of these applications relies on KSD convergence control, but past work only established convergence control for P with Lipschitz $\boldsymbol{s}_{\mathrm{p}}$ and strongly log concave tails (Gorham and Mackey 2017, Lem. 16; Chen et al. 2018, Thm. 3; Huggins and Mackey 2018, Thm. 3.2). Notably, these conditions imply generalized dissipativity (8) with $u = 1$ but exclude all P with tails lighter than a

Gaussian. Corollary 3 and Theorem 11 significantly relax these requirements by providing convergence control for all dissipative P with lighter-than-Laplace tails.

Much of the difficulty in analyzing KSDs stems from the fact that all known convergence-controlling KSDs are based on unbounded Stein kernels k_{P}. As a second illustration of the power of Corollary 3, Theorem 12 develops the first KSDs known to *metrize* P-*convergence* (i.e., $\mathrm{KSD}(\mathrm{Q}_n, \mathrm{P}) \to 0 \Leftrightarrow \mathrm{Q}_n \to \mathrm{P}$ weakly), by constructing **bounded** convergence-controlling Stein kernels. The following theorem is proved in Appendix R.

Theorem 12 (Metrizing P-convergence with bounded Stein kernels)**.** *Consider a target measure* $\mathrm{P} \in \mathcal{P}$ *with score* $\boldsymbol{s}_{\mathrm{P}}$ *that, for some dissipativity rate* $u > 1/2$ *and* $r, r_1, r_2 > 0$*, satisfies the generalized dissipativity condition* (8)*. Define the Stein kernel with base kernel* $K(x, y) = \mathrm{diag}\left(a(\|x\|)(x^i y^i + k(x, y))a(\|y\|)\right)$*, i.e.,*

$$k_{\mathrm{P}}(x, y) = \sum_{1 \le i \le d} \frac{\partial_{x^i} \partial_{y^i} (p(x)a(\|x\|)(x^i y^i + k(x, y))a(\|y\|)p(y))}{p(x)p(y)},$$

for k *characteristic to* $\mathscr{D}^1_{L^1}$ *with* $\mathcal{H}_k \subseteq \mathscr{C}^1_0$ *and* $a(\|x\|) \equiv (c^2 + \|x\|^2)^{-\gamma}$ *a tilting function with* $c > 0$ *and* $\gamma \le u$*. The following statements hold true:*

(a) *If* $\mathrm{P} \in \mathcal{P}_{K,0}$*, then* $\mathcal{H}_{k_{\mathrm{P}}}$ P*-dominates indicators and enforces tightness.*

(b) *If* $\mathrm{P} \in \mathcal{P}_{K,0}$*,* $\gamma \ge 0$*, and* $\|\boldsymbol{s}_{\mathrm{P}}(x)\| \le (c^2 + \|x\|^2)^{\gamma}$*, then* k_{P} *is bounded* P*-separating and controls* P*-convergence.*

(c) *If* $\|\boldsymbol{s}_{\mathrm{P}}(x)\| \cdot \|x\| \le (c^2 + \|x\|^2)^{\gamma}$ *and* $\boldsymbol{s}_{\mathrm{P}} \in \mathscr{C}$*, then* $\mathcal{H}_{k_{\mathrm{P}}} \subseteq \mathscr{C}_b$ *and* k_{P} *metrizes* P*-convergence.*

Application 4: Sampling with Stein Variational Gradient Descent

Stein variational gradient descent (SVGD) is a popular technique for approximating a target distribution P with a collection of n representative particles. The algorithm proceeds by iteratively updating the locations of the particles according to a simple rule determined by a user-selected KSD. Liu (2017) showed that the SVGD approximation converges weakly to P as the number of particles and iterations tend to infinity, provided that the chosen KSD controls P-convergence **and** that the Stein kernel is bounded. However, prior to this work, no bounded convergence-controlling Stein kernels were known. Theorem 12 therefore provides the first instance of a Stein kernel satisfying the SVGD convergence assumptions of Liu (2017).

4.2 Necessary conditions

We finally conclude with a necessary condition for an MMD to control weak convergence to P, which recovers and broadens the KSD failure derived by Gorham and Mackey (2017, Thm. 6). For each RKHS $\mathcal{H}_k \subseteq \mathrm{L}^1(\mathrm{P})$, define the P-centered RKHS $\mathcal{H}_{k^{\mathrm{P}}} \equiv \{h - \mathrm{P}h : h \in \mathcal{H}_k\}$ with P-centered kernel

$$k^{\mathrm{P}}(x, y) \equiv k(x, y) - \int k(x, y)\mathrm{dP}(y) - \int k(x, y)\mathrm{dP}(x) + \iint k(x, y)\mathrm{dP}(x)\mathrm{dP}(y).$$

Theorem 13 shows that k **fails** to control P-convergence whenever its P-centered RKHS functions all vanish at infinity; notably, this occurs whenever k^{P} is bounded with $k_x^{\mathrm{P}} \in \mathscr{C}_0$ for each x (Simon-Gabriel and Schölkopf, 2018, Prop. 3). The proof in Appendix S relies on the fact that k and k^{P} induce exactly the same MMD.

Theorem 13 (Decaying P-centered kernels fail to control P-convergence)**.** *Suppose that $\mathcal{X}$ is locally compact but not compact. If $\mathcal{H}_{k^{\mathrm{P}}} \subseteq \mathscr{C}_0$, then k does not control* P*-convergence.*

Implication 1: Standard KSDs fail for heavy-tailed P!

Since Stein kernels are already P-centered by design (i.e., $(k_{\mathrm{p}})^{\mathrm{P}} = k_{\mathrm{p}}$), Theorem 13 holds dire consequences for standard KSDs with heavy-tailed targets P. As noted by Gorham and Mackey (2017, Thm. 10), if the score function is bounded (as is common for super-Laplace distributions), then the KSD fails to control P-convergence whenever a $\mathscr{C}_0^1$ base kernel is used. Moreover, our more general Theorem 13 result implies that if the score function is decaying (as is true for any Student's t distribution), then the KSD fails to control P-convergence for any bounded base kernel. This result suggests that the standard KSD practice of using a $\mathscr{C}_0^1$ base kernel is unsuitable for heavy-tailed targets and that one should instead choose a base kernel with growth sufficient to counteract the decay of $\boldsymbol{s}_{\mathrm{p}}$.

5. Discussion

This article derived new sufficient and necessary conditions for kernel discrepancies to enforce P-separation and control P-convergence. We characterized all MMDs that separate P from Bochner embeddable measures, proposed novel sufficient conditions for separating all measures and enforcing tightness, strengthened all prior guarantees for KSD separation and convergence control on $\mathbb{R}^d$, and derived the first KSD known to exactly metrize (as opposed to strictly dominating) weak P-convergence on $\mathbb{R}^d$.

These developments point to several opportunities for further advances. First, while we have focused on weak convergence in this article, we believe many of the tools and constructions can be adapted to study the control of other modes of convergence. Natural candidates include α-*Wasserstein convergence* (Ambrosio et al., 2005), i.e., weak convergence plus the convergence of α moments, and $\mathscr{C}_{\sqrt{k}}$ *convergence*, i.e., expectation convergence for all continuous test functions bounded by $\sqrt{k}$. When $\sqrt{k}$ is unbounded, Theorem 2 exposes an important relationship between separation and $\mathscr{C}_{\sqrt{k}}$ convergence: P-separating Bochner embeddable measures is equivalent to controlling $\mathscr{C}_{\sqrt{k}}$ convergence to P for sequences that uniformly integrate $\sqrt{k}$. Hence, to control $\mathscr{C}_{\sqrt{k}}$ convergence, it remains to identify those kernels that simultaneously separate and enforce uniform integrability.

Second, while we have focused on canonical KSDs defined by the Langevin Stein operator and a bounded base kernel, our tools are amenable to analyzing other kernel-based Stein discrepancies like the diffusion KSDs of Gorham et al. (2019); Barp et al. (2019), the second-order KSDs studied in Barp et al. (2022b); Liu and Zhu (2018); Hodgkinson et al. (2020); Barp et al. (2022a), the gradient-free KSDs of Han and Liu (2018); Fisher et al. (2022), and the random feature Stein discrepancies of Huggins and Mackey (2018). In fact, employing

a diffusion KSD with an unbounded diffusion coefficients is one promising way to overcome the heavy-tailed-target failure mode highlighted in Implication 1.

Finally, while we have focused on KSDs for measures defined on $\mathbb{R}^d$, the very recent work of Wynne et al. (2022) provides a template for studying measure separation on infinite-dimensional Hilbert spaces.

Acknowledgments

We thank Bharath Sriperumbudur for suggesting the extent of the correspondence between L^2 integrally strictly positive definiteness and characteristicness for translation-invariant kernels and Heishiro Kanagawa for identifying several typos in an earlier version of this manuscript. AB and MG were 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 under the EPSRC grant [EP/R018413/2]. AB gratefully acknowledges support from the Turing-Roche strategic partnership.

Appendix A. Appendix Notation

Throughout we denote by $(e_\ell)_\ell$ the canonical basis of $\mathbb{R}^d$ and by $(e^\ell)_\ell$ its dual basis.

The spaces $\mathscr{C}_c^\ell(\mathbb{R}^d)$ and $\mathscr{C}_0^\ell(\mathbb{R}^d)$ will be equipped with their canonical topologies. However on $\mathscr{C}_b^\ell(\mathbb{R}^d)$ we will use the strict topology, written $\mathscr{C}_b^\ell(\mathbb{R}^d)_\beta$, because, for $\ell = 0$, its dual is the space of finite (Radon) measures (Conway, 1965) whenever $\mathcal{X}$ is a locally compact Hausdorff space (e.g., when $\mathcal{X} = \mathbb{R}^d$). Note in general, any topology between the weak paired topology and the Mackey topology yields the space of finite measures as its (continuous) dual (Buck, 1958, Sec. 4).

In fact we will often use a generalization of $\mathscr{C}_b^\ell(\mathbb{R}^d)$: given a continuous function $\theta : \mathbb{R}^d \to [c, \infty)$ for some $c > 0$, we will need to construct a generalisation of the space $\mathscr{C}_b^1(\mathbb{R}^d)_\beta$, denoted $\mathscr{C}_{b,\theta}^1(\mathbb{R}^d)_\beta$, and defined as the vector space of $\mathscr{C}^1(\mathbb{R}^d)$ functions for which $\theta f \in \mathscr{C}_b(\mathbb{R}^d)$, and $\partial f \in \mathscr{C}_b(\mathbb{R}^{d\times d})$, with the topology defined by the family of seminorms

$$\|f\| \equiv \sup_x \|\gamma(x)\theta(x)f(x)\|, \qquad \|f\| \equiv \sup_x \|\gamma(x)\partial_x^p f\|$$

where $\gamma \in \mathscr{C}_0$ and $|p| = 1$. In other words $f_\alpha \to f$ in $\mathscr{C}_{b,\theta}^1(\mathbb{R}^d)_\beta$ iff $(\theta f_\alpha, \partial f_\alpha) \to (\theta f, \partial f)$ in $\mathscr{C}_b(\mathbb{R}^d)_\beta \times \mathscr{C}_b(\mathbb{R}^d)_\beta$. We mention that in Lemma 10 we will similarly construct $\mathscr{B}_\theta^1(\mathbb{R}^d)$, a Banach space that plays a similar role to $\mathscr{C}_{b,\theta}^1(\mathbb{R}^d)$ but is simpler to work with (however it is not general enough for our purposes).

Given a topological vector space (TVS) $\mathcal{F}$, its (continuous) dual will be denoted $\mathcal{F}^*$. Given a subset $\mathcal{M} \subseteq \mathcal{F}^*$, and $D_\alpha, D \in \mathcal{M}$ we will write $D_\alpha \overset{\mathcal{M}}{\to} D$ when $D_\alpha(f) \to D(f), \forall f \in \mathcal{F}$ (i.e., D_α converges to D in weak star topology). When $\mathcal{F} = \mathscr{C}_b$, and $\mathcal{M} = \mathcal{P}$, we say that D_α *converges weakly to* D. More generally, we define *weak convergence in* $\mathcal{P}_{\sqrt{k}}$ (notice the "in $\mathcal{P}_{\sqrt{k}}$" part!) using $\mathscr{C}_{\sqrt{k}}$, where $\mathscr{C}_{\sqrt{k}}$ (resp. $\mathscr{C}_{0,\sqrt{k}}$) is the space of continuous functions f with $1 + \sqrt{k}$ growth, i.e., such that $f/(1+\sqrt{k})$ is bounded, (resp. in $\mathscr{C}_0$). Thus

$$\mathrm{Q}_n \overset{\mathcal{P}_{\sqrt{k}}}{\to} \mathrm{P} \quad \Leftrightarrow \quad \mathrm{Q}_n, \mathrm{P} \in \mathcal{P}_{\sqrt{k}}\,, \quad \text{and} \quad \mathrm{Q}_n(f) \to \mathrm{P}(f) \quad \forall f \in \mathscr{C}_{\sqrt{k}}.$$

Notice that $\mathscr{C}_b \subseteq \mathscr{C}_{\sqrt{k}}$ and $\mathcal{P}_{\sqrt{k}} \subseteq \mathcal{P}$ with equality *if and only if (iff)* k is bounded. Recall here that, for a $\mathbb{R}^\ell$-valued function f such as $\sqrt{k}$, $\mathcal{P}_f \equiv \{\mathrm{Q} \in \mathcal{P} : \|f\| \in \mathrm{L}^1(\mathrm{Q})\}$.

Given TVSs $\mathcal{F}_1$ and $\mathcal{F}_2$, we denote by $\mathcal{B}(\mathcal{F}_1, \mathcal{F}_2)$ the set of continuous linear functionals from $\mathcal{F}_1$ to $\mathcal{F}_2$. The transpose of a continuous linear functional T is denoted T^*.

Given a Radon measure μ on $\mathbb{R}^d$, its distributional x^i-derivative will be denoted $\partial_{x^i}\mu : \mathscr{C}_c^\infty \to \mathbb{R}$. Recall that the distributional derivative is equal to $\partial_{x^i}\mu = -\mu \circ \partial_{x^i}$ on $\mathscr{C}_c^\infty$.

Appendix B. Vector-Valued RKHSes and Stein RKHSes

Let $\mathcal{X}$ an open subset of $\mathbb{R}^d$. Let $\Gamma(Y)$ denotes the set of maps $\mathcal{X} \to Y$. Matrix-valued kernels are typically defined via a feature map, i.e., a map $\xi^* : \mathcal{X} \to \mathcal{B}(\mathcal{H}, \mathbb{R}^d)$ (see Definition 6), which generates the kernel

$$K(x, y) \equiv \xi^*(x) \circ \xi(y).$$

In particular if $\mathcal{H} \subseteq \Gamma(\mathbb{R}^d)$ is a RKHS of $\mathbb{R}^d$-valued functions, i.e., a Hilbert space on which the evaluation functionals $\delta_x \in \mathcal{B}(\mathcal{H}_K, \mathbb{R}^d)$ are continuous, then $\mathcal{H} \equiv \mathcal{H}_K$ where $K(x, y) \equiv \delta_x \circ \delta_y^* \in \mathbb{R}^{d\times d}$. The transpose of δ_y is usually denoted $K_y \equiv \delta_y^*$, and $K_y^v \equiv \delta_y^*(v) \in \mathcal{H}_K$, so

$K_x v(y) = \delta_y K_x v = \delta_y \delta_x^* v = K(x,y)^* v = K(y,x)v$ for any $v \in \mathbb{R}^d$, thus $K_x = K(\cdot, x)$. We can tilt matrix-valued kernels via a matrix-valued function $m \in \Gamma(\mathbb{R}^{d\times d})$, indeed $\xi_m^* \equiv m \circ \xi^*$ is a new feature map, and its kernel is

$$K_m(x,y) \equiv \xi_m^*(x) \circ \xi_m(y) = m(x) \circ \xi^*(x) \circ \xi(y) \circ m^T(y) = m(x) K(x,y) m(y)^T.$$

Given an RKHS $\mathcal{H}_K$ of continuously differentiable $\mathbb{R}^d$-valued functions, we can obtain a scalar-valued kernel via the Stein operator $\mathcal{S}_{\mathrm{p}}$.[6] Let $\tilde{\xi}_{\mathrm{P}}^m \equiv \mathcal{S}_{\mathrm{p}} \circ m \circ \tilde{\xi} : \mathcal{H}_K \to \Gamma(\mathbb{R})$, where $\tilde{\xi}(h)(x) \equiv \xi^*(x)(h)$. Then $\xi_{\mathrm{P}}^m : \mathcal{X} \to \mathcal{H}_K$ is a feature map for the Stein kernel k_{p} (Barp et al., 2019), i.e.,

$$k_{\mathrm{p}}(x,y) = \langle \xi_{\mathrm{P}}^m(x), \xi_{\mathrm{P}}^m(y) \rangle_K.$$

Since the matrix m just corresponds to a change of matrix kernel $K \mapsto K_m$, we can restrict to the identity case $\xi_P \equiv \xi_{\mathrm{P}}^{\mathrm{Id}}$. In other words, for the family of "diffusion" Stein operators (Gorham et al., 2019)

$$\mathcal{S}_{\mathrm{p}}^m(v) \equiv \frac{1}{\mathrm{p}} \nabla \cdot (\mathrm{p} m v) \ ,$$

the matrix-valued function m can be thought of as a transformation of the base RKHS $\mathcal{H}_K$ into $\mathcal{H}_{mKm^T}$, i.e.,

$$\mathcal{S}_{\mathrm{p}}^m(\mathcal{H}_K) = \mathcal{S}_{\mathrm{p}}^{\mathrm{Id}}(m\mathcal{H}_K) = \mathcal{S}_{\mathrm{p}}^{\mathrm{Id}}(\mathcal{H}_{mKm^T}).$$

Since K is arbitrary, without loss of generality we may choose $m = \mathrm{Id}$, $\mathcal{S}_{\mathrm{p}} \equiv \mathcal{S}_{\mathrm{p}}^{\mathrm{Id}}$. Note that the matrix functions m obtained by the generator of P-preserving diffusions can be characterized on any manifold (Barp et al., 2021).

Similarly, the Stein kernel obtained via the second-order Stein operator (Barp et al., 2022b) can be recovered by setting K to be the diagonal matrix kernel of partial derivatives of a scalar kernel. We finally recall the equivalence between universality, characteristicness, and strict positive definiteness of (scalar-valued) kernels (Simon-Gabriel and Schölkopf, 2018, Thm. 6), noting it carries on to the case of matrix-valued kernels.

Appendix C. Embedding Schwartz Distributions in an RKHS

Given a continuous linear map T between TVS, we denote by T^* its transpose, and, similarly, if h belongs to a Hilbert space, we will denote by h^* the associated element in the dual space, i.e., $h^*(f) \equiv \langle h \, , \, f \rangle$ for any f in that Hilbert space.

Definition 6 (Kernel embeddings and Pettis integrals)**.** *Let D be a linear functional on a vector space $\mathcal{F}$ containing the RKHS $\mathcal{H}_K$ of a matrix-valued kernel K.*

(a) *We say that D embeds into $\mathcal{H}_K$ if $D|_{\mathcal{H}_K}$ is continuous, i.e., if there exists a function $\Phi_K(D) \in \mathcal{H}_K$ such that for all $h \in \mathcal{H}_K$: $D(h) = \langle \Phi_K(D), h \rangle_K$. We call Φ_K the kernel embedding and $\Phi_K(D)$ the (kernel or RKHS) embedding of D. It is given by*

$$\Phi_K(D)(x) = \textstyle\sum_i e_i D(K_x^{e_i}).$$

6. Note $\mathcal{S}_{\mathrm{p}}$ is a special instance of the canonical operator associated to "measures" equivalent to the Lebesgue one with differentiable densities (or more precisely, the canonical operator induced by positive 1-densities) Barp et al. (2022a).

(b) *Given a* feature map*, i.e., a function $\xi : \mathcal{X} \to \mathcal{B}(\mathbb{R}^d, \mathcal{H}_K)$, we denote by $\xi^* : \mathcal{X} \to \mathcal{B}(\mathcal{H}_K, \mathbb{R}^d)$ the map $x \mapsto \xi(x)^*$ and define the* feature operator *$\tilde{\xi} : \mathcal{H}_K \to (\mathcal{X} \to \mathbb{R}^d)$ as $\tilde{\xi}(h)(\cdot) \equiv \xi^*(\cdot)(h)$. We say ξ is* Pettis-integrable *with respect to D if $\tilde{\xi}(\mathcal{H}_K) \subseteq \mathcal{F}$ and the linear functional $D \circ \tilde{\xi}$ embeds into $\mathcal{H}_K$. The RKHS embedding, $\Phi_K(D \circ \tilde{\xi})$, of $D \circ \tilde{\xi}$ is known as the Pettis-integral of ξ with respect to D. We will also call the map from $D \mapsto \Phi_K(D \circ \tilde{\xi})$ the RKHS embedding of ξ.*

When $\mathcal{M}$ is a set of embeddable linear functionals, for any $D, \tilde{D} \in \mathcal{M}$ we can define

$$\mathrm{MMD}_K(D, \tilde{D}) \equiv \|D - \tilde{D}\|_K \equiv \|\Phi_K(D) - \Phi_K(\tilde{D})\|_K,$$

where $\Phi_K : \mathcal{M} \to \mathcal{H}_K$ is the kernel embedding, which recovers (2) when Q and P are embeddable probability measures. In that case, k separates P from $\mathcal{M}$ iff $\Phi_k(\cdot - \mathrm{P})|_{\mathcal{M}}$ vanishes only at P.

Hereafter, we will say that a kernel is *characteristic* to a set of embeddable linear functionals $\mathcal{M}$ when the RKHS embeddings of two distinct elements in $\mathcal{M}$ are always distinct.

Definition 7 (Characteristicness)**.** *Given a set $\mathcal{M}$ of embeddable linear functionals (see Definition 6), we say K is* characteristic to $\mathcal{M}$ *when Φ_K is injective over $\mathcal{M}$.*

When μ is a finite ($\mathbb{R}$-valued) measure on $\mathcal{X}$, then a natural set of functions that μ can act on is the set of finitely μ-integrable functions $\mathrm{L}^1(|\mu|)$. Now, if a function $\xi : \mathcal{X} \to \mathcal{H}_k$ is to be Pettis-integrable by μ, then the very least is that the functions $\tilde{\xi}(h)$ be contained in $\mathrm{L}^1(|\mu|)$ for every $h \in \mathcal{H}_k$. Interestingly, we will now see that, because $\mathcal{H}_k$ is a Hilbert space (not just Banach), this condition is also sufficient to guarantee μ-Pettis integrability.

Proposition 6 (Finite measures embed into $\mathcal{H}_k$ iff $\mathcal{H}_k$ is finitely integrable)**.** *Let μ be a finite $\mathbb{R}$-valued measure (e.g., a probability measure), seen as a linear functional over $\mathrm{L}^1(|\mu|)$. Then a function $\xi : \mathcal{X} \to \mathcal{H}_k$ is μ-Pettis integrable if and only if $\tilde{\xi}(\mathcal{H}_k) \subseteq \mathrm{L}^1(|\mu|)$. In particular, if $\mu = \mathrm{Q} \in \mathcal{P}$, then the following claims hold.*

1. *Using $\xi : x \to k_x$, it follows that Q is embeddable into $\mathcal{H}_k$ iff $\mathcal{H}_k \subseteq \mathrm{L}^1(\mathrm{Q})$.*
2. *If $\partial_{x^i}\mathcal{H}_k$ exists, then via $\xi : x \to \partial_{x^i} k_x$ we obtain that $\partial_{x^i}\mathrm{Q}$ embeds iff $\partial_{x^i}\mathcal{H}_k \subseteq \mathrm{L}^1(\mathrm{Q})$.*

Proof Since Hilbert spaces are canonically isomorphic to their dual (i.e., $\mathcal{H}^* = \mathcal{H}$), Gelfand-integration and Pettis-integration coincide. Therefore, Proposition 3.4 in Musiał (2002) – which asserts that every scalarly μ-integrable function $\xi : \mathcal{X} \to \mathcal{H}_k \cong \mathcal{H}_k^*$ is Gelfand μ-integrable – concludes the first part.

Then (1) follows directly from $\langle k_x \,,\, h\rangle_k = h(x)$. For (2), note that if $\partial_{x^i}\mathcal{H}_k$ exists, then $\xi : x \mapsto \partial_{x^i} k_x \in \mathcal{H}_k$ and $\langle h, \partial_{x^i} k_x\rangle_k = \partial_{x^i} h(x)$ by Lemma 4. Thus $\tilde{\xi} = \partial_{x^i}$ so $\partial_{x^i}\mathcal{H}_k \subseteq \mathrm{L}^1(\mathrm{Q})$ iff it is Gelfand Q-integrable, in which case

$$-\partial_i \mathrm{Q} h = \int \partial_i h \mathrm{dQ} = \int \langle h, \partial_{x^i} k_x\rangle_k \mathrm{dQ}(x) = \langle h, \int \xi \mathrm{dQ}\rangle_k$$

where $\int \xi \mathrm{dQ}$ is the Pettis integral. Hence $\partial_i \mathrm{Q}$ embeds into $\mathcal{H}_k$. ■

The embeddability of distribution in a Stein RKHS can be analysed in terms of the embeddability of the associated (via pull-back) Schwartz distributions in the base RKHS, as Lemma 2 shows, by generalizing (Simon-Gabriel and Schölkopf, 2018, Prop. 14).

Lemma 2 (Embedding functionals on RKHS defined by feature maps)**.** *Let $\mathcal{H}$ be a Hilbert space, $\mathcal{H}_K$ an RKHS of $\mathbb{R}^\ell$-valued functions on $\mathcal{X}$, and $\xi : \mathcal{X} \to \mathscr{B}(\mathbb{R}^\ell, \mathcal{H})$ be a feature map for K, i.e., $K(x,y) = \xi^*(x) \circ \xi(y)$. Then a linear functional $D : \mathcal{H}_K \to \mathbb{R}$ embeds into $\mathcal{H}_K$ iff $D \circ \tilde{\xi} : \mathcal{H} \to \mathbb{R}$ embeds into $\mathcal{H}$. Here $\tilde{\xi} : \mathcal{H} \to \mathcal{H}_K$ is the feature operator (see Definition 6).*

For any $\mathrm{Q} \in \mathcal{P}$,

$$\mathrm{Q}(\sqrt{k}) = \mathrm{Q}(\|\xi\|_{\mathcal{H}})$$

so Q is Bochner integrable in $\mathcal{H}_k$ iff $\|\xi\|_{\mathcal{H}} \in \mathrm{L}^1(\mathrm{Q})$.

Moreover, if D embeds into $\mathcal{H}_K$ then the transpose of $\tilde{\xi}$ is an isometry:

$$\|D\|_{\mathcal{H}_K} = \|D \circ \tilde{\xi}\|_{\mathcal{H}}.$$

In particular, if $D = \mathrm{Q} \in \mathcal{P}$ and $\mathcal{H}_k \subseteq \mathrm{L}^1(\mathrm{Q})$, then

$$\|\mathrm{Q}\|_{\mathcal{H}_k} = \| \int \xi \, \mathrm{dQ} \|_{\mathcal{H}}.$$

See Appendix C.1 for the proof. Applying this result to a Stein RKHS, we immediately obtain the following corollary.

Corollary 4 (Embedding measure in Stein RKHS and base RKHS)**.** *Consider a Stein kernel k_{p} (5) with base kernel K, and fix any $\mathrm{Q} \in \mathcal{P}$. The following are equivalent:*

(a) Q *embeds into $\mathcal{H}_{k_{\mathrm{p}}}$.*

(b) Q *embeds into $\mathcal{H}_K$ via the feature map $\xi_{\mathrm{P}} : \mathcal{X} \to \mathcal{H}_K$ with $\xi_{\mathrm{P}}(x) = K_x \boldsymbol{s}_{\mathrm{p}}(x) + \nabla_x \cdot K_x$.*

If either holds and $\mathrm{P} \in \mathcal{P}_{K,0}$, then

$$\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = \| \int \xi_{\mathrm{P}} \, \mathrm{dQ} \, \|_{\mathcal{H}_K},$$

where $\int \xi_{\mathrm{P}} \, \mathrm{dQ}$ is the Pettis integral.

We now formally introduce $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$, the d-dimensional product space of finite measures and their distributional derivatives.

Definition 8 (The space $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$)**.** *We write $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$ to represent the vector space of continuous linear functionals on $\mathscr{C}^1_0(\mathbb{R}^d)$ or, equivalently, on $\mathscr{C}^1_b(\mathbb{R}^d)_\beta$ and define $\mathscr{D}^1_{L^1} \equiv \mathscr{D}^1_{L^1}(\mathbb{R}^1)$. Notably, $D \in \mathscr{D}^1_{L^1}(\mathbb{R}^d)$ iff it can be expressed as a finite sum $D = \sum_{j=1}^l D_j e^j$ where each D_j is a finite Radon measure on $\mathbb{R}^d$ or a distributional derivative thereof. The topology on $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$ is the canonical dual topology induced by $\mathscr{C}^1_0(\mathbb{R}^d)$ (Schwartz, 1978, pg. 200).*

Following Definition 7, when the elements of $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$ embed into $\mathcal{H}_K$, for instance when $\mathcal{H}_K \subseteq \mathscr{C}^1_b(\mathbb{R}^d)$, we shall say that K is characteristic to $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$ when the kernel embedding $\Phi_K : \mathscr{D}^1_{L^1}(\mathbb{R}^d) \to \mathcal{H}_K$ is injective.

Importantly, for embeddable probability measures Q, the KSD is given by the norm of a vector D_{Q} that can be understood as a distributional derivative of Q with respect to a differential operator induced by P. When Q is smooth D_{Q}, will be a vector measure,

but when Q is not assumed to be smooth, D_{Q} will be a more general (vector) Schwartz distribution. The space $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$ assumes a central role in analysing D_{Q} and determining when kernel discrepancies separate D_{Q} from zero. This in turn helps us understand when KSDs effectively distinguish the target P from alternatives Q.

To define D_{Q}, note that since the feature operator of ξ_{P} in Corollary 4 is the Stein operator $\mathcal{S}_{\mathrm{p}}$, setting

$$D_{\mathrm{Q}}|_{\mathcal{H}_K} \equiv \mathrm{Q} \circ \mathcal{S}_{\mathrm{p}} : \mathcal{H}_K \to \mathbb{R}$$

we obtain that the KSD is given by evaluating the norm of D_{Q} in the base RKHS,

$$\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = \|D_{\mathrm{Q}}|_{\mathcal{H}_K}\|_{\mathcal{H}_K}.$$

More generally, D_{Q} can act on any function $f \in \mathscr{C}^1$ such that $\mathcal{S}_{\mathrm{p}}(f) \in \mathrm{L}^1(\mathrm{Q})$, and we will omit $|_{\mathcal{H}_K}$ when we do not specify its domain of definition. In addition, observe that when $\|\boldsymbol{s}_{\mathrm{p}}\|$ is integrable with respect to the probability measure Q, then both $s^i_{\mathrm{P}}\mathrm{Q}$ and Q are finite measures. Consequently, using the distributional derivative, we can write

$$D_{\mathrm{Q}} = \textstyle\sum_i (\boldsymbol{s}^i_{\mathrm{p}}\mathrm{Q} - \partial_{x^i}\mathrm{Q})e^i \equiv \sum_i D_i e^i$$

with $D_i \in \mathscr{D}^1_{L^1}$, the space of finite measures and their distributional derivatives. Hence, D_{Q} is a (vector) Schwartz distribution that belongs to the space $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$.

When Q also has a strictly positive differentiable density with respect to the Lebesgue measure, then D_{Q} simplifies to a vector measure absolutely continuous with respect to the Lebesgue measure,

$$D_{\mathrm{Q}} = \textstyle\sum_i (\boldsymbol{s}^i_{\mathrm{p}} - \boldsymbol{s}^i_{\mathrm{q}})\mathrm{Q}\, e^i.$$

The following lemma provides bounds on $\mathrm{Q}(\sqrt{k_{\mathrm{p}}})$ in terms of the base kernel K and the target score $\boldsymbol{s}_{\mathrm{p}}$, and thus sufficient conditions for a probability measure to be able to Bochner integrate k_{p}. See Appendix C.2 for the proof.

Proposition 7 (Bochner embeddability vs. score integrability)**.** *Consider a Stein kernel k_{p} (5) with base kernel K, and fix any $\mathrm{Q} \in \mathcal{P}$. We have*

$$\mathrm{Q}(\sqrt{k_{\mathrm{p}}}) \leq \textstyle\int (\|K_x\|_{\mathrm{op}}\|\boldsymbol{s}_{\mathrm{p}}(x)\|_{\mathbb{R}^d} + \|\nabla_x \cdot K_x\|_K)\,\mathrm{Q}(\mathrm{d}x), \quad \textit{and}$$
$$\mathrm{Q}(\sqrt{k_{\mathrm{p}}}) \geq \textstyle\int \big|\sqrt{\langle \boldsymbol{s}_{\mathrm{p}}(x), K(x,x)\boldsymbol{s}_{\mathrm{p}}(x)\rangle_{\mathbb{R}^d}} - \|\nabla_x \cdot K_x\|_K\big|\,\mathrm{Q}(\mathrm{d}x).$$

Now suppose $\mathcal{H}_K \subseteq \mathscr{C}^1_b(\mathbb{R}^d)$. Then the following claims hold.

- *The maps $x \mapsto \|K_x\|_{\mathrm{op}} = \sqrt{\|K(x,x)\|}$ and $x \mapsto \|\nabla_x \cdot K_x\|_K$ are bounded.*
- *If $\mathrm{Q}(\|\boldsymbol{s}_{\mathrm{p}}\|) < \infty$, then k_{p} is Bochner integrable by Q, i.e., $\mathrm{Q}(\sqrt{k_{\mathrm{p}}}) < \infty$.*
- *If $x \mapsto K(x,x)$ is uniformly positive definite (i.e., $\exists c > 0$ such that for all $v \neq 0 \in \mathbb{R}^d$, $v^T K(x,x)v \geq c\|v\|^2_{\mathbb{R}^d} > 0$ for all x), then $\mathrm{Q}(\sqrt{k_{\mathrm{p}}}) < \infty$ implies $\mathrm{Q}(\|\boldsymbol{s}_{\mathrm{p}}\|) < \infty$.*

Hence, if K is a diagonal kernel with each component satisfying $\inf_x k^i(x,x) > 0$, then $\mathrm{Q}(\sqrt{k_{\mathrm{p}}}) < \infty$ iff $\mathrm{Q}(\|\boldsymbol{s}_{\mathrm{p}}\|) < \infty$, i.e., $\mathcal{P}_{\sqrt{k_{\mathrm{p}}}} = \mathcal{P}_{\boldsymbol{s}_{\mathrm{p}}}$. In particular, if $K = k\mathrm{Id}$ where k is translation-invariant (and not equal to the null function), then $\mathcal{P}_{\sqrt{k_{\mathrm{p}}}} = \mathcal{P}_{\boldsymbol{s}_{\mathrm{p}}}$.

Remark 8 (Scalar base kernel norms)**.** *When $K = k\mathrm{Id}$, we have*

$$\|K_x\|_{\mathrm{op}} = \sqrt{k(x,x)}, \quad \textit{and} \quad \sqrt{\langle \boldsymbol{s}_{\mathrm{p}}(x), K(x,x)\boldsymbol{s}_{\mathrm{p}}(x)\rangle_{\mathbb{R}^d}} = \sqrt{k(x,x)}\|\boldsymbol{s}_{\mathrm{p}}(x)\|_{\mathbb{R}^d}.$$

C.1 Proof of Lemma 2: Embedding functionals on RKHS defined by feature maps

By (Carmeli et al., 2010, Prop 1), $\tilde{\xi}$ is a surjective partial isometry from $\mathcal{H}$ onto $\mathcal{H}_K$. Hence it is continuous, and $\tilde{\xi}|_{\ker \tilde{\xi}^T} : \ker \tilde{\xi}^T \to \mathcal{H}_K$ is an isometric isomorphism, where $\ker \tilde{\xi}^T$ is the orthogonal complement to the kernel of $\tilde{\xi}$. If D is continuous, so is $D \circ \tilde{\xi}$ since it is the composition of continuous maps. For the converse, note that $\tilde{\xi} \circ (\tilde{\xi}|_{\ker \tilde{\xi}^T})^{-1} : \mathcal{H}_K \to \mathcal{H}_K$ is the identity so $D = D \circ \tilde{\xi} \circ (\tilde{\xi}|_{\ker \tilde{\xi}^T})^{-1}$, which is continuous if $D \circ \tilde{\xi}$ is.

For the second claim, we apply the first claim to $D \equiv \mathrm{Q}|_{\mathcal{H}_k}$. Noting that the RKHS embedding $\Phi_k(\mathrm{Q}) \in \mathcal{H}_k$ is the function $x \mapsto \mathrm{Q}k_x$, we have

$$\begin{aligned}\|\mathrm{Q}\|_{\mathcal{H}_k}^2 &= \mathrm{Q}(x \mapsto \mathrm{Q}k_x) = \iint k(x,y)\mathrm{Q}(\mathrm{d}y)\mathrm{Q}(\mathrm{d}x) = \iint \langle \xi(x), \xi(y)\rangle_{\mathcal{H}} \mathrm{Q}(\mathrm{d}y)\mathrm{Q}(\mathrm{d}x) \\ &= \iint \tilde{\xi}(\xi(x))(y)\mathrm{Q}(\mathrm{d}y)\mathrm{Q}(\mathrm{d}x) = \int \mathrm{Q} \circ \tilde{\xi}(\xi(x))\mathrm{Q}(\mathrm{d}x) = \int \langle (\mathrm{Q} \circ \tilde{\xi})^*, \xi(x)\rangle_{\mathcal{H}} \mathrm{Q}(\mathrm{d}x) \\ &= \int \tilde{\xi}((\mathrm{Q} \circ \tilde{\xi})^*)(x)\mathrm{Q}(\mathrm{d}x) = \|\mathrm{Q} \circ \tilde{\xi}\|_{\mathcal{H}}^2,\end{aligned}$$

where as usual $(\mathrm{Q} \circ \tilde{\xi})^*$ denoted the embedding of $\mathrm{Q} \circ \tilde{\xi}$ into $\mathcal{H}$.

To generalise the above to Q being any embeddable functional D: letting $\xi_\delta : \mathcal{X} \to \mathcal{B}(\mathbb{R}^d, \mathcal{H}_K)$ denote the canonical feature map, $\xi_\delta(x) = K(\cdot, x)$, then

$$\xi_\delta = \tilde{\xi} \circ \xi,$$

since for any $x, y \in \mathcal{X}$, $c \in \mathbb{R}^d$ $(\xi_\delta(y)c)(x) = \xi_\delta^*(x)\xi_\delta(y)c = K(x,y)c = \xi^*(x)\xi(y)c = \tilde{\xi}(\xi(y)c)(x)$. Moreover

$$K_x^{e_i} = \tilde{\xi}\xi(x)e_i$$

and if S is embeds into $\mathcal{H}$ and $\xi_f : \mathcal{X} \to \mathcal{B}(\mathbb{R}^d, \mathcal{H})$ is a feature map for K, then

$$\tilde{\xi}_f(S^*) = e_i S \circ \xi_f(\cdot)e_i,$$

since $\tilde{\xi}_f(S^*)(x) = \xi_f^*(x)(S^*) = e_i(\xi_f^*(x)(S^*))^i = e_i\langle e_i, \xi_f^*(x)(S^*)\rangle = e_i S\xi_f(x)e_i$. Hence

$$\|D \circ \tilde{\xi}\|_{\mathcal{H}}^2 = D \circ \tilde{\xi}(D \circ \tilde{\xi})^* = De_iD \circ \tilde{\xi} \circ \xi(\cdot)e_i = De_iDK_\cdot^{e_i} = DD^* = \|D\|_{\mathcal{H}_k}^2.$$

C.2 Proof of Proposition 7: Bochner embeddability vs. score integrability

The fact that $x \mapsto \|K_x\|_{\mathrm{op}}$ and $x \mapsto \|\nabla_x \cdot K\|_K$ are bounded follows from Lemma 3:

Lemma 3 (RKHS boundedness conditions)**.** *If $\mathcal{H}_K$ is a RKHS of $\mathbb{R}^d$-valued functions, then the following claims hold.*

(a) *$\mathcal{H}_K \subseteq \mathrm{L}^\infty(\mathbb{R}^d)$ iff $x \mapsto \|K(x,x)\|$ is bounded.*

(b) *If $\partial_{x^\ell}\mathcal{H}_K$ exists, then $\partial_{x^\ell}\mathcal{H}_K \subseteq \mathrm{L}^\infty(\mathbb{R}^d)$ iff $x \mapsto \|\partial_{x^\ell}\partial_{y^\ell}K(x,y)\|$ is bounded.*

(c) *If $\partial_{x^\ell}\partial_{y^\ell}K$ exists, then $\partial_{x^\ell}\mathcal{H}_K$ exists.*

(d) *If $K \in \mathscr{C}_b^{(1,1)}(\mathbb{R}^d)$, then $\mathcal{H}_K \subseteq \mathscr{C}_b^1(\mathbb{R}^d)$.*

Proof (a) If $\mathcal{H}_K \subseteq \mathrm{L}^\infty(\mathbb{R}^d)$, proceeding as in Appendix P.1, we have $\|K_x^* h\| = \|h(x)\| \leq \|h\|_\infty$ for any $h \in \mathcal{H}_K$, so the Banach-–Steinhaus Theorem implies $\sup_x \|K_x^*\| = \sup_x \sqrt{\|K(x,x)\|}$ is finite. Conversely, when $x \mapsto \|K(x,x)\|$ is bounded, then $\|h(x)\| = \|K_x^* h\| \leq \|h\|_K \|K_x^*\| \leq \|h\|_K \sqrt{\|K(x,x)\|} \leq \|h\|_K \sup_x \sqrt{\|K(x,x)\|}$.

(b) Similarly, if $\mathcal{H}_K$ is a RKHS of differentiable functions, then $\partial_{x^\ell} H_K$ is a RKHS with matrix-valued kernel $(x,y) \mapsto \partial_{x^\ell}\partial_{y^\ell} K(x,y)$ by Lemma 4. Thus, from above, $\partial_{x^\ell} H_K \subseteq \mathrm{L}^\infty(\mathbb{R}^d)$ iff $x \mapsto \|\partial_{1^\ell}\partial_{2^\ell} K(x,x)\|$ is bounded.

(c) If $\partial_{1^\ell}\partial_{2^\ell} K$ exists, then the argument of Micheli and Glaunes (2013, p. 8 near Eq. (5)) shows $\partial^\ell h$ exists for all $h \in \mathcal{H}_K$.

(d) If $K \in \mathscr{C}_b^{(1,1)}(\mathbb{R}^d)$, then $\mathcal{H}_K \subseteq \mathscr{C}^1(\mathbb{R}^d)$ by Micheli and Glaunes (2013, Thm. 2.11), and by above $\mathcal{H}_K \subseteq \mathscr{C}_b(\mathbb{R}^d)$. Proceeding as above, we have of any $v \in \mathbb{R}^d$, $|\langle v, \partial^p h(x)\rangle| = |\langle h, \partial_2^p K^v(.,x)\rangle| \leq \|h\|_K \|\partial_2^p K^v(.,x)\|_K = \|h\|_K \sqrt{v^T \partial_1^p \partial_2^p K(x,x) v}$ which is bounded in x, and thus $\partial^p h$ is bounded. ■

Now, by definition, since ξ_{P} is a feature map for k_{p},

$$\mathrm{Q}(\sqrt{k_{\mathrm{p}}}) = \int \sqrt{\langle \xi_{\mathrm{P}}(x), \xi_{\mathrm{P}}(x)\rangle_K} \mathrm{dQ} = \mathrm{Q}(\|\xi_{\mathrm{P}}\|_K).$$

Recall $\xi_{\mathrm{P}}(x) = K_x \boldsymbol{s}_{\mathrm{p}}(x) + \nabla_x \cdot K$. By the triangle inequalities

$$\int \big| \|K_x \boldsymbol{s}_{\mathrm{p}}(x)\|_K - \|\nabla_x \cdot K\|_K \big| \, \mathrm{Q}(\mathrm{d}x) \leq \mathrm{Q}(\|\xi_{\mathrm{P}}\|_K) \leq \int \left(\|K_x \boldsymbol{s}_{\mathrm{p}}(x)\|_K + \|\nabla_x \cdot K\|_K\right) \mathrm{Q}(\mathrm{d}x).$$

The result follows by continuity of $K_x = \delta_x^* \in \mathscr{B}(\mathbb{R}^d, \mathcal{H}_K)$, and the assumptions on K:

$$\|K_x\|_{\mathrm{op}} \|\boldsymbol{s}_{\mathrm{p}}(x)\|_{\mathbb{R}^d} \geq \|\delta_x^* \boldsymbol{s}_{\mathrm{p}}(x)\|_K = \sqrt{\langle \delta_x^* \boldsymbol{s}_{\mathrm{p}}(x), \delta_x^* \boldsymbol{s}_{\mathrm{p}}(x)\rangle_K} = \sqrt{\langle \boldsymbol{s}_{\mathrm{p}}(x), K(x,x) \boldsymbol{s}_{\mathrm{p}}(x)\rangle_{\mathbb{R}^d}}.$$

C.3 Proof of Proposition 5: Score vs. Bochner embeddability

The result follows by Proposition 7.

C.4 Proof of Proposition 1: Embeddability conditions

That $\mathcal{P}_{\sqrt{k}} \subseteq \mathcal{P}_{\mathcal{H}_k}$ follows directly from the embeddability of the Dirac measures: $\mathrm{P}|h| = \mathrm{P}|\xi_\delta^*(h)| \leq \|h\|_k \mathrm{P}\|\xi_\delta^*\|_{\mathrm{op}} = \|h\|_k \mathrm{P}\sqrt{k}$, where $\xi_\delta^* : x \mapsto \delta_x|_{\mathcal{H}_k}$ is the canonical feature map. When $\mathcal{H}_k$ is separable, then by Carmeli et al. (2006, Cor. 4.3 and Prop. 4.4) a sufficient condition for Q to embed is $|k| \in \mathrm{L}^1(\mathrm{Q} \otimes \mathrm{Q})$. Note that $\mathrm{Q} \in \mathcal{P}_{\sqrt{k}}$ implies $|k| \in \mathrm{L}^1(\mathrm{Q} \otimes \mathrm{Q})$, as $\iint |k(x,y)| \mathrm{dQ}(x)\mathrm{dQ}(y) = \iint |\langle k_x, k_y\rangle_k| \mathrm{dQ}(x)\mathrm{dQ}(y) \leq \iint \|k_x\|_k \|k_y\|_k \mathrm{dQ}(x)\mathrm{dQ}(y) = \iint \sqrt{k(x,x)}\sqrt{k(y,y)} \mathrm{dQ}(x)\mathrm{dQ}(y) = (\mathrm{Q}\sqrt{k})^2$.

The following example is adapted from (Berlinet and Thomas-Agnan, 2004, p.204). Take $k(i,j) = \mathbb{I}[i=j]$ for $i,j \in \mathbb{N}^*$, i.e. $\mathcal{H}_k = \ell^2(\mathbb{N}^*)$, and consider the Radon measure $\mu(i) = 1/i$. Then it is easy to see that μ is Pettis-embeddable (and satisfies $|\mu| \otimes |\mu|(k) < \infty$), but that it is not Bochner-embeddable into $\mathcal{H}_k$, since $|\mu|(\sqrt{k}) = \infty$.

C.5 Proof of Theorem 1: KSD as MMD

First let us show the following differential reproducing property, which is a mild generalization of results in (Steinwart and Christmann, 2008; Zhou, 2008; Micheli and Glaunes, 2013).

In contrast to the results provided in these references, Lemma 4 does not assume continuity of the derivatives. We note that the proof is essentially identical to the continuous derivative case in Micheli and Glaunes (2013, Thm. 2.11) (which also deals with matrix-valued kernels). This generalized form is important for our results as it allows us to establish that MMD and KSD coincide under no additional conditions than those required for them to be well-defined.

Lemma 4 (Differential reproducing property)**.** *Suppose that $\partial_{x^\ell}\mathcal{H}_K$ exists. Then for all $c \in \mathbb{R}^d$ and $h \in \mathcal{H}_K$*

$$\langle \partial_{x^\ell} K(\cdot, x)c, h\rangle_K = \langle c, \partial_{x^\ell} h(x)\rangle.$$

Moreover $\partial_{x^\ell}\mathcal{H}_K$ is a RKHS with kernel $(x, y) \mapsto \partial_{x^\ell}\partial_{y^\ell} K(x, y)$.

Proof Given a sequence $\epsilon_n > 0$ with $\epsilon_n \to 0$, define $\Delta_{\epsilon_n} \equiv (K(\cdot, x + \epsilon_n e_\ell)c - K(\cdot, x)c)/\epsilon_n \in \mathcal{H}_K$. Since $K(\cdot, x)c \in \mathcal{H}_K$ its partial derivative in direction e_ℓ exists, and thus $K(y, \cdot)c$ also has a partial derivative in direction e_ℓ, hence Δ_{ϵ_n} converges pointwise to $\partial_{x^\ell} K(\cdot, x)c$. Moreover, for any $h \in \mathcal{H}_K$, $\langle \Delta_{\epsilon_n}, h\rangle_K = \langle c, (h(x + \epsilon_n e_\ell) - h(x))/\epsilon_n\rangle$ converges to $\langle c, \partial_{x^\ell} h(x)\rangle$ as $\epsilon_n \to 0$ (since $\partial_{x^\ell} h$ exists). By the Banach–Steinhaus theorem $\{\Delta_{\epsilon_n}\}_n$ is thus a bounded subset of $\mathcal{H}_K$, and by Micheli and Glaunes (2013, Cor. 2.8) it follows that $\partial_{x^\ell} K(\cdot, x)c \in \mathcal{H}_K$ and $\langle c, \partial_{x^\ell} h(x)\rangle = \lim_{n\to\infty}\langle \Delta_{\epsilon_n}, h\rangle_K = \langle \partial_{x^\ell} K(\cdot, x)c, h\rangle_K$ for all $h \in \mathcal{H}_K$.

We now show $\xi^* \equiv \partial_{x^\ell} : \mathcal{X} \to \mathcal{B}(\mathcal{H}_K, \mathbb{R}^d)$, with $\xi^*(x) \equiv \partial_{x^\ell}|_x$, is a feature map with associated kernel $\partial_{y^\ell}\partial_{x^\ell} K$. By above $\langle e_i, \xi^*(x)\rangle$ is continuous for all x, i, so indeed ξ^* is $\mathcal{B}(\mathcal{H}_K, \mathbb{R}^d)$-valued. Note $\tilde{\xi} = \partial_{x^\ell}$, so by Carmeli et al. (2006, Prop. 2.4) $\partial_{x^\ell}\mathcal{H}_K$ is a RKHS with kernel $\overline{K}$ s.t.,

$$\overline{K}(x, y)c = \xi^*(x)\xi(y)c = \partial_{x^\ell}|_x \partial_{y^\ell} K(\cdot, y)c = \partial_{x^\ell}\partial_{y^\ell} K(x, y)c.$$

■

Now we apply Lemma 4 to the RKHS $\mathrm{p}\mathcal{H}_K$ whose kernel is $\mathrm{p}(x)K(x, y)\mathrm{p}(y)$ (see Appendix B), in order to show that $\xi_\mathrm{p} : \mathcal{X} \to \mathcal{H}_K$ defined by

$$\xi_\mathrm{p}(x) \equiv \tfrac{1}{\mathrm{p}(x)}\nabla_x \cdot (\mathrm{p}K) \equiv \tfrac{1}{\mathrm{p}(x)}\textstyle\sum_i \partial_{x^i}(\mathrm{p}K(\cdot, x)e_i)$$

is a feature map for k_p. Indeed by Lemma 4, and using (Carmeli et al., 2010, Prop. 1) to relate the inner products of K and $\mathrm{p}K\mathrm{p}$,

$$\begin{aligned}\mathcal{S}_\mathrm{p}(h)(x) &= \textstyle\sum_i \frac{1}{\mathrm{p}(x)}\partial_{x^i}(\mathrm{p}h^i) = \sum_i \frac{1}{\mathrm{p}(x)}\langle \partial_{x^i}(\mathrm{p}(\cdot)K(\cdot, x)\mathrm{p}(x)e_i), \mathrm{p}h\rangle_{\mathrm{p}K\mathrm{p}} \\ &= \textstyle\sum_i \frac{1}{\mathrm{p}(x)}\langle \partial_{x^i}(K(\cdot, x)\mathrm{p}(x)e_i), h\rangle_K = \langle \sum_i \frac{1}{\mathrm{p}(x)}\partial_{x^i}(K(\cdot, x)\mathrm{p}e_i), h\rangle_K.\end{aligned}$$

This shows $\mathcal{S}_\mathrm{p}$ is the feature operator associated to ξ_p, and thus $\mathcal{H}_{k_\mathrm{p}} \equiv \mathcal{S}_\mathrm{p}(\mathcal{H}_K)$ is a RKHS with kernel $k_\mathrm{p}(x, y) \equiv \langle \xi_\mathrm{p}(x), \xi_\mathrm{p}(y)\rangle_K$ (Carmeli et al., 2010, Prop. 1).

The following lemma Lemma 5 concludes.

Lemma 5 (Feature operators preserve unit balls)**.** *Suppose $\tilde{\xi} : \mathcal{H}_K \to \mathcal{H}_{\overline{K}}$ is a feature operator. Then $\tilde{\xi}(\mathcal{B}_K) = \mathcal{B}_{\overline{K}}$.*

Proof Recall $\tilde{\xi}$ is a surjective partial isometry (Carmeli et al., 2010, Prop. 1), in particular $\|\xi(h)\|_{\overline{K}} \leq \|h\|_K$, so $\tilde{\xi}(\mathcal{B}_K) \subseteq \mathcal{B}_{\overline{K}}$. Moreover, since $\tilde{\xi}|_A$ is an isometric isomorphism from the orthogonal complement of its kernel $A \equiv \ker \tilde{\xi}^{\perp}$ onto $\mathcal{H}_{\overline{K}}$, it follows that for any $g \in \mathcal{B}_{\overline{K}}$ there exists $g \in A$ s.t., $1 \geq \|g\|_{\overline{K}} = \|h\|_K$, which concludes. ■

C.6 Proof of Proposition 3: Stein embeddability conditions

The result is an immediate consequence of Proposition 7 and the following proposition.

Proposition 8 (Stein RKHS with vanishing P-expectations)**.** *Suppose* $\mathcal{H}_{k_{\mathrm{p}}}$ *and* $\mathcal{H}_K \subseteq \mathscr{C}^1$ *are subsets of* $\mathrm{L}^1(\mathrm{P})$*. Then* $\mathrm{P}h = 0$ *for all* $h \in \mathcal{H}_{k_{\mathrm{p}}}$*.*

Proof The result follows from Pigola and Setti (2014, Thm. 2.36), after observing that the distributional and usual derivatives of $\mathscr{C}^1$ functions coincide. ■

Appendix D. Proof of Theorem 2: Bochner P-separation with MMDs

In the proof we will use the fact that if we define the tilted reproducing kernel $\tilde{k}(x,y) \equiv {}^{k(x,y)}\!/_{(1+\sqrt{k}(x))(1+\sqrt{k}(y))}$, whose RKHS is ${}^{\mathcal{H}_k}\!/_{1+\sqrt{k}}$, we have the following immediate relation between the MMDs of k and $\tilde{k}$, which, for instance, may be used to generalize some results from bounded to unbounded kernels:

Proposition 9 (Kernel tilting)**.** *With the notation above*

$$\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) = \mathrm{MMD}_{\tilde{k}}(\tilde{\mathrm{Q}}_n, \tilde{\mathrm{P}}),$$

where $\tilde{\mathrm{Q}} \equiv (1+\sqrt{k})\mathrm{Q}$ *for any* $\mathrm{Q} \in \mathcal{P}_{\sqrt{k}}$*, and*

$$\mathrm{Q}_n f \to \mathrm{P} f \text{ for any } f \in \mathscr{C}_{\sqrt{k}} \;\Leftrightarrow\; \tilde{\mathrm{Q}}_n g \to \tilde{\mathrm{P}} g \text{ for any } g \in \mathscr{C}_b.$$

The rationale for the above proposition is that the map $f \mapsto (1+\sqrt{k})f$ is a vector space isomorphism from $\mathscr{C}_b$ (resp. $\mathscr{C}_0$) to $\mathscr{C}_{\sqrt{k}}$ (resp. $\mathscr{C}_{0,\sqrt{k}}$), which induces TVS and isometric isomorphisms once appropriate topologies have been introduced. In that case the map $\mathrm{Q} \mapsto \tilde{\mathrm{Q}}$ identifies $\mathscr{C}^*_{\sqrt{k}}$ (resp. $\mathscr{C}^*_{0,\sqrt{k}}$) with $\mathscr{C}^*_b$ (resp. $\mathscr{C}^*_0$).

Coming back to the proof of Theorem 2, note the kernel k needs to be characteristic to $\mathrm{P} \in \mathcal{P}_{\sqrt{k}}$, since if it was not, there would be a measure $\mathrm{Q} \in \mathcal{P}_{\sqrt{k}}$ not equal to P such that $\|\mathrm{Q} - \mathrm{P}\|_k = 0$, hence $(\mathrm{Q}_n) \equiv (\mathrm{Q})$ would satisfy (a), and (b) since every distribution is tight on a Radon space, while (c) holds since $(1+\sqrt{k})\mathrm{Q}$ is a finite measure; yet (Q_n) does not converge weakly to P in $\mathcal{P}_{\sqrt{k}}$, since it does not converge weakly in $\mathcal{P}$ as a result of the fact $\mathscr{C}_b$ is a separating set on any metrisable space.

Conversely, let us assume that k is characteristic to $\mathrm{P} \in \mathcal{P}_{\sqrt{k}}$. We will show that, given (c), (a)-(b) is equivalent to (usual!) weak convergence in $\mathcal{P}$. So, applying Lemma 5.1.7 of Ambrosio et al. (2005) to every $f \in \mathscr{C}_{\sqrt{k}}$, gives the equivalence in (6) and concludes. Intuitively speaking, (c) lifts weak convergence in $\mathcal{P}$ to weak convergence in $\mathcal{P}_{\sqrt{k}}$.

Assume (a)-(c). By (b) any subsequence of (Q_n) is tight, so, by Prokhorov's theorem Ambrosio et al. (2005, Thm. 5.1.3), it is relatively compact in $\mathcal{P}$ (equipped with the weak topology) and thus contains yet another subsequence (P_l) that converges weakly in $\mathcal{P}$ to some probability distribution P'. Since $1 + \sqrt{k}$ is continuous, using condition (c) and (5.1.23b) in Ambrosio et al. (2005, Lem. 5.1.7) further implies that $\mathrm{P}' \in \mathcal{P}_{\sqrt{k}}$. Moreover, by (a) and continuity of the inner product, $\langle \mathrm{P}' , f \rangle_k = \lim_l \langle \mathrm{P}_l , f \rangle_k = \langle \mathrm{P} , f \rangle_k$ for any $f \in \mathcal{H}_k$. So, by the Pettis property, the embeddings of P' and P coincide: $\Phi_k(\mathrm{P}') = \Phi_k(\mathrm{P})$. Since k is characteristic to $\mathrm{P} \in \mathcal{P}_{\sqrt{k}}$, we get $\mathrm{P}' = \mathrm{P}$. So we have shown that, out of any subsequence of (Q_n), we can extract a (sub)subsequence that converges weakly to P in $\mathcal{P}$. By a classical argument, the original sequence $(\mathrm{Q}_n)_n$ thus converges weakly to P in $\mathcal{P}$.

For the converse we essentially rely on Proposition 9. Note that if we assume that $\mathrm{Q}_n \to \mathrm{P}$ in $\mathcal{P}$, then, by Lemma 5.1.7 of Ambrosio et al. (2005), (c) is equivalent to weak convergence in $\mathcal{P}_{\sqrt{k}}$. Define the measures $\tilde{\mathrm{P}}$ and $\tilde{\mathrm{Q}}_n$ as $\tilde{\mathrm{P}}(A) \equiv \int_A (1+\sqrt{k}) \,\mathrm{dP}$ and $\tilde{\mathrm{Q}}_n(A) \equiv \int_A (1+\sqrt{k}) \,\mathrm{dQ}_n$ for any measurable Borel set $A \subseteq \mathcal{X}$, and let $\tilde{k}(x, y) \equiv {k(x,y)}/{(1+\sqrt{k}(x))(1+\sqrt{k}(y))}$. By LeCam (1957) since $\mathcal{X}$ is Radon, weak convergence in $\mathcal{P}$ implies tightness, i.e. (b). Since $\mathrm{Q}_n(f) \to \mathrm{P}(f)$ for any $f \in \mathscr{C}_{\sqrt{k}}$ can be re-written as $\tilde{\mathrm{Q}}_n(g) \to \tilde{\mathrm{P}}(g)$ for any $g \in \mathscr{C}_b$, it follows that $(\tilde{\mathrm{Q}}_n)$ converges weakly to $\tilde{\mathrm{P}}$ (in the usual sense). Moreover, (c) also shows that $\tilde{\mathrm{Q}}_n$ and $\tilde{\mathrm{P}}$ are finite (non-negative) measures. So we can apply Prop. 2.3.3 of Berg et al. (1984), which says that the tensor product of finite, non-negative measures is weakly continuous, and get

$$\begin{aligned}\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P})^2 &= (\mathrm{Q}_n - \mathrm{P}) \otimes (\mathrm{Q}_n - \mathrm{P})(k) = (\tilde{\mathrm{Q}}_n - \tilde{\mathrm{P}}) \otimes (\tilde{\mathrm{Q}}_n - \tilde{\mathrm{P}})(\tilde{k}) \\ &= \tilde{\mathrm{P}} \otimes \tilde{\mathrm{P}}(\tilde{k}) - 2\tilde{\mathrm{Q}}_n \otimes \tilde{\mathrm{P}}(\tilde{k}) + \tilde{\mathrm{Q}}_n \otimes \tilde{\mathrm{Q}}_n(\tilde{k}) \quad \longrightarrow \quad 0.\end{aligned}$$

D.1 Weak convergence in $\mathcal{P}_{\sqrt{k}}$ and Wasserstein metric

The following proposition first gives an alternative characterization of $\mathscr{C}_{\sqrt{k}}$, $\mathcal{P}_{\sqrt{k}}$ and weak convergence in $\mathcal{P}_{\sqrt{k}}$. See also (Kanagawa et al., 2022, Section 3.1).

Proposition 10 (Wasserstein vs $\mathscr{C}_{\sqrt{k}}$ convergence)**.** *Let k be a continuous strictly positive definite kernel over a separable metric space $\mathcal{X}$. Let $d_k(x, y) \equiv \|\delta_x - \delta_y\|_k$ be the metric induced by k over $\mathcal{X}$. Then $\mathscr{C}_{\sqrt{k}}$ is the set of functions with 1-growth and $\mathcal{P}_{\sqrt{k}}$ the probability measures with finite first-order moments, both w.r.t. the metric d_k.*

Let $W^1_{d_k}$ denote the Wasserstein-1 distance over $\mathcal{P}_{\sqrt{k}}$ w.r.t. d_k. Then $(\mathcal{X}, d_k)$ is a separable metric space, and $W^1_{d_k}$ metrizes weak convergence in $\mathcal{P}_{\sqrt{k}}$, i.e., for $\mathrm{P}_n, \mathrm{P} \in \mathcal{P}_{\sqrt{k}}$

$$\mathrm{P}_n h \to \mathrm{P} h \quad \forall h \in \mathscr{C}_{\sqrt{k}} \qquad \Longleftrightarrow \qquad W^1_{d_k}(\mathrm{P}_n, \mathrm{P}) \to 0.$$

Moreover $(\mathcal{P}_{\sqrt{k}}, W^1_{d_k})$ is complete whenever $(\mathcal{X}, d_k)$ is.

Note that $(\mathcal{X}, d_k)$ is complete whenever d_k is stronger than the original metric d (i.e. whenever there exists $C > 0$ such that $d(x, y) \leq C d_k(x, y)$), which is for example the case when $\mathcal{X} = \mathbb{R}^d$ equipped with its usual Euclidian metric, and k is polynomial kernel of order ≥ 1.

Proof We will prove that there exists constants $C, C' > 0$ such that

$$C(1 + \sqrt{k}(x, x)) \leq 1 + d_k(x, x_0) \leq C'(1 + \sqrt{k}(x, x)) \tag{9}$$

for all $x, x_0 \in \mathcal{X}$. This shows that $\mathscr{C}_{\sqrt{k}}$ is indeed the set of functions with 1-growth for the metric d_k in the sense of (5.1.21) in Ambrosio et al. (2005); and that $\mathcal{P}_{\sqrt{k}}$ is indeed the set of probability measures P with finite first-order moments, i.e. such that for an arbitrary (and then any) $x_0 \in \mathcal{X}$, $\mathrm{P}(d_k(., x_0)) < \infty$.

The space $(\mathcal{X}, d_k)$ is separable whenever $\mathcal{X}$ is (independently of characteristicness), because, since k is continuous, d_k is also continuous, so the topology defined by d_k is weaker than the original one. Finally, when k is characteristic over $\mathcal{P}_{\sqrt{k}}$, then d_k becomes a metric (i.e. additionally satisfies $d_k(x, y) = 0$ iff $x = y$). So Theorem 7.1.5 of Ambrosio et al. (2005) concludes on the completeness condition, and on the equivalence between weak convergence in $\mathcal{P}_{\sqrt{k}}$ and Wasserstein-1 convergence $W^1_{d_k}$.

We know prove (9). Let $k_0 \equiv k(x_0, x_0)$ and $k_x \equiv k(x, x)$. First, notice that

$$d_k(x, x_0)^2 = k_0 - 2k(x, x_0) + k_x \leq k_0 + 2\sqrt{k_0}\sqrt{k_x} + k_x = (\sqrt{k_0} + \sqrt{k_x})^2$$

Therefore $1 + d_k(x, x_0) \leq 1 + |\sqrt{k_0} + \sqrt{k_x}| \leq 1 + \sqrt{k_0} + \sqrt{k_x} \leq C'(1 + \sqrt{k_x})$ with $C' \equiv 1 + \sqrt{k_0}$. Conversely and similarly, $d_k(x, x_0)^2 = (\sqrt{k}(x_0, x_0) - \sqrt{k}(x, x))^2$. Therefore

$$1 + d_k(x, x_0) \geq 1 + |\sqrt{k_x} - \sqrt{k_0}| \geq 1 + \max(\sqrt{k_x} - \sqrt{k_0}, 0) \geq C(1 + \sqrt{k_x})$$

with $C = 1/(1 + \sqrt{k_0})$. ■

Appendix E. Proof of Theorem 3: Score P-separation with KSDs

When $\|\boldsymbol{s}_{\mathrm{p}}\|$ is finitely integrable with respect to a probability measure Q, then both $\boldsymbol{s}^i_{\mathrm{p}}\mathrm{Q}$ and Q are finite measures, so

$$D_{\mathrm{Q}} = \textstyle\sum_i (\boldsymbol{s}^i_{\mathrm{p}}\mathrm{Q} - \partial_{x^i}\mathrm{Q})e^i \equiv \sum_i D_i e^i$$

where $D_i \in \mathscr{D}^1_{L^1}$ and thus $D_{\mathrm{Q}} \in \mathscr{D}^1_{L^1}(\mathbb{R}^d)$.

Using Corollary 4

$$\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = \| \textstyle\int \xi_{\mathrm{P}} \mathrm{Q} \|_{\mathcal{H}_K} = \|D_Q\|_{\mathcal{H}_K}.$$

Hence $\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = 0$ iff $D_{\mathrm{Q}} = 0$. Now since the matrix kernel K is characteristic to $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$, and $D_{\mathrm{P}} = 0$ by the divergence theorem, we finally obtain $\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = 0$ iff $D_{\mathrm{Q}} = 0$ iff $\mathrm{Q} = \mathrm{P}$.

Appendix F. Proof of Theorem 14: Characteristicness of transformed kernel

Theorem 14 (Characteristicness of transformed kernel)**.** *Let $\phi : \mathscr{E} \to \mathscr{F}$ be a linear continuous map that restricts to a feature operator $\overline{\phi} : \mathcal{H}_K \to \mathcal{H}_{\overline{K}}$. Suppose the following diagram commutes, where all maps are continuous*

$$\begin{array}{ccc} \mathcal{H}_K & \xrightarrow{\iota_K} & \mathscr{E} \\ {\scriptstyle\overline{\phi}}\downarrow & & \downarrow{\scriptstyle\phi} \\ \mathcal{H}_{\overline{K}} & \xrightarrow{\iota_{\overline{K}}} & \mathscr{F} \end{array}$$

If $\phi(\mathscr{C})$ is dense in $\mathscr{F}$, then $\overline{K}$ is characteristic to $\mathscr{F}^*$ when K is characteristic to $\mathscr{C}^*$.

Proof Taking the transpose of the commutative diagram $\phi \circ \iota_K = \iota_{\overline{K}} \circ \overline{\phi} : \mathcal{H}_K \to \mathcal{F}$ yields

$$\iota_K^* \circ \phi^* = \overline{\phi}^* \circ \iota_{\overline{K}}^* : \mathscr{F}^* \to \mathcal{H}_K^*.$$

We want to show $\iota_{\overline{K}}^*$ injective given that ι_K^* is injective. Note that $\overline{\phi}^* \circ \iota_{\overline{K}}^*$ injective implies $\iota_{\overline{K}}^*$ injective, so it is sufficient to show $\iota_K^* \circ \phi^*$ injective. Hence it is sufficient to show $\phi^* : \mathscr{F}^* \to \mathscr{C}^*$ injective, which is equivalent to $\phi(\mathscr{C})$ dense in $\mathscr{F}$ by Treves (1967, Chap 18 Cor. 5). ■

Concretely, it will usually suffice to verify that the image $\phi(\mathscr{C})$ contains the smooth compactly supported functions, since these typically form a dense subset of $\mathscr{F}$.

F.1 Proof of Proposition 4: Preserving characteristicness

The result follows from our general theorem on characteristicness-preserving transformations Theorem 14.

For the first claim, note that $\phi : f \mapsto af$ is a continuous map from $\mathscr{C}_b^1(\mathbb{R}^d)_\beta$ to itself. Moreover $\phi(\mathscr{C}_c^1(\mathbb{R}^d)) = \mathscr{C}_c^1(\mathbb{R}^d)$ since $f/a \in \mathscr{C}_c^1(\mathbb{R}^d)$ for all $f \in \mathscr{C}_c^1(\mathbb{R}^d)$, and $\mathscr{C}_c^1(\mathbb{R}^d)$ is dense in the predual $\mathscr{C}_b^1(\mathbb{R}^d)_\beta$ of $\mathscr{D}_{L^1}^1(\mathbb{R}^d)$.

Recall the family of semi-norms defining the $\mathscr{C}_b^1(\mathbb{R}^d)_\beta$ are parametrized by $\gamma \in \mathscr{C}_0$ and have the form $f \mapsto \|\gamma f\|_\infty$, $f \mapsto \|\gamma \partial f\|_\infty$. We first show that $\phi : f \mapsto f \circ b$ is continuous from $\mathscr{C}_b^1(\mathbb{R}^d)_\beta$ to itself. Note that if $\gamma \in \mathscr{C}_0$ then $\gamma \circ b^{-1} \in \mathscr{C}_0$, since for any $\epsilon > 0$ we have $\overline{b \circ \gamma^{-1} \circ \|\cdot\|^{-1}[\epsilon, \infty)} = b \circ \overline{\gamma^{-1} \circ \|\cdot\|^{-1}[\epsilon, \infty)}$ because b is a homeomorhism, and the resulting set is compact since $\overline{\gamma^{-1} \circ \|\cdot\|^{-1}[\epsilon, \infty)}$ is compact (because γ is $\mathscr{C}_0$) and b preserves compactness by continuity. Thus

$$\|\gamma f \circ b\|_\infty = \|\gamma \circ b^{-1} \circ b\, f \circ b\|_\infty = \|\gamma \circ b^{-1} f\|_\infty$$

which is a semi-norm in $\mathscr{C}_b^1(\mathbb{R}^d)_\beta$. The same proof works for the family of semi-norms $\|\gamma \partial(f \circ b)\|_\infty$ once we have observed that $\partial(f \circ b) = \partial f \circ b \cdot \partial b$ (where $\cdot$ denotes matrix multiplication). Indeed

$$\|\gamma \partial(f \circ b)\|_\infty = \|\gamma \partial f \circ b \cdot \partial b\|_\infty = \|\gamma \circ b^{-1} \partial f \cdot \partial b \circ b^{-1}\|_\infty \leq \|\partial b\|_\infty \|\gamma \circ b^{-1} \partial f\|_\infty$$

since ∂b is bounded as b is Lipschitz. Thus ϕ is continuous. It remains to show that $\phi(\mathscr{C}_c^1(\mathbb{R}^d)) = \mathscr{C}_c^1(\mathbb{R}^d)$, which follows from the fact that $f \in \mathscr{C}_c^1(\mathbb{R}^d)$ implies $f \circ b^{-1} \in \mathscr{C}_c^1(\mathbb{R}^d)$ since $\mathrm{supp}(f \circ b^{-1}) = \overline{b \circ f^{-1}(\{0\}^c)} = b \circ \overline{f^{-1}(\{0\}^c)}$ which is compact.

Finally, (c) follows from the more general statement

Proposition 11 (Scalar vs. vector characteristicness)**.** *Consider a matrix-valued kernel K with $\mathcal{H}_K \hookrightarrow \mathcal{F}^d$ for some topological vector space $\mathcal{F}$. Then K is universal to $\mathcal{F}^d$ iff K_{ii} is universal to $\mathcal{F}$ for all i.*

Proof Recall that $h \in \mathcal{H}_K$ iff $h^i \in \mathcal{H}_{K_{ii}}$ for all i. Note $(f^1, \ldots, f^d) \in \mathcal{F}^d$ iff $f^i \in \mathcal{F}$ for any $i \in [d]$, and $h_n^i \to f^i$ in $\mathcal{F}$ for all i iff $(h_n^1, \ldots, h_n^d) \to (f^1, \ldots, f^d)$ in $\mathcal{F}^d$.

■

Appendix G. Proof of Theorem 4: L^2-ISPD conditions

(a) Let $\hat{\kappa}_j \mathrm{d}x$ be the Bochner measure of $\kappa_j(x-y) \equiv k_j(x,y)$. Then $\hat{\kappa}_j \mathrm{d}x$, has full support, i.e., $\operatorname{supp} \hat{\kappa}_j \mathrm{d}x = \mathbb{R}^d$, since this is equivalent to the characteristicness of κ_j (Simon-Gabriel and Schölkopf, 2018, Thm.17). Moreover $\hat{\kappa}_j \in \mathrm{L}^2$ since $\kappa_j \in \mathrm{L}^2$. Since $\mathcal{H}_{k_j} \subseteq \mathrm{L}^2$, then any measure of the form $f\mathrm{d}x$, with $f \in \mathrm{L}^2$ embeds into $\mathcal{H}_{k_j}$ by Proposition 6.

Then, if $g \in \mathrm{L}^2(\mathbb{R}^d)$, using Barp et al. (2019, Appendix 4)

$$\begin{aligned}\|\Phi_K(g\mathrm{d}x)\|_K^2 &= \textstyle\sum_i \|\langle e_i, \Phi_K(g\mathrm{d}x)\rangle\|_{k_i}^2 = \sum_i \|\Phi_{k_i}(g_i \mathrm{d}x)\|_{k_i}^2 \\ &= \textstyle\sum_i \iint g_i(x)\kappa_i(x-y)g_i(y)\mathrm{d}x\mathrm{d}y.\end{aligned}$$

Moreover, since Plancherel theorem and the convolution theorem are valid for L^2 functions (Schwartz, 1978, Remarque pg. 270), using the fact that g_j, κ_j and $\kappa \star g_j$ are in L^2 by Carmeli et al. (2006, Prop. 4.4), then

$$\begin{aligned}\iint g_i(x)\kappa_i(x-y)g_i(y)\mathrm{d}y\mathrm{d}x &= \textstyle\int g_i(x)\kappa_i \star g_i(x)\mathrm{d}x = \int \hat{g}_i(w)\hat{\kappa}_i(w)\hat{g}_i(w)\mathrm{d}w \\ &= \textstyle\int |\hat{g}_i(w)|^2 \hat{\kappa}_i(w)\mathrm{d}w.\end{aligned}$$

Hence, whenever $g\mathrm{d}x$ is non-zero, i.e., $\|g\|_{\mathrm{L}^2} > 0$, then

$$\|\Phi_K(g\mathrm{d}x)\|_K^2 = \textstyle\sum_i \int |\hat{g}_i(w)|^2 \hat{\kappa}_i(w)\mathrm{d}w > 0$$

since $\hat{\kappa}_i(w)\mathrm{d}w$ is a fully supported non-negative measure.

(b) This follows directly by the definition of ISPD, together with the fact htat $A\left(\mathrm{L}^2(\mathbb{R}^d)\right) \subseteq \mathrm{L}^2(\mathbb{R}^d)$ by boundedness, and that if $\|Ag\|_{\mathrm{L}^2(\mathbb{R}^d)} = 0$ then $\|Ag\| = 0$ a.e. so $\|g\| = 0$ a.e. and thus $\|g\|_{\mathrm{L}^2(\mathbb{R}^d)} = 0$.

(c) Let us first discuss the scalar (and matrix diagonal case), i.e., we show that for a scalar reproducing kernel k, if $\mathcal{H}_k$ is separable, $\sup_x \|k_x\|_{\mathrm{L}^1} < \infty$, and $k_x \in \mathrm{L}^2$ for each x, then $\mathcal{H}_{k\mathrm{Id}} \subseteq \mathrm{L}^2(\mathbb{R}^d)$.

Write $k^*f(x) \equiv (k^*f)(x) \equiv \int k(x,y)f(y)\mathrm{d}y$ when the integral is well-defined. Note, for each y, $k^*k_y \in \mathrm{L}^1$, since

$$\begin{aligned}\|k^*k_y\|_{\mathrm{L}^1} &= \textstyle\int |k^*k_y|(x)\mathrm{d}x = \int |\int k(x,z)k(z,y)\mathrm{d}z|\mathrm{d}x = \iint |k(x,z)k(z,y)|\mathrm{d}x\mathrm{d}z \\ &= \textstyle\int |k(z,y)| \|k_z\|_{\mathrm{L}^1}\mathrm{d}z \leq \|k_y\|_{\mathrm{L}^1} \sup_z \|k_z\|_{\mathrm{L}^1} < \infty.\end{aligned}$$

Note the integral swap is justified by Fubini's theorem. Moreover, $\sup_y \|k^*k_y\|_{\mathrm{L}^1} \leq \sup_z \|k_z\|_{\mathrm{L}^1}^2$.

Now, if $f \in \mathrm{L}^2$, then

$$\iint |f(x)^2 k^*k_x(z)|\mathrm{d}z\mathrm{d}x = \textstyle\int f(x)^2 \|k^*k_x\|_{\mathrm{L}^1}\mathrm{d}x \leq \sup_x \|k^*k_x\|_{\mathrm{L}^1}\|f\|_{\mathrm{L}^2},$$

so by Fubini $\iint |f(x)^2 k^*k_x(z)|\mathrm{d}x\mathrm{d}z < \infty$ and thus $\int |f(x)^2 k^*k_x(\cdot)|\mathrm{d}x$ is finite a.e., i..e, $\sqrt{|f^2 k^*k_z|} \in \mathrm{L}^2$ for almost all z. It follows that $z \mapsto \|\sqrt{|f^2 k^*k_z|}\|_{\mathrm{L}^2} \in \mathrm{L}^2$, since $\int \|\sqrt{|f^2 k^*k_z|}\|_{\mathrm{L}^2}\mathrm{d}z = \iint |f(x)^2 k^*k_x(z)|\mathrm{d}z\mathrm{d}x$, and similarly $z \mapsto |f(z)|\|\sqrt{|k^*k_z|}\|_{\mathrm{L}^2} \in \mathrm{L}^2$.

Hence, $\int\int |f(x)k^*k_z(x)f(z)|\mathrm{d}x\mathrm{d}z < \infty$ since $fk^*k_z \in \mathrm{L}^1$ (being the product of the L^2 functions $\sqrt{|f^2 k^*k_z|}$ and $\sqrt{|k^*k_z|}$) and

$$\iint |f(x)k^*k_z(x)f(z)|\mathrm{d}x\mathrm{d}z \leq \textstyle\int \left(\|\sqrt{|f^2 k^*k_z|}\|_{\mathrm{L}^2}|f(z)|\|\sqrt{|k^*k_z|}\|_{\mathrm{L}^2}\right)\mathrm{d}z < \infty.$$

Finally,

$$\begin{aligned}\|k^*f\|_{\mathrm{L}^2}^2 &= \int (k^*f)^2(y)\mathrm{d}y = \iint k(y,x)f(x)\mathrm{d}x \int k(y,z)f(z)\mathrm{d}z\mathrm{d}y \\ &= \iint f(x)f(z)k^*k_z(x)\mathrm{d}x\mathrm{d}z \leq \iint |f(x)f(z)k^*k_z(x)|\mathrm{d}x\mathrm{d}z < \infty.\end{aligned}$$

It follows by (Carmeli et al., 2006, Prop. 4.4) that $\mathcal{H}_k \subseteq \mathrm{L}^2$.

For a general matrix-valued kernel K, set $G_{ij}(x,z) \equiv \int \langle e_i, K(x,y)K(y,z)e_j\rangle \mathrm{d}y = \sum_l \int K_{il}(x,y)K_{lj}(y,z)\mathrm{d}y$. Note $G_{ij}(x,z) = G_{ji}(z,x)$, and set $G_{ij}^z \equiv G_{ij}(\cdot,z)$. Now for $f \in \mathrm{L}^2(\mathbb{R}^d)$, if $\iint |f_i(x)f_j(z)G_{ij}^z(x)|\mathrm{d}x\mathrm{d}z < \infty$ we have

$$\begin{aligned}\|K^*f\|_{\mathrm{L}^2(\mathbb{R}^d)}^2 &= \int \|K^*f\|^2(y)\mathrm{d}y = \textstyle\sum_l \int |\langle e_l, K^*f\rangle|^2(y)\mathrm{d}y = \sum_l \int \langle e_l, K^*f\rangle|(y)\langle e_l, K^*f\rangle|(y)\mathrm{d}y \\ &= \textstyle\sum_{lij} \iiint K_{li}(y,x)f_i(x)K_{lj}(y,z)f_j(z)\mathrm{d}x\mathrm{d}z\mathrm{d}y \\ &= \textstyle\sum_{lij} \iiint K_{il}(x,y)f_i(x)K_{lj}(y,z)f_j(z)\mathrm{d}x\mathrm{d}z\mathrm{d}y \\ &= \textstyle\sum_{ij} \iint f_i(x)f_j(z)G_{ij}^z(x)\mathrm{d}x\mathrm{d}z.\end{aligned}$$

We can now proceed as in the scalar case with G_{ij}^z taking the place of k^*k_z. Indeed

$$\|G_{ij}^z\|_{\mathrm{L}^1} \leq \|K_{ij}^z\|_{\mathrm{L}^1} \sup_x \|K_{ij}^x\|_{\mathrm{L}^1},$$

and note $|K_{ij}^z| = |\langle e_i, K_z e_j\rangle \leq \|e_i\|\|K_z e_j\|$ so $\|K_{ij}^z\|_{\mathrm{L}^1} \leq \|e_i\|\|K_z e_j\|_{\mathrm{L}^1(\mathbb{R}^d)}$.

(d) Note that if K is a matrix-valued kernel, then $|K_{ij}(x,y)| = |\langle e_i, \xi(x)^*\xi(y)e_j\rangle| \leq \|\xi(x)e_i\|\|\xi(y)e_j\| = \sqrt{K_{ii}(x,x)}\sqrt{K_{jj}(y,y)}$ (for any feature map ξ), so if K is translation-invariant then K is bounded.

In general, if $K_x u$ is both finitely-integrable and bounded, then $\|K_x u\|_{\mathrm{L}^2(\mathbb{R}^d)} \leq \|K_x u\|_{\mathrm{L}^1(\mathbb{R}^d)}\|K_x u\|_{\mathrm{L}^\infty(\mathbb{R}^d)}$.

Appendix H. Proof of Theorem 5: L^2 P-separation with KSDs

(a) Let us first show that under the assumptions of the result, the alternative definition of KSD used in Liu et al. (2016) is equivalent to ours.

By Lemma 2, since Q embeds we have $\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = \|D_Q\|_K$. Moreover since $\mathrm{Q} \in \mathcal{P}_{K,0}$, we know $T_Q \equiv \mathrm{Q} \circ \mathcal{S}_\mathrm{q}$ embeds to zero in $\mathcal{H}_K$, $\Phi_K(T_\mathrm{Q}) = 0$, and for any $h \in \mathcal{H}_K$, since $\mathcal{S}_\mathrm{p}(h)$ and $\mathcal{S}_\mathrm{q}(h)$ are finitely Q-integrable, we have $D_Q - T_Q(h) = \mathrm{Q}\mathcal{S}_\mathrm{p}(h) - \mathrm{Q}\mathcal{S}_\mathrm{q}(h) = \mathrm{Q}(\mathcal{S}_\mathrm{p}(h) - \mathcal{S}_\mathrm{q}(h)) = (\boldsymbol{s}_\mathrm{p} - \boldsymbol{s}_\mathrm{q})\mathrm{Q}(h)$. Hence

$$\begin{aligned}\mathrm{KSD}_{K,\mathrm{P}}^2(\mathrm{Q}) = \|D_Q\|_K^2 &= \|D_Q - T_Q\|_K^2 = \|(\boldsymbol{s}_\mathrm{p} - \boldsymbol{s}_\mathrm{q})\mathrm{Q}\|_K^2 \\ &= \iint \langle \boldsymbol{s}_\mathrm{p}(y) - \boldsymbol{s}_\mathrm{q}(y), K(y,x)(\boldsymbol{s}_\mathrm{p}(x) - \boldsymbol{s}_\mathrm{q}(x))\rangle \mathrm{d}\mathrm{Q}(y)\mathrm{d}\mathrm{Q}(x).\end{aligned}$$

Now, recall K is $\mathrm{L}^2(\mathbb{R}^d)$ ISPD iff it is characteristic to $\mathrm{L}^2(\mathbb{R}^d)$ (Simon-Gabriel and Schölkopf, 2018, Thm. 6), so the result follows from $\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = \mathrm{MMD}_K((\boldsymbol{s}_\mathrm{p} - \boldsymbol{s}_\mathrm{q})\mathrm{Q}, 0)$, and $(\boldsymbol{s}_\mathrm{p} - \boldsymbol{s}_\mathrm{q})\mathrm{Q} \in \mathrm{L}^2(\mathbb{R}^d)$.

(b) We want to show that $\mathcal{H}_K$ and $\mathcal{S}_\mathrm{q}(\mathcal{H}_K)$ are subsets of $\mathrm{L}^1(\mathrm{Q})$ in order to apply (g) in Proposition 3. By assumption $\mathcal{H}_K \subseteq \mathrm{L}^\infty(\mathbb{R}^d) \subseteq \mathrm{L}^1(\mathrm{Q})$, and $\mathcal{S}_\mathrm{p}(\mathcal{H}_K) \subseteq \mathrm{L}^1(\mathrm{Q})$. Note $\mathcal{S}_\mathrm{q}(h) = \mathcal{S}_\mathrm{p}(h) + \langle \boldsymbol{s}_\mathrm{q} - \boldsymbol{s}_\mathrm{p}, h\rangle$, so we have for any $h \in K$

$$\mathrm{Q}|\mathcal{S}_\mathrm{q}(h)| = \mathrm{Q}|\mathcal{S}_\mathrm{p}(h) + \langle \boldsymbol{s}_\mathrm{q} - \boldsymbol{s}_\mathrm{p}, h\rangle| \leq \mathrm{Q}|\mathcal{S}_\mathrm{p}(h)| + \mathrm{Q}|\langle \boldsymbol{s}_\mathrm{q} - \boldsymbol{s}_\mathrm{p}, h\rangle|.$$

Moreover $\mathrm{Q}|\langle \boldsymbol{s}_\mathrm{q} - \boldsymbol{s}_\mathrm{p}, h\rangle| = \langle \mathrm{q}(\boldsymbol{s}_\mathrm{q} - \boldsymbol{s}_\mathrm{p}), h\rangle_{\mathrm{L}^1(\mathbb{R}^d)} \leq \|\mathrm{q}(\boldsymbol{s}_\mathrm{q} - \boldsymbol{s}_\mathrm{p})\|_{\mathrm{L}^2(\mathbb{R}^d)}\|h\|_{\mathrm{L}^2(\mathbb{R}^d)}$, which concludes.

(c) For any $h \in \mathcal{H}_K$ we have

$$\mathrm{Q}|\mathcal{S}_\mathrm{p}(h)| \leq \textstyle\sum_i \mathrm{Q}|\partial_i h^i| + \mathrm{Q}|\langle \boldsymbol{s}_\mathrm{p}, h\rangle| \leq \sum_i \|\partial_i h^i\|_\infty + \|\mathrm{q}\boldsymbol{s}_\mathrm{p}\|_{\mathrm{L}^2(\mathbb{R}^d)}\|h\|_{\mathrm{L}^2(\mathbb{R}^d)} < \infty.$$

Appendix I. Fourier Transforms

If μ is a finite measure, its Fourier transform is $\hat{\mu}(x) \equiv \int e^{-ix^T w}\mathrm{d}\mu(w)$, which is a positive definite function when μ is a non-negative measure. More generally, if T is a tempered distribution (a.k.a. slowly increasing distribution, see Treves 1967, Chap.25), i.e., an element of the dual of the Schwartz space (a.k.a. space of rapidly decaying functions, see Treves 1967, Chap.10, Example IV), we define its distributional Fourier transform $\hat{T}$ by

$$\hat{T}(\gamma) \equiv T(\hat{\gamma}) \ ,$$

for any function γ in the Schwartz space. In particular if $T = \Phi(x)\mathrm{d}x$, with Φ continuous and slowly increasing, and there exists $f \in \mathrm{L}^2_{\mathrm{loc}}(\mathbb{R}^d/\{0\})$ such that $\hat{T} = f\mathrm{d}x$, then f is known as the generalized Fourier transform (of order 0) of Φ, denoted $\hat{\Phi}$ (Wendland, 2004, Def. 8.9). The above formula then reads $\int \hat{\Phi}(x)\gamma(x)\mathrm{d}x = \int \hat{\gamma}(x)\Phi(x)\mathrm{d}x$.

Appendix J. Proof of Theorem 6: Controlling tight convergence with bounded separation

Since k is continuous, $\mathcal{H} \subseteq \mathscr{C}$, and since $h \in \mathcal{H}_b \subseteq \mathscr{C}_b$ is integrable by any finite measure, we have (with $0/0 = 0$)

$$|\mathrm{Q}h - \mathrm{P}h| \leq \left|\mathrm{Q}\tfrac{h}{\|h\|_k} - \mathrm{P}\tfrac{h}{\|h\|_k}\right| \|h\|_k \leq \sup_{h\in\mathcal{B}_k\cap \mathrm{L}^1(\mathrm{Q})} |\mathrm{Q}h - \mathrm{P}h|\|h\|_k \leq \mathrm{MMD}_k(\mathrm{Q},\mathrm{P})\|h\|_k.$$

Hence if $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0$ then $\mathrm{Q}_n h \to \mathrm{P}h$ for all $h \in \mathcal{H}_b$. Taking $\mathrm{Q}_n = \mathrm{Q}$ for all n, the P-bounded separating assumption implies that k is characteristic to $\mathrm{P} \in \mathcal{P}$. On the other hand, if (Q_n) is a tight sequence, then it is sequentially compact, i.e., any subsequence has a $\mathscr{C}_b$-convergent subsequence, whose weak limit must be P by P-characteristicness (Ethier and Kurtz, 2009, Lemma 4.3), which in turn implies that $\mathrm{Q}_n \to_{\mathscr{C}_b} \mathrm{P}$ (see also proof of Theorem 2: Bochner P-separation with MMDs for additional details). Hence k controls tight convergence.

Appendix K. Proof of Theorem 7: Controlling tight convergence with bounded Stein kernels

By Lemma 11 the shifted kernel $K(x,y)/\theta(x)\theta(y)$ is universal to (i.e., dense in) $\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)$. Moreover, by assumption on the score growth and base kernel the associated Stein RKHS consists of bounded functions. The following lemma concludes:

Lemma 6 (Controlling tight convergence with bounded Stein kernels)**.** *Suppose a matrix-valued kernel K with $\mathcal{H}_K \subseteq \mathscr{C}^1_b(\mathbb{R}^d)$ is characteristic to $\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)^*_\beta$. If $\mathrm{P} \in \mathcal{P}_{K,0}$ and k_p is bounded, then k_p is P-separating and controls tight P-convergence. Moreover, k_p is bounded iff $x \mapsto \sqrt{\langle \boldsymbol{s}_\mathrm{p}(x), K(x,x)\boldsymbol{s}_\mathrm{p}(x)\rangle}$ is bounded.*

Proof We apply Proposition 12 with $\theta(x) \equiv \|\boldsymbol{s}_{\mathrm{p}}(x)\| + 1$. Note that if $\|\boldsymbol{s}_{\mathrm{p}}(x)\|\mathcal{H}_K$ is bounded, then $\mathcal{H}_K \subseteq \mathscr{C}^1_{b,\theta}(\mathbb{R}^d)$. Moreover $\|\boldsymbol{s}_{\mathrm{p}}(x)\|\mathcal{H}_K = \mathcal{H}_{\tilde{K}}$ where $\tilde{K}(x,y) \equiv \|\boldsymbol{s}_{\mathrm{p}}(x)\|K(x,y)\|\boldsymbol{s}_{\mathrm{p}}(y)\|$, and $\mathcal{H}_{\tilde{K}}$ is a RKHS of bounded functions iff $x \mapsto \|\boldsymbol{s}_{\mathrm{p}}(x)\|\sqrt{\|K(x,x)\|}$ is bounded by Lemma 3.

Moreover, since k_{p} consists of bounded functions, it then follows by Theorem 6 that when k_{P} is P-separating then it controls tight convergence to P.

■

Appendix L. Proof of Theorem 8: Translation-invariant kernels have rapidly decreasing sub-RKHSes

We will use the following result proved in Appendix L.1.

Lemma 7 (Convolution decay bound)**.** *Fix any $u, v \in \mathrm{L}^1$ and any subadditive function ρ satisfying $|u(x)| \leq U(\rho(x))$ and $|v(x)| \leq V(\rho(x))$ for non-increasing U, V with $U \circ \rho, V \circ \rho$ finitely integrable. Then*

$$u \star v\,(x) \leq \inf_{\alpha \in [0,1]} \|u\|_{\mathrm{L}^1} V(\alpha\rho(x)) + \|v\|_{\mathrm{L}^1} U((1-\alpha)\rho(x))$$

Let us quote Bochner's theorem (Wendland, 2004, Thm. 6.6) (we refer to Appendix I for definitions of Fourier transforms).

Theorem 15 (Bochner's theorem)**.** *A continuous $\mathbb{R}$-valued function on $\mathbb{R}^d$ is positive definite if and only if it is the Fourier transform of a non-negative finite measure.*

Moreover we will use the following lemma, which follows by combining Lemma 4 with the fact that vector-valued RKHSes of continuous functions have locally bounded kernels (Carmeli et al., 2006, Prop. 5.1), or by the closed graph theorem as shown Appendix L.2.

Lemma 8 (Continuity of RKHS inclusion)**.** *Let $\mathcal{F}$ be a complete metrizable TVS, continuously included in the space of functions $\mathcal{X} \to \mathbb{R}^d$. Then $\mathcal{H}_K \subseteq \mathcal{F}$ implies $\mathcal{H}_K \hookrightarrow \mathcal{F}$. In particular $\mathscr{C}^s(\mathbb{R}^d)$ is a complete metrizable TVS.*

We now show k is characteristic to $\mathscr{D}^1_{L^1}$. Indeed, since $\mathcal{H}_k \subseteq \mathscr{C}^1$, then $\mathcal{H}_k \hookrightarrow \mathscr{C}^1$, so we know $\partial_i \partial_{i+d} k$ exists and is separately continuous for all i by Micheli and Glaunes (2013, Thm. 2.11). Since $\partial_i \partial_{i+d} k$ is translation-invariant, it is further continuous. Thus $k(x,y)$ is $\mathscr{C}^{(1,1)}$, and is characteristic to $\mathscr{D}^1_{L^1}$ by Simon-Gabriel and Schölkopf (2018, Thm. 17).

Let $\kappa(x) \equiv k(x,0)$. Now, given the spectral density $\hat{\kappa} : \mathbb{R}^d \to [0,\infty]$, we define the *ironed* radial kernel κ_{iron} on $\mathbb{R}^d$ by

$$\hat{\kappa}_{\mathrm{iron}}(y) \equiv \inf_{w \in \mathbb{R}^d :\, 0 \leq \|w\| \leq \|y\|} \hat{\kappa}(w)$$

and show it is characteristic to $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$. Note $\hat{\kappa}_{\mathrm{iron}} : \mathbb{R}^d \to [0,\infty]$, and is finite-valued except at the origin (since $\hat{\kappa}$ is). Since $\hat{\kappa}_{\mathrm{iron}} \leq \hat{\kappa}$, $\hat{\kappa} \in \mathrm{L}^1(\mathbb{R}^d)$, $\infty > \int \hat{\kappa}(w)\mathrm{d}w \geq \int \hat{\kappa}_{\mathrm{iron}}(w)\mathrm{d}w = \|\hat{\kappa}_{\mathrm{iron}}\|_{\mathrm{L}^1(\mathbb{R}^d)}$, so $\hat{\kappa}_{\mathrm{iron}}(w)\mathrm{d}w$ is a finite non-negative measure whose Fourier transform defines

a continuous positive definite (radial) kernel κ_{iron} by Theorem 15. Moreover $\hat{\kappa}_{\text{iron}}$ is strictly positive since $\hat{\kappa}$ is bounded away from zero on the compact set $\overline{\mathcal{B}_r}$ for any $r \geq 0$. In particular κ_{iron} is characteristic to $\mathscr{D}^0_{L^1}$ (Simon-Gabriel and Schölkopf, 2018, Thm. 17).

Moreover, by Steinwart and Christmann (2008, Sec. 4.3) $\partial_{x^i}\partial_{y^i}k(x,y)$ is a continuous translation-invariant kernel, and $\partial_{x^i}\partial_{y^i}k(x,y) = -\partial_i^2\kappa(x-y)$. This implies that $\int \|w\|^2\hat{\kappa}(w)\mathrm{d}w < \infty$. Indeed $\widehat{(-\partial_i^2\kappa)}(w) = w_i^2\hat{\kappa}(w)$ (Treves, 1967, Thm. 25.7), and by Theorem 15 the generalized Fourier transform of a continuous translation-invariant kernel is integrable, i.e., $w \mapsto w_i^2\hat{\kappa}(w) \in \mathrm{L}^1(\mathbb{R}^d)$. From $\kappa \geq \kappa_{\text{iron}}$, it follows that $w \mapsto \|w\|^2\hat{\kappa}_{\text{iron}}(w) \in \mathrm{L}^1(\mathbb{R}^d)$, and thus $\kappa_{\text{iron}} \in \mathscr{C}^2$, since by Leibniz integral rule the second partial derivatives exist and are continuous. Hence k_i is characteristic to $\mathscr{D}^1_{L^1}$ by Simon-Gabriel and Schölkopf (2018, Thm. 17).

Now define a radial kernel κ_s on $\mathbb{R}^d$ by $\hat{\kappa}_s(x) = \hat{\kappa}_{\text{iron}}(2x)$ which is strictly positive so also characteristic to $\mathscr{D}^0_{L^1}$ (more generally we can compose $\hat{\kappa}_{\text{iron}}$ with any homeomorphism, as their preimage commutes with closure), and $\kappa_s \in \mathscr{C}^2$ so it characteristic to $\mathscr{D}^1_{L^1}$ (by an argument analogous to that of the previous paragraph).

We now discuss a general mechanism to construct a Schwartz function f that is strictly positive, and has a non-negative Fourier transform with compact support (which even makes f an *entire* function). Later on we will apply this construction to obtain a particular f with root exponential decay. Let us first choose a function $g \in C^\infty$ that is non-negative and compactly supported (we will identify an explicit choice of such a function later in the proof). Then we set $f \equiv \hat{G} \star \hat{G}$ where $G \equiv g \star g$. Note that G is non-negative, smooth and compactly supported with non-negative Fourier transform since by the convolution theorem $\hat{G} = (\hat{g})^2$. Moreover the Schwartz's Paley–Wiener theorem (Treves, 1967, Thm. 29.1) then asserts that its Fourier transform (more precisely, its real part restricted to real inputs, i.e., complex numbers with zero imaginary part) $\hat{G}$ is an indefinitely (real) differentiable function that decays faster than any polynomial, that is for all positive integer m we have constant $C_m > 0$ such that $|\hat{G}(x)| \leq \frac{C_m}{(1+\|x\|)^m}$. Hence so does f by Lemma 7 applied with U, V of the form $r \mapsto (1+r)^{-m}$. Moreover $\hat{f}(x) = (G(-x))^2$ which is non-negative with compact support, and f is strictly positive since $\hat{G}$ (viewed as function of arbitrary complex variables) is entire by the Schwartz's Paley–Wiener theorem, and thus holomorphic, and thus has finitely many isolated zeros, the set of which has Lebesgue-measure zero - hence f is the integral of an almost-everywhere strictly positive function. Moreover, the derivative of f decays faster than any polynomial. Indeed, since $f = \hat{G} \star \hat{G}$, where $\hat{G}$ is a smooth function decaying faster than any polynomial, Leibniz' integral rule yields $\partial_{x^j} f = (\partial_{x^j}\hat{G}) \star \hat{G}$. By Wendland (2004, Thm. 5.16 (6)), $\partial_{x^j}\hat{G}(x) = (\widehat{-iy^jG(y)})(x)$, and since $y \mapsto -iy^jG(y)$ is smooth and compactly supported, Schwartz's Paley–Wiener theorem implies that its Fourier transform will decay faster than any polynomial. The convolution bound Lemma 7 then implies that $\partial_{x^j}\hat{G} \star \hat{G}$ will also decay faster than any polynomial. The above argument can then be iterated to show that f belongs to the Schwartz space.

If in the above we specifically choose the function $g = \psi \star \psi$ with $\psi(x) \equiv \varphi(x^1)\cdots\varphi(x^d)$ where $\varphi(x^1) \equiv \exp(-1/(1-c^2|x^1|^2))\mathbb{I}\left[|x^1| < 1\right]$ for some $c > 0$, then $\hat{\psi}(x) = \hat{\varphi}(x^1)\cdots\hat{\varphi}(x^d)$, which implies $|\hat{\psi}(x)| = O(e^{-c\sum_i\sqrt{|x^i|}})$ by Johnson (2018). It follows that $|\hat{g}(x)| = O(e^{-2c\sum_i\sqrt{|x^i|}})$, and thus $\hat{G} = (\hat{g})^2 = O(e^{-4c\sum_i\sqrt{|x^i|}})$, so $f = \hat{G} \star \hat{G} = O(e^{-2c\sum_i\sqrt{|x^i|}})$ by the convolution bound Lemma 7 with $\alpha = 1/2$ and the subadditive function $\rho(x) = \sum_i\sqrt{|x^i|}$.

Similarly, $|\partial_{x^j}\hat{G}| = 2|\hat{g}\partial_{x^j}\hat{g}| \leq 2\|\partial_{x^j}\hat{g}\|_\infty|\hat{g}|$, and $\partial_{x^j}\hat{g}$ is bounded by the Schwartz's Paley-Wiener theorem as above, so $\partial_{x^j}\hat{G} = O(e^{-2c\sum_i\sqrt{|x^i|}})$, so $\partial_{x^j}f = (\partial_{x^j}\hat{G})\star\hat{G} = O(e^{-c\sum_i\sqrt{|x^i|}})$ by the convolution bound Lemma 7 with $\alpha = 1/2$ and the subadditive function $\rho(x) = \sum_i\sqrt{|x^i|}$.

Now define $k_f(x,y) \equiv f(x)k_s(x,y)f(y)$, and we will show that $\mathcal{H}_{k_f} \subseteq \mathcal{H}_k$, by leveraging the translation-invariance of the kernel $k_{fs}(x,y) \equiv \kappa_s(x-y)f(x-y)$, which is a kernel since f is a positive definite function (since its Fourier transform is a positive function) and thus defines a reproducing kernel. Then $\mathcal{H}_{k_f} \subseteq \mathcal{H}_{k_{fs}}$. Indeed, the former RKHS is simply the set of functions $f\mathcal{H}_{k_s}$ (Paulsen and Raghupathi, 2016, Prop. 6.2).

On the other hand, Aronszajn (1950, Thm. II Sec. 8) implies the latter product RKHS $\mathcal{H}_{k_{fs}} = \mathcal{H}_{k_s} \times \mathcal{H}_f$ consists of the functions in the tensor product RKHS $\mathcal{H}_{k_s} \otimes \mathcal{H}_f$ restricted to the diagonal set $\{(x,x)\} \subseteq \mathbb{R}^d \times \mathbb{R}^d$, while Berlinet and Thomas-Agnan (2004, Thm. 13) shows the tensor product RKHS contains the functions of the form $(x,y) \mapsto g(x)h(y)$ where $g \in \mathcal{H}_{k_s}$ and $h \in \mathcal{H}_f$ (i.e., it is the pullback RKHS defined by the diagonal map $x \mapsto (x,x)$), and hence $\mathcal{H}_{k_{fs}}$ contains all the functions of the form $x \mapsto g(x)h(x)$. Restricting to $h(x) = f(x-0) \in \mathcal{H}_f$ yields the subset inclusion $\mathcal{H}_{k_f} \subseteq \mathcal{H}_{k_{fs}}$, which is moreover a continuous inclusion $\mathcal{H}_{k_f} \hookrightarrow \mathcal{H}_{k_{fs}}$ because inclusions of RKHS are always continuous (Schwartz, 1964, Prop. 2).

Hence to show that $\mathcal{H}_{k_f}$ is a subset of $\mathcal{H}_k$, it is sufficient to show that $\mathcal{H}_{k_{fs}} \subseteq \mathcal{H}_k$. But, conveniently, since k_{fs} is translation invariant, we can now apply Lemma 9, proved in Appendix L.3, to verify this inclusion.

Lemma 9 (RKHS inclusion of product RKHS)**.** *Let k, k_2 be kernels and $\star$ denote the convolution operator. The following claims hold true.*

(a) *If there exists a $\lambda \geq 0$ for which $\lambda k - kk_2$ is a kernel, then $h\mathcal{H}_k \subseteq \mathcal{H}_k$ for any $h \in \mathcal{H}_{k_2}$.*

(b) *Suppose k, k_2 are continuous translation invariant kernels. By Bochner's theorem (Theorem 15), such kernels are the Fourier transform of some finite positive measures which we will call μ and ν. If $\mu \star \nu \ll \mu$ and the density $\frac{\mathrm{d}\mu\star\nu}{\mathrm{d}\mu}$ belongs to $\mathrm{L}^\infty(\mu)$, then $h\mathcal{H}_k \subseteq \mathcal{H}_k$ for any $h \in \mathcal{H}_{k_2}$.*

(c) *Moreover, if μ (resp. ν) above is equivalent to (resp. absolutely continuous with respect to) the Lebesgue measure on $\mathbb{R}^d$, with density q_μ (resp. q_ν), then $h\mathcal{H}_k \subseteq \mathcal{H}_k$ for any $h \in \mathcal{H}_{k_2}$ if $q_\mu \star q_\nu / q_\mu \in \mathrm{L}^\infty(\mathbb{R}^d)$.*

(d) *Similarly, if $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous positive definite function with generalized Fourier transform $\hat{f}$, and $q_\mu \star \hat{f}/q_\mu \in \mathrm{L}^\infty(\mathbb{R}^d)$, then $h\mathcal{H}_k \subseteq \mathcal{H}_k$ for any $h \in \mathcal{H}_{k_2}$, where $k_2(x,y) \equiv f(x-y)$.*

Since $\hat{k}$ is strictly positive, the result states that $\mathcal{H}_{k_{fs}} \subseteq \mathcal{H}_k$ iff $\hat{\kappa}_{fs}/\hat{\kappa} \in \mathrm{L}^\infty$, which, by the convolution theorem, can be written as $\hat{\kappa}_s \star \hat{f}/\hat{\kappa} \in \mathrm{L}^\infty$. To show this we will use Lemma 7 to obtain the convolution bound

$$|(\widehat{\kappa_s} \star \hat{f})(w)| \leq \|\widehat{\kappa}_s\|_{\mathrm{L}^1} U(\tfrac{1}{2}\|w\|) + \|\hat{f}\|_{\mathrm{L}^1} V(\tfrac{1}{2}\|w\|) \ ,$$

where U and V are non-increasing functions that upper bound $\hat{f}$ and $\hat{\kappa}_s$ respectively. We let U be the envelope above f, $U(r) \equiv \sup\{\hat{f}(w) : \|w\| \geq r\}$, and since by construction $\hat{\kappa}_s$

is non-increasing, we can set $V(r) \equiv \hat{\kappa}_s(r)$. Thus

$$|(\widehat{\kappa_s} \star \hat{f})/\hat{\kappa}(w)| \leq \|\widehat{\kappa_s}\|_{\mathrm{L}^1} U(\tfrac{1}{2}\|w\|)/\hat{\kappa}(w) + \|\hat{f}\|_{\mathrm{L}^1} \hat{\kappa}_s(\tfrac{1}{2}\|w\|)/\hat{\kappa}(w) \ .$$

By construction, for any $w \in \mathbb{R}^d$ we have $\hat{\kappa}_s(\frac{1}{2}w) = \hat{\kappa}_{\mathrm{iron}}(w) \leq \hat{\kappa}(w)$, and thus $\hat{\kappa}_s(\frac{1}{2}\cdot)/\hat{\kappa}(\cdot) \in \mathrm{L}^\infty$. Moreover $\hat{f}$ has compact support, and thus so does U, hence $|(\widehat{\kappa_s} \star \hat{f})/\hat{\kappa}| \in \mathrm{L}^\infty$ and $\mathcal{H}_{k_{fs}} \subseteq \mathcal{H}_k$ as claimed. Moreover $\mathcal{H}_{k_f} \hookrightarrow \mathcal{H}_k$ (Schwartz, 1964, Prop. 2).

L.1 Proof of Lemma 7: Convolution decay bound

Fix any $\alpha \in [0,1]$ and let $S_x \equiv \{y \in \mathbb{R}^d : \rho(x-y) \leq \alpha\rho(x)\}$. On this set $\rho(y) \geq (1-\alpha)\rho(x)$ by subadditivity. Now

$$u \star v(x) = \int u(y)v(x-y)\mathrm{d}y = \int_{S_x} u(y)v(x-y)\mathrm{d}y + \int_{S_x^c} u(y)v(x-y)\mathrm{d}y.$$

Moreover,

$$\int_{S_x} |u(y)v(x-y)|\mathrm{d}y \leq \int_{S_x} U(\rho(y))|v(x-y)|\mathrm{d}y \leq \int_{S_x} U((1-\alpha)\rho(x))|v(x-y)|\mathrm{d}y \leq U((1-\alpha)\rho(x))\|v\|_{\mathrm{L}^1}.$$

On the other hand

$$\int_{S_x^c} |u(y)v(x-y)|\mathrm{d}y \leq \int_{S_x^c} |u(y)|V(\rho(x-y))\mathrm{d}y \leq \int_{S_x^c} |u(y)|V(\alpha\rho(x))\mathrm{d}y \leq V(\alpha\rho(x))\|u\|_{\mathrm{L}^1}.$$

L.2 Proof of Lemma 8: Continuity of RKHS inclusion

Consider a convergent sequence $(h_n, h_n) \to (h, f)$ in $\mathcal{H}_K \times \mathcal{F}$. Since $\mathcal{H}_K$ and $\mathcal{F}$ are continuously included in the space of functions $\mathcal{X} \to \mathbb{R}^d$, (h_n) converges pointwise to both h and f, hence $h = f \in \mathcal{H}_K$, and the graph of $\iota : \mathcal{H}_K \to \mathcal{F}$ is closed. Thus (Treves, 1967, Cor. 4, Chap. 17) implies it is continuous, since the product of metrizable (resp. complete) TVS is metrizable (resp. complete).

In particular $\mathscr{C}^s(\mathbb{R}^d)$ is a complete metrizable space by (Treves, 1967, Ex. 1 Chap. 10).

L.3 Proof of Lemma 9: RKHS inclusion of product RKHS

The first result follows by the characterization of Aronszajn (1950), once we note that the product kernel $(x,y) \mapsto k(x,y)k_2(x,y)$ contains the functions of the form $x \mapsto h(x)f(x)$ with $h \in \mathcal{H}_{k_2}$ and $f \in \mathcal{H}_k$, since it is the pullback under the diagonal map $x \mapsto (x,x)$ of the tensor product kernel $k \otimes k_2$, and the latter is the completion of the inner product space of functions $(x,y) \mapsto h(x)f(y)$.

The second and third result follow by Zhang and Zhao (2013, Prop. 3.1) and the convolution theorem, which implies that the (translation invariant) product kernel $k(r)k_2(r) = \hat{\mu}(r)\hat{\nu}(r) = \widehat{\mu \star \nu}(r)$. Finally, for the final result, observe that Theorem 15 implies f is the Fourier transform of a non-negative finite measure ν, which satisfies for any γ in the Schwartz space

$$\hat{f}\mathrm{d}x[\gamma] \equiv f\mathrm{d}x[\hat{\gamma}] = \hat{\nu}\mathrm{d}x[\hat{\gamma}] = \nu[\gamma \circ R] \equiv R_*\nu[\gamma],$$

where R_* is the pushforward, and thus $R_*\nu = \hat{f}\mathrm{d}x$ (i.e., $R_*\nu$ is the generalized Fourier transform of f), which implies $\nu = \hat{f} \circ R\mathrm{d}x$. By Wendland (2004, Thm. 6.2) f is even. In

fact $\hat{f} \circ R$ is also the generalized Fourier transform of f, and $\hat{f} \circ R = \hat{f}$ almost everywhere. Indeed, on the one hand $\int \hat{f}\gamma \mathrm{d}x = \int \hat{f} \circ R\gamma \circ RR_* \mathrm{d}x = \int \hat{f} \circ R\gamma \circ R \mathrm{d}x$. On the other hand

$$\int \hat{f}\gamma \mathrm{d}x = \int f\hat{\gamma} \mathrm{d}x = \int f \circ R\hat{\gamma} \mathrm{d}x = \int f \circ R\hat{\gamma} \circ R \circ R \mathrm{d}x = \int f\hat{\gamma} \circ R \mathrm{d}x = \int f\widehat{\gamma \circ R} \mathrm{d}x.$$

Since composition with R is a bijection from the Schwartz space to itself, this shows that $\hat{f} \circ R$ is the generalized Fourier transform of f. The result then follows by the third result.

Appendix M. Proof of Theorem 9: Controlling tight convergence with KSDs

Before proving the result, let us introduce the Banach space $\mathscr{B}^1_\theta(\mathbb{R}^d)$, a generalization of $\mathscr{C}^1_0(\mathbb{R}^d)$, which is easier to handle than the topological vector space $\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)_\beta$ (the analogous generalization of $\mathscr{C}^1_b(\mathbb{R}^d)_\beta$).

Lemma 10 (Definition of $\mathscr{B}^1_\theta(\mathbb{R}^d)$)**.** *Given a continuous function $\theta : \mathbb{R}^d \to [c, \infty)$, for some $c > 0$, let $\mathscr{B}^1_\theta(\mathbb{R}^d)$ be the completion of $\mathscr{C}^1_c(\mathbb{R}^d)$ with respect to*

$$\|f\|_{\mathscr{B}^1_\theta} \equiv \sup \|\theta(x)f(x)\| + \sum_{|p|=1} \sup \|\partial^p_x f\|. \tag{10}$$

Then $\mathscr{B}^1_\theta(\mathbb{R}^d) \cong \{f \in \mathscr{C}^1(\mathbb{R}^d) : \theta f \in \mathscr{C}_0(\mathbb{R}^d), \partial f \in \mathscr{C}_0(\mathbb{R}^{d\times d})\}$.

Proof We first show that if $f \in \{f : \mathscr{C}^1(\mathbb{R}^d) : \theta f \in \mathscr{C}_0(\mathbb{R}^d), \partial f \in \mathscr{C}_0(\mathbb{R}^{d\times d})\}$, then $\exists c_n \in \mathscr{C}^1_c(\mathbb{R}^d)$ such that $c_n \overset{\mathscr{B}^1_\theta}{\to} f$. By definition $\forall \epsilon > 0$ there exists compact subsets S_1, S_2 such that $\|\theta(x)f(x)\| < \epsilon$ for $x \in S_1^c$ and $\|\partial f(x)\| < \epsilon$ for $x \in S_2^c$. Since $S_1 \cup S_2$ is compact, there exists a ball of radius r such that $S \equiv S_1 \cup S_2 \subseteq \mathcal{B}_r$. Using Lemma 14 in Gorham and Mackey (2017) we can find a function $c_\epsilon \in \mathscr{C}^1_c(\mathbb{R}^d)$ with

$$c_\epsilon : \mathbb{R}^d \to [0, 1], \quad c_\epsilon|_{\overline{\mathcal{B}_r}} = 1, \quad c_\epsilon|_{\overline{\mathcal{B}_{r+2\delta}}^c} = 0, \quad \|\partial c_\epsilon\| \leq \mathbb{I}[\overline{\mathcal{B}_{r+2\delta}}/\overline{\mathcal{B}_r}]$$

for some $\delta > 0$. In particular $\partial c_\epsilon = 0$ and $c_\epsilon = 1$ on $S \subseteq \overline{\mathcal{B}_r}$. Now let $f_\epsilon \equiv fc_\epsilon \in \mathscr{C}^1_c(\mathbb{R}^d)$. Then on S we have $\|\theta(x)f(x) - \theta(x)f(x)c_\epsilon(x)\| = 0$ and

$$\|\partial f - \partial(fc_\epsilon)\| = \|\partial f - c_\epsilon \partial f - f \otimes \partial c_\epsilon\| = \|\partial f - c_\epsilon \partial f\| = 0.$$

On S^c, we have $|\theta(x)f(x) - \theta(x)f(x)c_\epsilon(x)| \leq 2|\theta(x)f(x)| \leq 2\epsilon$ and

$$\|\partial f - \partial(fc_\epsilon)\| \leq \|\partial f\| + \|f \otimes \partial c_\epsilon\| + \|c_\epsilon \partial f\| \leq 3\epsilon.$$

Thus $\mathscr{C}^1_c(\mathbb{R}^d)$ is dense in $\{f : \mathscr{C}^1 : \theta f \in \mathscr{C}_0, \partial f \in \mathscr{C}_0\}$.

On the other hand, suppose we have a Cauchy sequence $c_n \in \mathscr{C}^1_c(\mathbb{R}^d)$ for the norm (10). Then, since $\theta \geq c > 0$, $(c_n)_n$ is a fortriori a $\mathscr{C}^1_0(\mathbb{R}^d)$-Cauchy sequence, and thus $\|\cdot\|_{\mathscr{C}^1_0}$-converges to a function $f \in \mathscr{C}^1_0(\mathbb{R}^d)$. Now we show that c_n also converges to f in the norm defined in (10). Indeed $\forall \epsilon > 0$ $\exists \ell$ such that $n, m \geq \ell$ implies $\|\theta(x)c_n(x) - \theta(x)c_m(x)\| \leq \epsilon$ for all x, and thus taking $m \to \infty$ gives $\|\theta(x)c_n(x) - \theta(x)f(x)\| \leq \epsilon$ for all x, i.e., $\|\theta(c_n - f)\|_\infty \leq \epsilon$. An analogous argument shows $\|\partial c_n - \partial f\|_\infty \to 0$, and thus $c_n \overset{\mathscr{B}^1_\theta}{\to} f$.

Finally, note that $\|\theta f - \theta c_n\|_\infty \to 0$ and $\theta c_n \in \mathscr{C}_c(\mathbb{R}^d)$ imply $\theta f \in \mathscr{C}_0(\mathbb{R}^d)$. Similarly $\|\partial f - \partial c_n\|_\infty \to 0$ implies $\partial f \in \mathscr{C}_0(\mathbb{R}^{d\times d})$ since $\partial c_n \in \mathscr{C}_c(\mathbb{R}^{d\times d})$.

■

We will now first prove the non-tilted case, with $a(x) = 1$. We will use the following result, proved in Appendix M.1.

Lemma 11 (Characteristicness of tilted bounded kernels)**.** *Using the notations of Theorem 14, let ϕ be the multiplication by $1/\theta$, where θ is a strictly positive $\mathscr{C}^1$ function such that $1/\theta, \partial(1/\theta)$ are bounded. If K is universal to $\mathscr{C}^1_0(\mathbb{R}^\ell)$ (resp. $\mathscr{C}^1_b(\mathbb{R}^\ell)_\beta$), then $K(x,y)/(\theta(x)\theta(y))$ is universal to $\mathscr{B}^1_\theta(\mathbb{R}^\ell)$ (resp. $\mathscr{C}^1_{b,\theta}(\mathbb{R}^\ell)_\beta$).*

Construct k_s and k_f satisfying the conditions of Theorem 8. Note k_f is characteristic to $(\mathscr{C}^1_{b,\theta})^*_\beta$, as can been seen by applying Lemma 11 to k_s with $\theta(x) \equiv 1/f(x)$, and recalling that the RKHS of $k_f(x,y) = f(x)k_s(x,y)f(y)$ is $f\mathcal{H}_{k_s}$.

Summarizing, we have proven that $\mathcal{H}_{k_f} \subseteq \mathcal{H}_k$ and that k_f is characteristic to $(\mathscr{C}^1_{b,\theta})^*_\beta$. Hence, $\mathcal{H}_{\tilde{K}} \subseteq \mathcal{H}_K$, where $\tilde{K} \equiv k_f\mathrm{Id}$ is characteristic to $\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)^*_\beta$ by Proposition 11. The Stein RKHS associated to $\tilde{K}$ consists of bounded functions and is characteristic to $\mathrm{P} \in \mathcal{P}$ by Proposition 12, because $D_\mathrm{Q} \in \mathscr{C}^1_{b,\theta}(\mathbb{R}^d)^*_\beta$ for any probability measure Q. Moreover, since $\mathcal{H}_{k_\mathrm{P}}$ is a superset of a bounded P-separating sub-RKHS, the result then follows from Theorem 6.

Finally, consider a tilting function a. In the Appendix L we have constructed a Schwartz function f that is strictly positive and s.t., f and its partial derivatives have root exponential decay. Moreover we have shown that $\mathcal{H}_{k_f} \subseteq \mathcal{H}_k$, where $k_f(x,y) \equiv f(x)k_s(x,y)f(y)$ with k_s a kernel obtained by ironing and scaling k, and shown that k_f is universal to $(\mathscr{C}^1_{b,\theta}(\mathbb{R}^d))_\beta$ (with $\theta(x) \equiv 1/f(x)$). Hence $a\mathcal{H}_{k_f} \subseteq a\mathcal{H}_k$. Since af and $\partial(af)$ are bounded, and k_s is universal in $(\mathscr{C}^1_b(\mathbb{R}^d))_\beta$, then by Lemma 11 $a(x)k_f(x,y)a(y)$ is universal to $(\mathscr{C}^1_{b,\frac{\theta}{a}}(\mathbb{R}^d))_\beta$.

M.1 Proof of Lemma 11: Characteristicness of tilted bounded kernels

Note $\phi : \mathscr{C}^1_0(\mathbb{R}^\ell) \to \mathscr{B}^1_\theta(\mathbb{R}^\ell)$ is continuous, indeed (here $\theta^{-1} \equiv 1/\theta$)

$$\|\phi(f)\|_{B^1_\theta} \leq \|f\|_\infty + \|\theta^{-1}\partial f + \partial\theta^{-1}\otimes f\|_\infty \leq \|f\|_\infty + \|\theta^{-1}\|_\infty\|\partial f\|_\infty + \|\partial\theta^{-1}\otimes f\|_\infty \leq C\|f\|_{\mathscr{C}^1_0},$$

for some $C > 0$ (where we have used the boundedness of $|\theta^{-1}|$ and $\|\partial\theta^{-1}\|$), where $\otimes$ is the outer product. Similarly, ϕ is continuous as a map $\mathscr{C}^1_b(\mathbb{R}^\ell)_\beta \to \mathscr{C}^1_{b,\theta}(\mathbb{R}^\ell)_\beta$, since for any $\gamma \in \mathscr{C}_0$

$$\|\gamma\theta\phi(f)\|_\infty = \|\gamma f\|_\infty$$

and

$$\|\gamma\partial_i\phi(f)\|_\infty \leq \|\gamma\theta^{-1}\partial_i f\|_\infty + \|\gamma f\partial_i\theta^{-1}\|_\infty \leq \|\theta^{-1}\|_\infty\|\gamma\partial_i f\|_\infty + \|\partial_i\theta^{-1}\|_\infty\|\gamma f\|_\infty.$$

Moreover $\phi(\mathscr{C}^1_c(\mathbb{R}^\ell)) = \mathscr{C}^1_c(\mathbb{R}^\ell)$ since $\theta^{-1} \in \mathscr{C}^1$ is strictly positive, and $\mathscr{C}^1_c(\mathbb{R}^\ell)$ is dense in $\mathscr{C}^1_{b,\theta}(\mathbb{R}^\ell)$, and in $\mathscr{B}^1_\theta(\mathbb{R}^\ell)$ since the latter is its completion (and metric spaces are dense in their completion). The result then follows by Theorem 14.

Appendix N. Proof of Theorem 10: Controlling P-convergence by dominating indicators

Fix any $\epsilon > 0$, and pick any function $h \in \mathcal{F}$ and compact set C satisfying

$$h - \mathrm{P}h \geq \mathbb{I}\left[C^c\right] - \epsilon/2.$$

Moreover, suppose $d_{\mathcal{F}}(\mathrm{Q}_n, \mathrm{P}) \to 0$. For each n, we have (note h is bounded below, so $h_+ \in \mathrm{L}^1(\mathrm{Q})$ for all $\mathrm{Q} \in \mathcal{P}$)

$$\mathrm{Q}_n(C^c) \leq \epsilon/2 + \mathrm{Q}_n h - \mathrm{P}h,$$

and, since $h \in \mathcal{F}$ and $d_{\mathcal{F}}(\mathrm{Q}_n, \mathrm{P}) \to 0$, we further have $|\mathrm{Q}_n h - \mathrm{P}h| \leq \epsilon/2$ for all n larger than some N. Hence, $\mathrm{Q}_n(C^c) \leq \epsilon$ for all n sufficiently large. Since $\epsilon > 0$ was arbitrary, $(\mathrm{Q}_n)_{n\geq 1}$ is tight.

Finally, if k enforces tightness and controls tight weak convergence, then $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0$ implies (Q_n) is tight, so $\mathrm{Q}_n \to \mathrm{P}$, i.e., k controls weak convergence.

Appendix O. Proof of Lemma 1: Coercive functions dominate indicators

Since $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_k}$, $\mathrm{P}h$ is finite and hence $h - \mathrm{P}h$ is also coercive and bounded below. Since $h - \mathrm{P}h$ is bounded below, there exists $C > 0$ such that $(h - \mathrm{P}h)/C \geq -1$. Moreover, for any $\epsilon > 0$, writing $h_\epsilon \equiv h\epsilon/C \in \mathcal{H}_k$, then $h_\epsilon - \mathrm{P}h_\epsilon \geq -\epsilon$, and since $h_\epsilon - \mathrm{P}h_\epsilon$ is coercive, there exists a compact set S for which $\inf_{x \in S^c} h_\epsilon - \mathrm{P}h_\epsilon \geq 1 - \epsilon$, and therefore $\mathcal{H}_k$ P-dominates indicators.

Appendix P. From Separating Measures in $\mathcal{H}_{k_\mathrm{P}}$ to Separating Schwartz Distributions in $\mathcal{H}_K$

For a bounded RKHS we have the following result shown in Appendix P.2:

Proposition 12 (Separating measures with bounded Stein RKHSes)**.** *Suppose* $\|\boldsymbol{s}_\mathrm{P}(x)\| \leq \theta(x)$*, and* $\mathcal{H}_K \subseteq \mathscr{C}^1_{b,\theta}(\mathbb{R}^d)$*. Then* k_P *is* P*-separating iff* K *is characteristic to* 0 *in* $\{D_\mathrm{Q} : \mathrm{Q} \in \mathcal{P}\} \subseteq (\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)_\beta)^*$*, that is for any* $\mathrm{Q} \in \mathcal{P}$*,* $D_\mathrm{Q}|_{\mathcal{H}_K} = 0 \implies \mathrm{Q} = \mathrm{P}$*.*

In order to prove this it will be convenient to first prove in Appendix P.1 the analogous but simpler result, Proposition 13, which relies on the Banach space defined in Lemma 10.

Proposition 13 (Separating measures with $\mathscr{C}_0$ Stein RKHSes)**.** *Suppose* $\|\boldsymbol{s}_\mathrm{P}(x)\| \leq \theta(x)$*, and* $\mathcal{H}_K \subseteq \mathscr{B}^1_\theta(\mathbb{R}^d)$*. Then* k_P *is* P*-separating iff* K *separates* 0 *from* $\{D_\mathrm{Q} : \mathrm{Q} \in \mathcal{P}\} \subseteq \mathscr{B}^1_\theta(\mathbb{R}^d)^*$*, i.e., for any* $\mathrm{Q} \in \mathcal{P}$*,* $D_\mathrm{Q}|_{\mathcal{H}_K} = 0 \implies \mathrm{Q} = \mathrm{P}$*.*

P.1 Proof of Proposition 13: Separating measures with $\mathscr{C}_0$ Stein RKHSes

Proceeding as in Proposition 9, we can define a Banach space $\mathscr{B}^0_\theta(\mathbb{R}^d)$ such that division by θ yields an isometric isomorphism $\mathscr{C}^0_0(\mathbb{R}^d) \cong \mathscr{B}^0_\theta(\mathbb{R}^d)$. Note the continuous inclusion $\mathscr{B}^1_\theta(\mathbb{R}^d) \hookrightarrow \mathscr{B}^0_\theta(\mathbb{R}^d)$.[7] In particular $\theta\mathrm{Q}$, is a continuous linear functional on $\mathscr{B}^1_\theta(\mathbb{R}^d)$, and

7. When $\theta \geq 1$ we also have $\mathscr{B}^1_\theta(\mathbb{R}^d) \hookrightarrow \mathscr{C}^1_0(\mathbb{R}^d)$.

hence so is $\boldsymbol{s}_{\mathrm{p}}\mathrm{Q}$, since

$$|\boldsymbol{s}_{\mathrm{p}}\mathrm{Q}(f)| \equiv |\sum_i \mathrm{Q}(s^i_{\mathrm{P}} f_i)| \leq \sum_i \mathrm{Q}(|s^i_{\mathrm{P}} f_i|) \leq \tilde{C} \sum_i \mathrm{Q}(\theta|f_i|) \leq C \sup \|\theta f\| \leq C \, \|f\|_{\mathscr{B}^1_\theta}$$

for some constants $\tilde{C}, C > 0$ (that arise from the equivalence of norms on $\mathbb{R}^d$). Moreover, $\partial_i \mathrm{Q}$, and hence D_Q, acts continuously on $\mathscr{B}^1_\theta(\mathbb{R}^d)$, since

$$|\partial_i \mathrm{Q}(f)| = |\mathrm{Q}\partial_i f| \leq \|\partial_i f\|_\infty \leq \|f\|_{\mathscr{B}^1_\theta(\mathbb{R}^d)}.$$

Moreover $\mathcal{H}_K \subseteq \mathscr{B}^1_\theta(\mathbb{R}^d)$ implies $\mathcal{H}_K \hookrightarrow \mathscr{B}^1_\theta(\mathbb{R}^d)$. Indeed, recalling that $K^*_x \in \mathcal{B}(\mathcal{H}_K, \mathbb{R}^d)$ is the evaluation functional, $\|\theta(x)K^*_x h\| = \|\theta(x)h(x)\| \leq \|h\|_{\mathscr{B}^0_\theta(\mathbb{R}^d)}$ for all $h \in \mathcal{H}_K$, so by the Banach–Steinhaus Theorem $\|\theta(x)K^*_x\| \leq C$ for some $C < \infty$. From this we find that $\sup_x \|\theta(x)h(x)\| = \sup_x \|\theta(x)K^*_x h\| \leq \sup_x \|\theta(x)K^*_x\| \|h\|_{\mathcal{H}_K} \leq C\|h\|_{\mathcal{H}_K}$. Proceeding analogously with the derivative contribution, we can use $|\langle \partial_2^p K^{e_i}(.,x), h\rangle_k| \leq \|h\|_{\mathscr{B}^1_\theta}$ to show $\|\langle \partial_2^p K^{e_i}(.,x), \cdot\rangle_k\| \leq A^i$, for some $A^i < \infty$, and then $\sup_x \|\partial^p h(x)\| \leq B \sup_x \max_i |\partial^p h^i(x)| \leq dB \max_i A_i \|h\|_{\mathcal{H}_K}$, which yields the continuity of the inclusion.

Now, note $\mathcal{H}_{k_{\mathrm{P}}} \subseteq \mathscr{C}_0$, so any probability measure Q embeds into the Stein RKHS by Proposition 6. Moreover, since by assumption the embedding of P into $\mathcal{H}_{k_{\mathrm{p}}}$ is the null function, from Corollary 4

$$\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = \|\mathrm{Q}\|_{\mathcal{H}_{k_{\mathrm{P}}}} = \|\mathrm{Q} \circ \mathcal{S}_{\mathrm{p}}\|_{\mathcal{H}_K}.$$

We thus want to show that "$\mathrm{Q} \circ \mathcal{S}_{\mathrm{p}}|_{\mathcal{H}_K} = 0$ implies $\mathrm{Q} \circ \mathcal{S}_{\mathrm{p}}|_{\mathscr{B}^1_\theta} = 0$" iff "$\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = 0$ implies $\mathrm{Q} = \mathrm{P}$".

If $\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = 0$ implies $\mathrm{Q} = \mathrm{P}$, then $\mathrm{Q} \circ \mathcal{S}_{\mathrm{p}}|_{\mathcal{H}_K} = 0$ implies $\mathrm{Q} = \mathrm{P}$, so we want to show $\mathrm{P} \circ \mathcal{S}_{\mathrm{p}}|_{\mathscr{B}^1_\theta} = 0$. For this we can use $\mathrm{P} \circ \mathcal{S}_{\mathrm{p}}|_{\mathscr{C}^1_c} = 0$ by the divergence theorem and the fact $\mathscr{C}^1_c(\mathbb{R}^d)$ is dense in $\mathscr{B}^1_\theta(\mathbb{R}^d)$ since the latter is its completion (and metric spaces are dense in their completion).

Conversely, we have that $\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = 0$ implies $D_{\mathrm{Q}}|_{\mathscr{B}^1_\theta} = 0$, that is the distributional Stein equation

$$\textstyle\sum_i (\boldsymbol{s}^i_{\mathrm{p}}\mathrm{Q} - \partial_{x^i}\mathrm{Q}) e^i = 0,$$

where $(e^i)_{i=1}^{i=d}$ is the dual basis to the canonical basis of $\mathbb{R}^d$. Applying to this vectorial distributional PDE compactly supported smooth vector fields of the form $f = (0, \ldots, 0, l, 0, \ldots)$ with $l \in \mathscr{C}^\infty_c$, yields the system of (scalar) distributional PDEs $\partial_{x^i}\mathrm{Q} = s^i_{\mathrm{P}}\mathrm{Q}$. In particular, solving for the function $q : \mathbb{R}^d \to \mathbb{R}$ the classical PDE $\partial_{x^i} q = s^i_{\mathrm{P}} q$, implies q is the target probability density, $\mathrm{q} = \mathrm{p}$. We then look for solutions via the method of variation of constants. We write the the form $\mathrm{Q} = \mathrm{p}T$. Subbing in and using $\partial_{x^i}(\mathrm{p}T) = \partial_{x^i}\mathrm{p}\, T + \mathrm{p}\partial_{x^i} T$ we obtain the equivalent distributional PDE $\partial_{x^i} T = 0$, which implies that T is a translation-invariant measure, and hence proportional to the Lebesgue measure, $T = C\mathrm{d}x$, by Schwartz (1978, Thm. VI of Chap. II). Since Q is a probability measure we must have $C = 1$, and thus $\mathrm{Q} = \mathrm{P}$.

P.2 Proof of Proposition 12: Separating measures with bounded Stein RKHSes

Since $\mathcal{H}_K \subseteq \mathscr{C}^1_{b,\theta}(\mathbb{R}^d)$, then $\mathcal{H}_{k_\mathrm{p}} \subseteq \mathscr{C}_b$, indeed $|\mathcal{S}_\mathrm{p}(h)(x)| \leq \|\boldsymbol{s}_\mathrm{p}(x)\|\|h(x)\| + |\nabla_x \cdot h| \leq \theta(x)\|h(x)\| + |\nabla_x \cdot h|$ and the latter is a bounded function of x. Hence any probability measure embeds into $\mathcal{H}_{k_\mathrm{p}}$ by Proposition 6, and we can proceed as above once we have shown that D_Q is continuous on $\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)_\beta$. First note that, using the fact that the dual of a finite product of TVS is isomorphic to the finite product of their duals (see, e.g., Treves (1967, p. 259))

$$\mathscr{D}^1_{L^1}(\mathbb{R}^d) \equiv (\mathscr{C}^1_b(\mathbb{R}^d))^*_\beta = (\textstyle\prod_{i=1}^d \mathscr{C}^1_b(\mathbb{R})_\beta)^* = \oplus_{i=1}^d \mathscr{D}^1_{L^1} \subseteq \mathscr{C}^1_{b,\theta}(\mathbb{R}^d)^*_\beta,$$

indeed, $\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)_\beta \hookrightarrow \mathscr{C}^1_b(\mathbb{R}^d)_\beta$ since $\theta \geq c > 0$, and $\mathscr{C}^1_b(\mathbb{R}^d)^*_\beta = \mathscr{D}^1_{L^1}(\mathbb{R}^d)$ by Conway (1965, Sec. 1). Thus $\partial_{x^i} \mathrm{Q}\, e^i \in (\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)^*_\beta$. [8]

Moreover, we have $\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)_\beta \hookrightarrow \mathscr{C}_{b,\theta}(\mathbb{R}^d)_\beta \cong \mathscr{C}_b(\mathbb{R}^d)_\beta$, where the latter isomorphism of topological vector spaces is given by multiplication by θ. Since $(\mathscr{C}_b)^*_\beta$ is the space of finite Radon measures, this shows that $\theta \mathrm{Q} e^i \in \mathscr{C}_{b,\theta}(\mathbb{R}^d)^*_\beta \subseteq \mathscr{C}^1_{b,\theta}(\mathbb{R}^d)^*_\beta$.

Appendix Q. Proof of Theorem 11: IMQ KSDs control P-convergence

Our proof parallels that of Gorham and Mackey (2017, Lem. 16). Fix any $c > 0$, $\gamma \in (0, 2u-1)$, $a > c/2$, and $\alpha \in (1-u, \frac{1}{2}(1-\gamma))$, and consider the functions

$$g_j(x) = -x_j(a^2 + \|x\|^2)^{\alpha-1} \quad \text{for} \quad 1 \leq j \leq d.$$

By Gorham and Mackey (2017, proof of Lem. 16), $g = (g_1, \ldots, g_d) \in \mathcal{H}_K$ for $K = k\mathrm{Id}$. Moreover, the Stein operator applied to g takes the form

$$\mathcal{S}_\mathrm{p}(g)(x) = -\tfrac{\langle \boldsymbol{s}_\mathrm{p}(x), x\rangle}{(a^2+\|x\|^2)^{1-\alpha}} - \tfrac{d}{(a^2+\|x\|^2)^{1-\alpha}} + \tfrac{2(1-\alpha)\|x\|^2}{(a^2+\|x\|^2)^{2-\alpha}}.$$

Since $\alpha < 1$, the final two terms in this expression are uniformly bounded in x. Meanwhile, our generalized dissipativity assumption (8) implies that $-\langle \boldsymbol{s}_\mathrm{p}(x), x\rangle = \Omega(\|x\|^{2u})$ as $\|x\| \to \infty$, so $-\frac{\langle \boldsymbol{s}_\mathrm{p}(x), x\rangle}{(a^2+\|x\|^2)^{1-\alpha}} = \Omega(\|x\|^{2u-2+2\alpha}) = \omega(1)$ since $\alpha > 1-u$. Hence, $\mathcal{S}_\mathrm{p}(g)$ is coercive. In addition, the generalized disspativity condition (8) implies that $-\langle \boldsymbol{s}_\mathrm{p}(x), x\rangle$ is bounded below and hence that $\mathcal{S}_\mathrm{p}(g)$ is bounded below.

Let $K = k\mathrm{Id}$. Since $\boldsymbol{s}_\mathrm{p}$ is well defined and continuous on $\mathbb{R}^d$, the density p is strictly positive and continuously differentiable. In addition, since $\mathrm{P} \in \mathcal{P}_{\boldsymbol{s}_\mathrm{p}}$, $K \in \mathscr{C}_b^{(1,1)}(\mathbb{R}^d)$, and $K \in \mathrm{L}^1(\mathrm{P})$, Proposition 3 and Theorem 1 imply that $\mathrm{p}\mathcal{H}_K \subseteq \mathscr{C}^1(\mathbb{R}^d)$, $\mathrm{P} \in \mathcal{P}_{K,0}$, $\mathrm{KSD}_{K,\mathrm{P}} = \mathrm{MMD}_{k_\mathrm{p}}(\cdot, \mathrm{P})$, and $\mathcal{S}_\mathrm{p}(\mathcal{H}_K) = \mathcal{H}_{k_\mathrm{p}}$. Since $\mathcal{S}_\mathrm{p}(g) \in \mathcal{H}_{k_\mathrm{p}}$, $\mathcal{H}_{k_\mathrm{p}}$ P-dominates indicators by Lemma 1 and enforces tightness by Theorem 10.

Finally, since $k \in \mathscr{C}_b^{(1,1)}$ is translation-invariant with a spectral density bounded away from zero in a neighborhood around the origin (Wendland, 2004, Thm. 8.15), we conclude that $\mathcal{H}_k \subseteq \mathscr{C}^1_b$ by Proposition 3 and that k_p controls P convergence by Corollary 3.

8. A more direct proof that establishes the continuity of $\partial_i \mathrm{Q} e^i$ on $\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)_\beta$ reads as follows: if $f \in \mathscr{C}^1_{b,\theta}(\mathbb{R}^d)$, then $|\partial_i \mathrm{Q} e^i f| = |\mathrm{Q}\partial_i f^i| \leq A \max_{j\in[n]} \|\gamma_j \partial_i f^i\|_\infty$ for some $\gamma_j \in \mathscr{C}_0$, $n \in \mathbb{N}$, $A > 0$, by continuity of Q on $(\mathscr{C}_b)_\beta$. Since $f \mapsto \|\gamma_j \partial f\|_\infty$ are semi-norms on $\mathscr{C}^1_{b,\theta}(\mathbb{R}^d)_\beta$ the result follows.

Appendix R. Proof of Theorem 12: Metrizing P-convergence with bounded Stein kernels

Our aim is to identify a function in $\mathcal{H}_{k_\mathrm{p}}$ that satisfies the indicator bounding property (7) for each $\epsilon > 0$. To this end, for each $m \in \mathbb{N}$, define the compact set $C_m = \{x \in \mathbb{R}^d : \|x\| \leq m\}$, and fix any $m > 1$ for which $-\langle \boldsymbol{s}_\mathrm{p}(x), x\rangle$ is nonnegative on C_{m-1}^c and

$$-\langle \boldsymbol{s}_\mathrm{p}(x), x\rangle - r_0 \left\|\boldsymbol{s}_\mathrm{p}(x)\right\|_1 - 1 - r_0 - 2|\gamma|(1 + \sqrt{d}/(c^2 + m)) \geq (r_1/2)\left\|x\right\|^{2u} \tag{11}$$

holds on C_m^c. These properties hold for all m sufficiently large (specifically, for all m such that $\frac{1}{2} r_1 m^{2u} \geq 1 + r_0 - r_2 + 2|\gamma|(1 + \frac{\sqrt{d}}{c^2+m})$) due to generalized dissipativity (8) with $u > 0$. Fix also any $\epsilon_m \in (0, s]$ satisfying

$$\epsilon_m \sup_{x \in C_m} \max\left(\left\|\boldsymbol{s}_\mathrm{p}(x)\right\|_1 a(\|x\|), \tfrac{2|\gamma|\|x\|_1}{c^2+\|x\|^2} a(\|x\|), \tfrac{\|x\|_1}{c^2+\|x\|^2}, a(\|x\|)\right) \leq \tfrac{a(m)}{m}. \tag{12}$$

Consider the smoothed indicator function

$$\tilde{f}_m(x) = \sigma(m - \|x\|) \quad \text{for} \quad \sigma(r) = 2\max(0, r)^2 \mathbb{I}\left[r < .5\right] + (1 - 2\max(0, 1 - r)^2)\mathbb{I}\left[r \geq .5\right]$$

which satisfies $\tilde{f}_m(x) = 1$ on C_{m-1}, $\tilde{f}_m(x) = 0$ on C_m^c,

$$\mathbb{I}\left[x \in C_{m-1}\right] \leq \tilde{f}_m(x) \leq \mathbb{I}\left[x \in C_m\right], \quad \text{and} \quad -\mathbb{I}\left[x \in C_m \backslash C_{m-1}\right] \leq \partial_{x^i} \tilde{f}_m(x) \leq 0.$$

Moreover, for each $i \in \{1, \ldots, d\}$, $x \mapsto x^i \tilde{f}_m(x) \in \mathscr{C}_0^1$.

Since $\mathcal{H}_k \subseteq \mathscr{C}_0^1$, then $\mathcal{H}_k \hookrightarrow \mathscr{C}_0^1$, so by Simon-Gabriel and Schölkopf (2018, Thm. 6) $\mathcal{H}_k$ is dense in $\mathscr{C}_0^1$. Hence, for each $i \in \{1, \ldots, d\}$ there exists $\tilde{g}_{mi} \in \mathcal{H}_k$ satisfying

$$\sup_{x \in \mathbb{R}^d} \max(|\tilde{g}_{mi}(x) - x^i \tilde{f}_m(x)|, |\partial_{x^i} \tilde{g}_m(x) - \partial_{x^i}(x^i \tilde{f}_m(x))|) \leq \epsilon_m. \tag{13}$$

Moreover, the function $w_i(x) = x^i$ belongs to $\mathcal{H}_{\tilde{k}_i}$ for $\tilde{k}_i(x, y) \equiv x^i y^i$. Since $\mathcal{H}_k \subseteq \mathcal{H}_{k+\tilde{k}_i}$ and $\mathcal{H}_{\tilde{k}_i} \subseteq \mathcal{H}_{k+\tilde{k}_i}$ (see for example Carmeli et al. (2010, Prop. 5)), the functions $\tilde{g}_{mi}, w_i$, and $g_{mi} = w_i - \tilde{g}_{mi}$ are all elements of $\mathcal{H}_{k+\tilde{k}_i}$.

Consider now the Stein function

$$\begin{aligned} h_m(x) &= \textstyle\sum_{i=1}^d \frac{-\partial_{x^i}(p(x) a(\|x\|) g_{mi}(x)))}{p(x)} \\ &= \underbrace{-\langle \boldsymbol{s}_\mathrm{p}(x), \boldsymbol{g}_m(x)\rangle a(\|x\|)}_{(i)} \underbrace{-\langle \boldsymbol{g}_m(x), \nabla a(\|x\|)\rangle}_{(ii)} \underbrace{-a(\|x\|)\nabla \cdot \boldsymbol{g}_m(x)}_{(iii)}, \end{aligned} \tag{14}$$

where $\boldsymbol{g}_m$ is the vector valued function $(g_{mi})_{i=1}^d$. By construction, $h_m \in \mathcal{S}_\mathrm{p}\mathcal{H}_K$, and thus in $\mathcal{H}_{k_\mathrm{p}}$ by Theorem 1. Therefore, the zero-mean embedding assumption $\mathrm{P} \in \mathcal{P}_{K,0}$ and Proposition 3 imply that $\mathrm{P}h_m = 0$. We will show that a rescaled version of h_m satisfies the indicator bound property (7) for a choice of $\tilde{\epsilon}_m$ that decays to 0 as $m \to \infty$. We begin by lower-bounding each of the components in the expansion (14).

To lower-bound term (i), we first record several properties of $\boldsymbol{g}_m$. First, our approximation guarantee (13) implies

$$\sup_{x \in \mathbb{R}^d} |g_{mi}(x) - x^i f_m(x)| \leq \epsilon_m \quad \text{for each} \quad i \in \{1, \ldots, d\}, \tag{15}$$

where $f_m \equiv 1 - \tilde{f}_m$ satisfies

$$\mathbb{I}\left[x \in C_{m-1}^c\right] \geq f_m(x) \geq \mathbb{I}\left[x \in C_m^c\right] \quad \text{and} \quad \mathbb{I}\left[x \in C_m \backslash C_{m-1}\right] \geq \partial_{x^i} f_m(x) \geq 0. \tag{16}$$

Since a is nonnegative and $f_m(x) = 0$ on C_{m-1}, Hölder's inequality, the guarantee (15), the assumed nonnegativity of $-\langle \boldsymbol{s}_{\mathrm{p}}(x), x\rangle$ on C_{m-1}^c, generalized dissipativity, and our choice (12) of ϵ_m implies that

$$\begin{aligned}
&-\langle \boldsymbol{s}_{\mathrm{p}}(x), \boldsymbol{g}_m(x)\rangle a(\|x\|) = -\langle \boldsymbol{s}_{\mathrm{p}}(x), x\rangle f_m(x) a(\|x\|) - \langle \boldsymbol{s}_{\mathrm{p}}(x), \boldsymbol{g}_m(x) - x f_m(x)\rangle a(\|x\|) \\
&\quad \geq -\langle \boldsymbol{s}_{\mathrm{p}}(x), x\rangle f_m(x) a(\|x\|) - \|\boldsymbol{s}_{\mathrm{p}}(x)\|_1 \|\boldsymbol{g}_m(x) - x f_m(x)\|_\infty a(\|x\|) \\
&\quad \geq -\langle \boldsymbol{s}_{\mathrm{p}}(x), x\rangle f_m(x) a(\|x\|) - \|\boldsymbol{s}_{\mathrm{p}}(x)\|_1 a(\|x\|)\epsilon_m \\
&\quad \geq -\langle \boldsymbol{s}_{\mathrm{p}}(x), x\rangle \mathbb{I}\left[x \in C_m^c\right] a(\|x\|) - \|\boldsymbol{s}_{\mathrm{p}}(x)\|_1 a(\|x\|)\epsilon_m \\
&\quad = (-\langle \boldsymbol{s}_{\mathrm{p}}(x), x\rangle - \|\boldsymbol{s}_{\mathrm{p}}(x)\|_1 \epsilon_m)\mathbb{I}\left[x \in C_m^c\right] a(\|x\|) - \|\boldsymbol{s}_{\mathrm{p}}(x)\|_1 \mathbb{I}\left[x \in C_m\right] a(\|x\|)\epsilon_m \\
&\quad \geq (-\langle \boldsymbol{s}_{\mathrm{p}}(x), x\rangle - \|\boldsymbol{s}_{\mathrm{p}}(x)\|_1 s)\mathbb{I}\left[x \in C_m^c\right] a(\|x\|) - a(m)/m.
\end{aligned}$$

To lower bound (ii), we again employ Hölder's inequality, the approximation guarantee (15), and the ϵ_m properties (12) to find that

$$\begin{aligned}
&-\langle \boldsymbol{g}_m(x), \nabla a(\|x\|)\rangle = 2\gamma \langle \boldsymbol{g}_m(x), x\rangle / (c^2 + \|x\|^2)^{\gamma+1} \\
&\quad = 2\gamma \|x\|^2 f_m(x)/(c^2 + \|x\|^2)^{\gamma+1} + 2\gamma \langle \boldsymbol{g}_m(x) - x f_m(x), x\rangle / (c^2 + \|x\|^2)^{\gamma+1} \\
&\quad \geq 2\gamma \|x\|^2 f_m(x)/(c^2 + \|x\|^2)^{\gamma+1} - 2\gamma \|\boldsymbol{g}_m(x) - x f_m(x)\|_\infty \|x\|_1 / (c^2 + \|x\|^2)^{\gamma+1} \\
&\quad \geq 2\gamma \|x\|^2 f_m(x)/(c^2 + \|x\|^2)^{\gamma+1} - 2|\gamma| \|x\|_1 \epsilon_m / (c^2 + \|x\|^2)^{\gamma+1} \\
&\quad \geq -2|\gamma| \|x\|^2 \mathbb{I}\left[x \in C_{m-1}^c\right] / (c^2 + \|x\|^2)^{\gamma+1} - 2|\gamma| \|x\|_1 \epsilon_m / (c^2 + \|x\|^2)^{\gamma+1} \\
&\quad = -2|\gamma| \tfrac{\|x\|^2}{c^2 + \|x\|^2} a(\|x\|)(\mathbb{I}\left[x \in C_m^c\right] + \mathbb{I}\left[x \in C_m \backslash C_{m-1}\right]) \\
&\quad - \tfrac{2|\gamma| \|x\|_1 \epsilon_m}{c^2 + \|x\|^2} a(\|x\|)(\mathbb{I}\left[x \in C_m^c\right] + \mathbb{I}\left[x \in C_m\right]) \\
&\quad \geq -2|\gamma| \left(\tfrac{\|x\|^2 + \|x\|_1 \epsilon_m}{c^2 + \|x\|^2}\right) a(\|x\|)\mathbb{I}\left[x \in C_m^c\right] - 2|\gamma| \max(a(m-1), a(m)) - a(m)/m \\
&\quad \geq -2|\gamma|(1 + \sqrt{d}/(c^2 + m)) a(\|x\|)\mathbb{I}\left[x \in C_m^c\right] - 2|\gamma| \max(a(m-1), a(m)) - a(m)/m.
\end{aligned}$$

To lower bound (iii), we first note that the derivative approximation (13) implies

$$\sup_{x \in \mathbb{R}^d} |\partial_{x^i} g_{mi}(x) - \partial_{x^i}(x^i f_m(x))| \leq \epsilon_m \quad \text{for each} \quad i \in \{1, \ldots, d\}.$$

Moreover, we have

$$\partial_{x^i}(x^i f_m(x)) = f_m(x) + x^i \partial_{x^i} f_m(x) \leq \mathbb{I}\left[x \in C_{m-1}^c\right] + |x^i| \mathbb{I}\left[x \in C_m \backslash C_{m-1}\right]$$

by our $\partial_{x^i} f_m$ constraints (16). Therefore, the nonnegativity of a and the ϵ_m properties (12) give the bound

$$\begin{aligned}
&-a(\|x\|)\nabla \cdot \boldsymbol{g}_m(x) \geq -a(\|x\|)(\mathbb{I}\left[x \in C_{m-1}^c\right] + \|x\|_1 \mathbb{I}\left[x \in C_m \backslash C_{m-1}\right] + \epsilon_m) \\
&\quad = -a(\|x\|)(1 + \epsilon_m)\mathbb{I}\left[x \in C_m^c\right] - a(\|x\|)(1 + \|x\|_1)\mathbb{I}\left[x \in C_m \backslash C_{m-1}\right] - a(\|x\|)\epsilon_m \mathbb{I}\left[x \in C_m\right] \\
&\quad \geq -a(\|x\|)(1 + s)\mathbb{I}\left[x \in C_m^c\right] - \max(a(m-1), a(m))(1 + \sqrt{d}m) - a(m)/m.
\end{aligned}$$

Our assumption $\gamma \leq u$ implies that $\|x\|^{2u} a(\|x\|) \geq m^{2u} a(m)$ whenever $\|x\| \geq m$. This fact combined with our collected results and the assumed growth (11) induced by our choice of m now imply that

$$\begin{aligned} h_m(x) &\geq (-\langle \boldsymbol{s}_{\mathrm{p}}(x), x\rangle - \|\boldsymbol{s}_{\mathrm{p}}(x)\|_1 r_0 - 1 - r_0 - 2|\gamma|(1+\sqrt{d}/(c^2+m)))\mathbb{I}\left[x \in C_m^c\right] a(\|x\|) \\ &\quad - 3a(m)/m - (1+\sqrt{d}m + 2|\gamma|)\max(a(m-1), a(m)) \\ &\geq (r_1/2)\|x\|^{2u}\mathbb{I}\left[x \in C_m^c\right] a(\|x\|) - 3a(m)/m - (1+\sqrt{d}m + 2|\gamma|)\max(a(m-1), a(m)) \\ &\geq (r_1/2)m^{2u}a(m)\mathbb{I}\left[x \in C_m^c\right] - 3a(m)/m - (1+\sqrt{d}m + 2|\gamma|)\max(a(m-1), a(m)). \end{aligned}$$

Hence, the rescaled Stein function $\tilde{h}_m = h_m/((r_1/2)m^{2u}a(m))$, satisfies the indicator approximation property (7) for the compact set C_m and the approximation factor

$$\tilde{\epsilon}_m = 6/(r_1 m^{2u+1}) + (1+\sqrt{d}m + 2|\gamma|)\max(a(m-1), a(m))/((r_1/2)m^{2u}a(m)).$$

Since $u > 1/2$, $\tilde{\epsilon}_m$ vanishes as $m \to \infty$, and hence $\mathcal{H}_{k_{\mathrm{p}}}$ P-dominates indicators. Thus by Theorem 10 the Stein kernel enforces tightness.

For (b), we use Lemma 12:

Lemma 12 (Universal KSDs tilted by score growth control tight convergence)**.** *Suppose that $\|\boldsymbol{s}_{\mathrm{p}}(x)\| \leq (c^2 + \|x\|^2)^{\gamma}$, where $c \neq 0$, $\gamma \geq 0$, and K is characteristic to $\mathfrak{D}^1_{L^1}(\mathbb{R}^d)$. Then the Stein kernel induced by $(c^2+\|x\|^2)^{-\gamma}K(x,y)(c^2+\|y\|^2)^{-\gamma}$ is* P*-separating and controls tight* P*-convergence.*

Proof The result follows by Theorem 7. Indeed the function $\theta(x) \equiv (c + \|x\|^2)^{\gamma}$ has $\partial^1/\theta(x) = -2\gamma x(c+\|x\|^2)^{-\gamma-1}$ satisfies the assumption of Theorem 7, so the result follows. ■

This shows that we can easily construct bounded Stein kernels that control *tight* weak convergence to P in $\mathcal{P}$ by simply tilting the base kernel through a function that bounds the score. By Lemma 12 the Stein kernel induced by the tilted base kernel $a(\|x\|)k(x,y)a(\|y\|)$ controls tight weak convergence, and thus so does the overall Stein kernel (which further controls weak convergence since it enforces tightness) as it may be viewed as the sum of two Stein kernels. Indeed, as proved in Appendix R.1, we have the following general bound between MMDs when an RKHS contains another one:

Lemma 13 (MMD controls subset MMDs)**.** *Suppose $\mathcal{H}_k \subseteq \mathcal{H}_{\tilde{k}}$ and that $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_{\tilde{k}}}$. Then $\exists c \geq 0$ such that for all $\mathrm{Q} \in \mathcal{P}$*

$$\mathrm{MMD}_k(\mathrm{Q}, \mathrm{P}) \leq c\,\mathrm{MMD}_{\tilde{k}}(\mathrm{Q}, \mathrm{P}).$$

Hence,

(i) If k is P*-separating then $\tilde{k}$ is* P*-separating.*

(ii) If k controls (tight) weak P*-convergence, then $\tilde{k}$ controls (tight) weak* P*-convergence.*

Finally, for (c), first note that $\mathcal{H}_{k_\mathrm{p}} \subseteq \mathscr{C}_b$. Indeed for any $h \in \mathcal{H}_{k_\mathrm{p}}$ we have $h(x) = \mathcal{S}_\mathrm{p}(ag) = \langle \boldsymbol{s}_\mathrm{p}(x), ag\rangle + \nabla \cdot (ag) = \langle \boldsymbol{s}_\mathrm{p}, ag\rangle + a\nabla \cdot g + \langle g, \partial a\rangle$, for some vector-valued function g with $g_i \in \mathcal{H}_{k+\tilde{k}_i}$, so h is continuous. Moreover it is bounded since (i) $|g_i(x)| \leq \|g_i\|_{k+\tilde{k}_i}(\sqrt{k(x,x)} + |x^i|)$, implies

$$|\langle g, \partial a\rangle| \leq \|g_i\|_{k+\tilde{k}_i} \textstyle\sum_i 2\gamma \frac{|x^i| \sup_x \sqrt{k(x,x)} + |x^i|^2}{(c^2+\|x\|^2)^{\gamma+1}}.$$

(ii) $a\nabla \cdot g$ is bounded since $\partial_i g \in \mathcal{H}_{\partial_i \partial_{i+d} k+1} \subseteq \mathscr{C}_b$. (iii) $\langle \boldsymbol{s}_\mathrm{p}(x), ag\rangle$ is bounded since

$$|\langle \boldsymbol{s}_\mathrm{p}(x), a(x)g(x)\rangle| \leq \|\boldsymbol{s}_\mathrm{p}(x)\||a(x)|\|g(x)\| \leq \|\boldsymbol{s}_\mathrm{p}(x)\|\|x\||a(x)| \leq 1.$$

Finally, $\mathrm{P} \in \mathcal{P}_{K,0}$ since $\mathcal{H}_{k_\mathrm{p}} \subseteq \mathscr{C}_b \subseteq \mathrm{L}^1(\mathrm{P})$ by above, and $\mathcal{H}_K \subseteq \mathrm{L}^1(\mathrm{P})$ as for large enough x

$$\|\boldsymbol{s}_\mathrm{p}\|\|x\| \geq -\langle \boldsymbol{s}_\mathrm{p}(x), x\rangle \geq r_1\|x\|^{2u} - r_2 \geq A\|x\|^{2u}$$

for some $A > 0$, so $\|\boldsymbol{s}_\mathrm{p}\| \geq A\|x\|^{2u-1}$ for x large enough, so $\|x\|a(x) \leq 1/\|\boldsymbol{s}_\mathrm{p}(x)\| \leq \frac{1}{A\|x\|^{2u-1}}$ for x large enough, which implies that $\mathcal{H}_K \subseteq \mathscr{C}_b$.

R.1 Proof of Lemma 13: MMD controls subset MMDs

By Schwartz (1964, Prop. 2), $\mathcal{H}_k \subseteq \mathcal{H}_{\tilde{k}}$ implies that there exists $c \geq 0$ such that, for all $h \in \mathcal{H}_k$, $\|h\|_{\mathcal{H}_{\tilde{k}}} \leq c\|h\|_{\mathcal{H}_k}$, so $c^{-1}\mathcal{B}_k \subseteq \mathcal{B}_{\tilde{k}}$. Hence

$$\mathrm{MMD}_k(\mathrm{Q}, \mathrm{P}) \equiv \sup_{h \in \mathcal{B}_k \,:\, h_+ \in \mathrm{L}^1(\mathrm{Q})} |\mathrm{Q}h - \mathrm{P}h| \leq c\,\mathrm{MMD}_{\tilde{k}}(\mathrm{Q}, \mathrm{P}).$$

Since $\mathcal{H}_k \subseteq \mathcal{H}_{\tilde{k}}$, the P-separation and tightness results are immediate.

Appendix S. Proof of Theorem 13: Decaying P-centered kernels fail to control P-convergence

We will use the following result based on a construction from Simon-Gabriel et al. (2023, Section 5).

Theorem 16 (Vanishing mean-zero kernels fail to control P-convergence)**.** *Suppose that $\mathcal{X}$ is locally compact but not compact. If $\mathcal{H}_k \subseteq \mathscr{C}_0$ and k maps $\mathrm{P} \in \mathcal{P}$ to $0 \in \mathcal{H}_k$, i.e., $\Phi_k(\mathrm{P}) = 0$, then k cannot control weak convergence to $\mathrm{P} \in \mathcal{P}$.*

Proof Since P is a regular measure, we can find a compact set $C \subseteq \mathcal{X}$ for which $\mathrm{P}(C) \geq 1/2$. By Simon-Gabriel et al. (2023, Lemma 9 and 10), we can find an open set U and compact set C' such that $C \subseteq U \subseteq C'$ and sequence of probability measures $(\mathrm{Q}_n)_n$ such that $\|\mathrm{Q}_n\|_k \to 0$ and $\mathrm{Q}_n(C') = 0$ for all n. Then $\|\mathrm{Q}_n - \mathrm{P}\|_k = \|\mathrm{Q}_n\|_k \to 0$, so Q_n converges to P in maximum mean discrepancy but not in weak convergence since $\mathrm{P}(U) \geq \mathrm{P}(C) > \mathrm{Q}_n(U) = 0$.

■

Now note that $\Phi_k(\mathrm{P}) = 0$ implies that every function in $\mathcal{H}_k$ has vanishing P-integral. Given any RKHS $\mathcal{H}_k \subseteq \mathrm{L}^1(\mathrm{P})$, we can construct a new RKHS whose functions have zero

expectation under P and has the same MMD between embeddable measures, as we now show by simply applying the projection operator $\Pi_{\mathrm{P}}(h) = h - \mathrm{P}h$. Since

$$|h(x) - \mathrm{P}h| = |\langle h\,,\, k_x\rangle_k - \langle \Phi_k(\mathrm{P})\,,\, h\rangle_k| \leq (\|k_x\|_k + \|\Phi_k(\mathrm{P})\|_k)\|h\|_k,$$

(Carmeli et al., 2006, Prop. 2.4) implies $\Pi_{\mathrm{P}}(\mathcal{H}_k)$ is a RKHS with kernel (using the fact $\xi^*_{\mathrm{P}}(x)(h) = \Pi_{\mathrm{P}}(h)(x)$ where $\xi^*_{\mathrm{P}}(x) \equiv \delta_x - \mathrm{P}$)

$$k^{\mathrm{P}}(x,y) = \langle \Phi_k(\delta_x - \mathrm{P})\,,\, \Phi_k(\delta_y - \mathrm{P})\rangle_k\,.$$

Thus, the elements of $\mathcal{H}_{k^{\mathrm{P}}}$ have the form $h - \mathrm{P}h$ for some $h \in \mathcal{H}_k$, and hence $\mathrm{P}(\mathcal{H}_{k^{\mathrm{P}}}) = \{0\}$.

Importantly, k^{P} and k generate the same MMD. First let us show this for embeddable measures: since for any finite measure μ with $\int \mu = 0$ that embeds into $\mathcal{H}_k$, we have using Lemma 2 and $\mu \circ \Pi_{\mathrm{P}}|_{\mathcal{H}_k} = \mu|_{\mathcal{H}_k}$ (from $\int \mu = 0$) that

$$\|\mu\|_{k^{\mathrm{P}}} = \|\mu \circ \Pi_{\mathrm{P}}\|_k = \|\mu\|_k.$$

Hence for any two embeddable probability measures Q, P we have $\mathrm{MMD}_{k^{\mathrm{P}}}(\mathrm{Q},\mathrm{P}) = \mathrm{MMD}_k(\mathrm{Q},\mathrm{P})$. In general, for any $\mathrm{Q} \in \mathcal{P}$, note that $h_+ \in \mathrm{L}^1(\mathrm{Q})$ iff $(\Pi_{\mathrm{P}}(h))_+ \in \mathrm{L}^1(\mathrm{Q})$. Moreover $\mathcal{B}_{k^{\mathrm{P}}} = \Pi_{\mathrm{P}}(\mathcal{B}_k)$ by Lemma 5. Thus, writing $S_k(\mathrm{Q}) \equiv \{h \in \mathcal{B}_k : h_+ \in \mathrm{L}^1(\mathrm{Q})\}$, we have $S_{k^{\mathrm{P}}}(\mathrm{Q}) = \Pi_{\mathrm{P}}(S_k(\mathrm{Q}))$

$$\begin{aligned}\mathrm{MMD}_{k^{\mathrm{P}}}(\mathrm{Q},\mathrm{P}) &= \sup_{f \in S_{k^{\mathrm{P}}}(\mathrm{Q})} |\mathrm{Q}(f) - \mathrm{P}(f)| = \sup_{f \in \Pi_{\mathrm{P}} S_k(\mathrm{Q})} |\mathrm{Q}(f) - \mathrm{P}(f)| \\ &= \sup_{h \in S_k(\mathrm{Q})} |\mathrm{Q}(\Pi_{\mathrm{P}} h) - \mathrm{P}(\Pi_{\mathrm{P}} h)| = \sup_{h \in S_k(\mathrm{Q})} |\mathrm{Q}h - \mathrm{P}h| = \mathrm{MMD}_k(\mathrm{Q},\mathrm{P}).\end{aligned}$$

Combining with Theorem 16 we obtain the other advertised result.

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows - In Metric Spaces and in the Space of Probability Measures*. Birkhäuser Verlag, Springer, 2005.

Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, et al. Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 1950.

Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum Stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pages 12964–12976, 2019.

Alessandro Barp, So Takao, Michael Betancourt, Alexis Arnaudon, and Mark Girolami. A unifying and canonical description of measure-preserving diffusions. *arXiv:2105.02845*, 2021.

Alessandro Barp, Lancelot Da Costa, Guilherme França, Karl Friston, Mark Girolami, Michael I Jordan, and Grigorios A Pavliotis. Geometric methods for sampling, optimization, inference, and adaptive agents. In *Handbook of Statistics*, volume 46, pages 21–78. Elsevier, 2022a.

Alessandro Barp, Chris J Oates, Emilio Porcu, and Mark Girolami. A Riemann–Stein kernel method. *Bernoulli*, 2022b.

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

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

Salomon Bochner. *Vorlesungen über Fouriersche integrale*, volume 12. Akademische Verlagsgesellschaft mbh, 1932.

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

Creighton R. Buck. Bounded continuous functions on a locally compact space. *Michigan Mathematical Journal*, 1958.

Jorge Buescu, AC Paixao, F Garcia, and I Lourtie. Positive-definiteness, integral equations and fourier transforms. *The Journal of Integral Equations and Applications*, pages 33–52, 2004.

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 2006.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 2010.

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. *arXiv:1905.03673*, 2019.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, 2020.

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

John Bligh Conway. *The Strict Topology and Compactness in the Space of Measures*. Louisiana State University and Agricultural & Mechanical College, 1965.

JLB Cooper. Positive definite functions of a real variable. *Proceedings of the London Mathematical Society*, 3(1):53–66, 1960.

Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and François-Xavier Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *AISTATS*, 2022.

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

Stewart N. Ethier and Thomas G. Kurtz. *Markov processes: Characterization and convergence*, volume 282. John Wiley & Sons, 2009.

Matthew A Fisher, Chris Oates, et al. Gradient-free kernel Stein discrepancy. *arXiv:2207.02636*, 2022.

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 Stein's method. In *NeurIPS*, 2015.

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

Jackson Gorham, Andrew B. Duncan, Sebastian J. Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 2019.

Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic Stein discrepancies. *Advances in Neural Information Processing Systems*, 2020.

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

Jun Han and Qiang Liu. Stein variational gradient descent without gradient. In *ICML*, 2018.

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

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

Steven G. Johnson. Saddle-point integration of C∞ "bump" functions. *arXiv:1508.04376*, 2018.

Heishiro Kanagawa, Alessandro Barp, Arthur Gretton, and Lester Mackey. Controlling moments with kernel stein discrepancies. *arXiv preprint arXiv:2211.05408*, 2022.

Lucien LeCam. Convergence in distribution of stochastic processes. *University of California Publications in Statistics*, 1957.

Chang Liu and Jun Zhu. Riemannian Stein variational gradient descent for Bayesian inference. In *AAAI*, 2018.

Qiang Liu. Stein variational gradient descent as gradient flow. In *NeurIPS*, 2017.

Qiang Liu and Jason 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.

Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, Chris Oates, et al. Robust generalised Bayesian inference for intractable likelihoods. *arXiv:2104.07359*, 2021.

Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, and Chris Oates. Generalised Bayesian inference for discrete intractable likelihood. *arXiv:2206.08420*, 2022.

Mario Micheli and Joan Alexis Glaunes. Matrix-valued kernels for shape deformation analysis. *arXiv:1308.5739*, 2013.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 1997.

Kazimierz Musiał. *Handbook of Measure Theory - Chapter 12 - Pettis Integral*. Elsevier, 2002.

Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *arXiv:1410.2392*, 2014.

Chris J. Oates, Jon Cockayne, François-Xavier Briol, and Mark Girolami. Convergence rates for a class of estimators based on Stein's method. *Bernoulli*, 2019.

Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

Tomos Phillips. *On positive and conditionally negative definite functions with a singularity at zero, and their applications in potential theory*. PhD thesis, Cardiff University, 2018.

Stefano Pigola and Alberto G. Setti. Global divergence theorems in nonlinear PDEs and geometry. *Ensaios Matemáticos*, 2014.

Marina Riabiz, Wilson Ye Chen, Jon Cockayne, Pawel Swietach, Steven A. Niederer, Lester Mackey, and Chris. J. Oates. Optimal thinning of MCMC output. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2022.

Laurent Schwartz. Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d'analyse mathématique*, 1964.

Laurent Schwartz. *Théorie des Distributions*. Hermann, 1978.

Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *JMLR*, 2018.

Carl-Johann Simon-Gabriel, Alessandro Barp, Bernhard Schölkopf, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *Journal of Machine Learning Research*, 24(184):1–20, 2023.

Bharath K. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 2016.

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

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *JMLR*, 2011.

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

James Stewart. Positive definite functions and generalizations, an historical survey. *The Rocky Mountain Journal of Mathematics*, 6(3):409–434, 1976.

Zhuo Sun, Alessandro Barp, and François-Xavier Briol. Vector-valued control variates. In *International Conference on Machine Learning*, pages 32819–32846. PMLR, 2023.

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

Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004.

George Wynne, Mikołaj Kasprzak, and Andrew B Duncan. A spectral representation of kernel Stein discrepancy with application to goodness-of-fit tests for measures on infinite dimensional Hilbert spaces. *arXiv:2206.04552*, 2022.

Haizhang Zhang and Liang Zhao. On the inclusion relation of reproducing kernel Hilbert spaces. *Analysis and Applications*, 2013.

Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, 220(1-2):456–463, 2008.