Bellman Contraction under MMD: A Unified Framework

Baptiste Debes
ESAT-PSI
KU Leuven
baptiste.debes@kuleuven.be

Tinne Tuytelaars
ESAT-PSI
KU Leuven
tinne.tuytelaars@kuleuven.be

Abstract

We introduce a mathematical framework for studying contraction of the distributional Bellman operator under Maximum Mean Discrepancy (MMD). We show that contraction reduces to conditional positive definiteness of an associated kernel constructed from the base kernel, which we call the Bellman difference kernel. Focusing on radial kernels, we use Schoenberg's theorem to connect this condition to Bernstein functions. This gives contraction criteria that hold in any dimension and can be verified analytically or tested numerically. Our framework unifies known results for the Gaussian RBF, power-distance, and multiquadric kernels. A numerical search identifies 11 new kernels that induce contraction. We also propose *chaining*, a general construction that produces kernels inducing contraction with any prescribed factor. To reconcile theory with practice, we show that contraction on a set only requires conditional positive definiteness on that set. This explains why the RBF kernel can remain a valid choice when rewards are bounded, despite its lack of global contraction. Finally, we identify a fundamental tradeoff between contraction rate and discriminative power, where kernels that contract faster tend to discriminate less between distributions. Code is available at `https://github.com/BaptisteDebes/BellmanContractionMMD`

1 Introduction

Distributional reinforcement learning (DRL) models the distribution over returns instead of their expectations [3, 4]. This in turn makes it possible to account for irreducible uncertainty in transitions and rewards exhibited by some environments [23]. It also enables risk-aware policy improvement [10, 32]. Distributional RL has shown substantial performance gains on a wide variety of tasks [2, 10, 18] and has been motivated theoretically [21, 27, 28]. A practical instantiation of DRL depends on both how the return distribution is parametrized and how discrepancies between probability distributions are measured [28]. A natural and flexible choice is to represent the return distribution as a set of equally weighted particles [10, 25, 33]. This representation enables the use of discrepancies computed from samples, such as Maximum Mean Discrepancy (MMD) [16, 19, 25, 38], Sinkhorn divergence [14, 34], or sliced divergences [13].

Maximum Mean Discrepancy depends on a kernel that controls which aspects of the distributions are compared and how sensitively. A central question is then whether the *distributional Bellman operator* [3] is contractive under the MMD induced by a given kernel, which governs whether the resulting learning objective is principled [3, 11]. Despite the flexibility afforded by kernel choice, existing understanding of which kernels induce contraction remains limited. In particular, it provides little guidance for identifying or constructing new ones. This in turn leaves practitioners with few theoretically motivated options.

The first contribution of this work is to connect contraction under MMD with a definiteness property of an associated kernel, namely conditional positive definiteness. For the important class of radial

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

kernels, we leverage Schoenberg's theorem [30], which links radial kernels to Bernstein functions [17, 29]. This perspective allows us to show that contraction can be verified analytically. This unifies known results from [12, 19, 25] under a single framework. More broadly, this framework provides practical recipes for identifying kernels that induce contraction and for constructing new ones, which we demonstrate in this work. Notably, all results hold in arbitrary dimension, directly covering the multivariate setting [13, 37, 38].

Secondly, a gap remains between theory and practice, as some kernels such as the RBF are known not to induce contraction in general, yet still perform among the best in practice [25]. We reconcile theory and practice by showing numerically that such kernels can in fact induce contraction when the return laws are supported on a bounded set. In reinforcement learning, this assumption arises naturally when rewards are bounded, since discounted returns are then bounded as well.

Finally, we identify a tradeoff between the contraction induced by a kernel and its ability to discriminate between distributions. In general, kernels with faster contraction tend to have weaker discriminative power, so these two aspects cannot be optimized independently. We study this tradeoff empirically by measuring the power of kernel two-sample tests.

2 Background and related work

We consider a discounted Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $P(\cdot \mid s, a)$ the transition kernel, $R(s, a)$ the random reward, and $\gamma \in (0, 1)$ the discount factor. Rewards may be multivariate, in which case returns take values in $\mathbb{R}^d$. For a fixed policy π, we define the random return as follows, with its expectation yielding the usual state-action value function:

$$Z^\pi(s, a) := \sum_{t \geq 0} \gamma^t R_t, \qquad Q^\pi(s, a) := \mathbb{E}[Z^\pi(s, a)],$$

where $S_0 = s$, $A_0 = a$, $R_t = R(S_t, A_t)$, $S_{t+1} \sim P(\cdot \mid S_t, A_t)$ and $A_{t+1} \sim \pi(\cdot \mid S_{t+1})$. We represent return laws by maps $\eta : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R}^d)$. For a fixed policy π, the distributional Bellman operator is defined, for $S' \sim P(\cdot \mid s, a)$, $A' \sim \pi(\cdot \mid S')$, and $X' \sim \eta(S', A')$, by

$$(\mathcal{T}^\pi \eta)(s, a) := \mathrm{Law}\big(R(s, a) + \gamma X'\big). \tag{1}$$

Policy evaluation. This work focuses on distributional policy evaluation for a fixed policy π, where one seeks to solve or approximate the fixed point equation $\eta^\pi = \mathcal{T}^\pi \eta^\pi$. Our goal is to study the contraction properties of the corresponding distributional Bellman operator $\mathcal{T}^\pi$, rather than policy improvement or control.

Maximum Mean Discrepancy. A kernel is a symmetric function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. For laws $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, the squared Maximum Mean Discrepancy (MMD) induced by k is defined by [16]

$$\mathbf{MMD}_k^2(\mu, \nu) = \mathbb{E}_{x, x' \sim \mu}[k(x, x')] + \mathbb{E}_{y, y' \sim \nu}[k(y, y')] - 2\,\mathbb{E}_{x \sim \mu,\, y \sim \nu}[k(x, y)]. \tag{2}$$

Some background on kernels and MMD is provided in Appendix A, including the biased estimator of $\mathbf{MMD}_k^2$ [16], which is the one commonly used in distributional RL [19, 25].

Contraction. Return laws are conditional distributions, so to compare two of them we use the supremum MMD over state-action pairs [3, 25], defined by

$$\overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2) := \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathbf{MMD}_k^2\big(\eta_1(s, a), \eta_2(s, a)\big). \tag{3}$$

For $\tau \in [0, 1)$, we say that $\mathcal{T}^\pi$ is a τ-contraction in $\overline{\mathbf{MMD}}_k^2$ if

$$\overline{\mathbf{MMD}}_k^2(\mathcal{T}^\pi \eta_1, \mathcal{T}^\pi \eta_2) \leq \tau\, \overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2) \qquad \forall\, \eta_1, \eta_2.$$

We call any such τ a *contraction factor* of $\mathcal{T}^\pi$ and denote by $\tau_k^\star(\mathcal{T}^\pi)$ the smallest one valid for every MDP with discount γ and every fixed policy. Throughout this work, all contraction factors refer to $\overline{\mathbf{MMD}}_k^2$. The corresponding smallest contraction factor in $\overline{\mathbf{MMD}}_k$ is the square root $\sqrt{\tau_k^\star(\mathcal{T}^\pi)}$.

In policy evaluation, $\tau_k^\star(\mathcal{T}^\pi) < 1$ guarantees that the distributional Bellman equation has a unique solution η^π. Repeated Bellman updates then converge to that solution in the supremum MMD at

Kernel	Formula	Contraction in $\overline{\mathbf{MMD}}_k^2$
Power-distance [25]	$k_\alpha(x,y) = -\|x-y\|^\alpha, \quad 0 < \alpha \leq 2$	Yes, factor γ^α
Gaussian RBF [25]	$k_{\mathrm{rbf}}(x,y) = \exp(-h^2\|x-y\|^2)$	No in general
Multiquadric [12, 19]	$k_{\mathrm{mq},h}(x,y) = -\sqrt{1+h^2\|x-y\|^2}$	Yes, factor γ

Table 1: Known contraction results for specific kernels in distributional RL with MMD prior to this work. The multiquadric corresponds to the case $\beta = 1/2$ of the generalized multiquadric $k_{\beta,h}(x,y) = -(1+h^2\|x-y\|^2)^\beta$.

a geometric rate, $\overline{\mathbf{MMD}}_k^2\big((\mathcal{T}^\pi)^n\eta, \eta^\pi\big) \leq \tau_k^\star(\mathcal{T}^\pi)^n\, \overline{\mathbf{MMD}}_k^2(\eta, \eta^\pi)$ for every η and every $n \geq 1$. More broadly, contraction is the standard way to show that the distributional Bellman equation is well posed under a given discrepancy [3, 4, 28].

Related work and known results. Existing work on distributional RL with MMD has established contraction results for a small set of kernels [19, 25]. In the univariate setting, sufficient conditions for Bellman contraction were established for kernels that are both shift invariant and scale sensitive of order $\alpha > 0$, meaning that for all x, y, c in the underlying space [25],

$$k(x+c, y+c) = k(x,y), \qquad k(cx, cy) = |c|^\alpha k(x,y). \tag{4}$$

Table 1 summarizes the main contraction results known for specific kernels prior to this work. Distributional RL with MMD was subsequently extended to multivariate rewards [37, 38]. The earlier of the two minimized MMD but proved contraction of the multivariate distributional Bellman operator only in the Wasserstein metric [38], leaving contraction in MMD open. A later work laying the foundations of multivariate distributional RL proved contraction in MMD, but only for kernels satisfying the conditions of Equation 4 [37]. Taken together, these works identify a small set of kernels for which contraction can be proved or disproved, but do not provide a general criterion that unifies these results. Providing such a framework is among the contributions of this work.

3 Contraction in Maximum Mean Discrepancy

In this section, we establish our first main result, showing that contraction of the distributional Bellman operator in $\overline{\mathbf{MMD}}_k^2$ reduces to a conditional positive definiteness condition on a kernel derived from k. We begin by recalling the notion of conditional positive definiteness.

Definition 3.1. Let $k : X \times X \to \mathbb{R}$ be symmetric. We say that k is *conditionally positive definite* (CPD), or *conditionally negative definite* (CND), if for every $n \geq 1$, every $x_1, \ldots, x_n \in X$, and every $w_1, \ldots, w_n \in \mathbb{R}$ with $\sum_{i=1}^n w_i = 0$, one has

$$\sum_{i,j=1}^n w_i w_j\, k(x_i, x_j) \geq 0 \qquad (\text{resp. } \leq 0).$$

Theorem 3.1 below makes precise the reduction of Bellman contraction to a CPD condition. Conditional positive definiteness is well studied, especially for radial kernels, as we exploit in Section 4. The proof is deferred to Appendix C.

Theorem 3.1. *Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a CPD kernel. For a reward value $q \in \mathbb{R}^d$ and a factor $\tau \in [0,1)$, define the* Bellman difference kernel

$$k_{\mathrm{diff}}^{q,\gamma}(x,y) := \tau k(x,y) - k(q+\gamma x, q+\gamma y). \tag{5}$$

Then the following two statements are equivalent. (i) $\mathcal{T}^\pi$ is a τ-contraction in $\overline{\mathbf{MMD}}_k^2$ for every MDP with discount γ and every fixed policy. (ii) The kernel $k_{\mathrm{diff}}^{q,\gamma}$ is CPD for every reward value $q \in \mathbb{R}^d$. In particular, the smallest contraction factor $\tau_k^\star(\mathcal{T}^\pi)$ is the smallest τ for which (ii) holds.

The CPD condition is therefore sufficient and necessary for $\mathcal{T}^\pi$ to be a τ-contraction for every MDP with discount γ and every fixed policy. If it fails, there exist an MDP with discount γ and a fixed policy for which $\mathcal{T}^\pi$ is not a τ-contraction.

4 Contraction with radial kernels

A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called *radial* if $k(x, y) = g(\|x-y\|)$ for some function $g : [0, \infty) \to \mathbb{R}$. We focus on this class for two reasons. First, it covers the main kernels studied in distributional RL with MMD, see Table 1. Second, radial kernels admit powerful contraction results through Schoenberg's theorem [30] and the theory of Bernstein functions [29], which we now develop.

Definition 4.1. A function $\phi : [0, \infty) \to [0, \infty)$ is a *Bernstein function* if it admits the Lévy–Khintchine representation

$$\phi(t) = a + b\,t + \int_{(0,\infty)} \left(1 - e^{-st}\right) \Lambda(ds), \quad t \geq 0$$

where $a, b \geq 0$ and Λ is a positive Borel measure on $(0, \infty)$, called the Lévy measure of ϕ, with $\int_{(0,\infty)} \min\{1, s\}\, \Lambda(ds) < \infty$. It is a *complete Bernstein function* if, moreover, $\Lambda(ds) = m(s)\, ds$ for a completely monotone density $m : (0, \infty) \to [0, \infty)$. In our setting $\phi(0) = a = 0$.

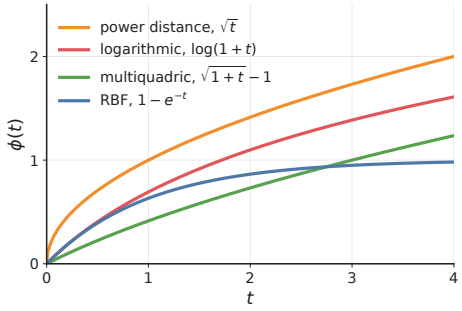

Figure 1: Examples of Bernstein functions.

We now make the bridge from the CPD condition in Theorem 3.1 to the Bernstein reduction that follows. To this end, we state Schoenberg's theorem on radial kernels [30], and use the elementary fact that CPD of a kernel k is equivalent to CND of $-k$.

Theorem 4.1 (Schoenberg [30])**.** *A kernel of the form $(x, y) \mapsto h(\|x - y\|)$, with h continuous and $h(0) = 0$, is CND on $\mathbb{R}^d$ for every $d \geq 1$ if and only if $h(r) = \phi(r^2)$ for some Bernstein function ϕ.*

Bernstein reduction. The next theorem applies Schoenberg's theorem to the Bellman difference kernel. Its proof is given in Appendix E.5.

Theorem 4.2. *Let $k(u, v) = g(\|u - v\|)$ be a radial kernel, and assume that $h(r) := g(0) - g(r)$ can be written as $h(r) = \phi(r^2)$ for a Bernstein function ϕ, which we call the* Bernstein profile *of k. For $\tau \in [0, 1)$, define the* Bellman difference profile

$$\psi_\tau(t) := \tau\phi(t) - \phi(\gamma^2 t), \qquad t \geq 0. \tag{6}$$

If ψ_τ is a Bernstein function, then $\mathcal{T}^\pi$ is a τ-contraction in $\overline{\mathrm{MMD}}_k^2$. Furthermore, this condition is necessary for τ-contraction to hold for every MDP with discount γ, every fixed policy, and every ambient dimension $d \geq 1$.

The necessity part of Theorem 4.2 concerns every dimension at once. If ψ_τ is not a Bernstein function, the Bellman difference kernel fails to be CPD in at least one dimension, but it may still be CPD in others. Theorem 3.1 then remains the right tool to check contraction in a given dimension.

Analytic reduction. Theorem 4.2 reduces Bellman contraction for radial kernels to checking whether the Bellman difference profile ψ_τ is a Bernstein function. Theorem 4.3 makes this condition explicit through the Lévy–Khintchine representation of the Bernstein profile ϕ. In the complete Bernstein case, it also yields an explicit upper bound on the smallest contraction factor. The argument is developed in Appendix E.3–E.6.

Theorem 4.3. *Adopt the setting of Theorem 4.2, and let b and Λ be the components of the Lévy–Khintchine representation of ϕ from Definition 4.1. Then the Bellman difference profile ψ_τ is a Bernstein function if and only if*

$$b(\tau - \gamma^2) \geq 0 \qquad \text{and} \qquad \tau\,\Lambda(B) \geq \Lambda(B/\gamma^2) \quad \text{for every Borel } B \subset (0, \infty),$$

where $B/\gamma^2 := \{s > 0 : \gamma^2 s \in B\}$. In particular, if these conditions hold for some $\tau \in [0, 1)$, then $\mathcal{T}^\pi$ is a τ-contraction in $\overline{\mathrm{MMD}}_k^2$. Furthermore, if ϕ is a complete Bernstein function, so that $\Lambda(ds) = m(s)\, ds$ for a density m, then the smallest contraction factor satisfies

$$\tau_k^\star(\mathcal{T}^\pi) \leq \overline{\tau}_{\mathrm{cbf}}(\phi, \gamma), \qquad \overline{\tau}_{\mathrm{cbf}}(\phi, \gamma) := \max\left\{\gamma^2\, \mathbf{1}_{\{b>0\}},\ \sup_{s>0} \frac{m(s/\gamma^2)}{\gamma^2 m(s)}\right\}.$$

Theorems 4.2 and 4.3 form the backbone of the remainder of this work, unifying known contraction results under a single criterion and providing the tools that the next sections use to identify and construct new kernels that induce contraction.

Explaining known results. Applying Theorem 4.3 to the power-distance and multiquadric kernels of [12, 19, 25] recovers their contraction results from a single analytic criterion and extends the multiquadric result from $\beta = 1/2$ to the full generalized range $0 < \beta \leq 1$. In both cases the radial profile is a complete Bernstein function, so contraction reduces to evaluating $\overline{\tau}_{\mathrm{cbf}}(\phi, \gamma)$ directly. Proofs are deferred to Propositions F.1 and F.2 in the appendix.

Proposition 4.1. *Consider the power-distance kernels, with parameter $0 < \alpha \leq 2$, and the generalized multiquadric kernels, with parameters $0 < \beta \leq 1$ and $h > 0$:*

$$k_\alpha(x, y) = -\|x - y\|^\alpha, \qquad \textit{and} \qquad k_{\beta,h}(x, y) = -\big(1 + h^2\|x - y\|^2\big)^\beta.$$

Their radial profiles are generated by the Bernstein functions

$$\phi_\alpha(t) = t^{\alpha/2}, \qquad \textit{and} \qquad \phi_{\beta,h}(t) = (1 + h^2 t)^\beta - 1.$$

Moreover, both ϕ_α and $\phi_{\beta,h}$ are complete Bernstein functions, and evaluating the bound $\overline{\tau}_{\mathrm{cbf}}(\phi, \gamma)$ from Theorem 4.3 yields

$$\overline{\tau}_{\mathrm{cbf}}(\phi_\alpha, \gamma) = \gamma^\alpha, \qquad \textit{and} \qquad \overline{\tau}_{\mathrm{cbf}}(\phi_{\beta,h}, \gamma) = \gamma^{2\beta}.$$

Hence $\mathcal{T}^\pi$ is a γ^α-contraction in $\overline{\mathbf{MMD}}_k^2$ for the power-distance kernels, and a $\gamma^{2\beta}$-contraction for the generalized multiquadric kernels.

A first obstruction: saturating profiles. Proposition 4.2 gives a first explanation for the non-contraction of the RBF kernel established in [25]. Saturating Bernstein profiles rule out contraction, and since the RBF profile shown in Figure 1 is saturating, the obstruction applies directly. The proof is given in Proposition E.3 in the appendix.

Proposition 4.2. *Adopt the setting of Theorem 4.2, and assume that the Bernstein profile ϕ is saturating, meaning that $\lim_{t\to\infty} \phi(t) < \infty$. If $\phi \not\equiv 0$, then there is no $\tau \in (0, 1)$ such that τ-contraction holds for every MDP with discount γ and every fixed policy.*

A second obstruction: finitely many atoms. The analytic criterion from Theorem 4.3 also rules out kernels whose Lévy measure has finitely many atoms, that is, finitely many Dirac masses, as stated in Proposition 4.3. This gives a second explanation for the failure of the RBF kernel. Its Bernstein profile is $\phi_{\mathrm{rbf},h}(t) = 1 - \exp(-h^2 t)$, whose Lévy measure $\Lambda = \delta_{h^2}$ has a single atom at h^2.

Proposition 4.3. *Adopt the setting of Theorem 4.3, and let Λ be the Lévy measure of ϕ. If Λ has at least one but finitely many atoms, then there is no $\tau \in (0, 1)$ such that τ-contraction holds for every MDP with discount γ, every fixed policy, and every ambient dimension $d \geq 1$.*

The same argument also shows that no finite sum of RBF kernels, such as those used in [25, 38], induces contraction. Later, we will show that this obstruction disappears for suitable infinite sums of RBF kernels, which can in fact induce contraction. We refer to Lemma E.3 and Theorem E.4 in the appendix for the proof of Proposition 4.3, and to Proposition F.3 for the case of the RBF kernel.

5 Contraction on a bounded support

The contraction theory developed in Sections 3 and 4 requires the Bellman difference kernel to be CPD on the whole ambient space $\mathbb{R}^d$, which can be unnecessarily restrictive in practice. The RBF kernel, which fails this global condition yet performs well empirically [25, 38], raises the question of whether contraction can hold under weaker conditions. In this section we show that under bounded rewards, CPD need only hold on the bounded region where return laws actually live.

In practical reinforcement learning settings, rewards can often be assumed bounded. In that case, the following proposition gives a bounded set that contains all discounted returns, and shows that return laws supported on this set remain supported on it after a Bellman update. The proof is given in Appendix C.2.

Proposition 5.1. *Assume the rewards are bounded by $r_{\max}$, and let E be the ball of radius M centered at the origin, $E := \{x \in \mathbb{R}^d : \|x\| \le M\}$, with $M \ge r_{\max}/(1-\gamma)$. Then every return law $\eta^\pi(s,a)$ is supported on E. Moreover, if $\eta(s,a)$ is supported on E for every (s,a), then $(\mathcal{T}^\pi\eta)(s,a)$ is also supported on E for every (s,a).*

This motivates a notion of contraction restricted to return laws supported on E, for which the difference kernel only needs to be CPD on E. The next theorem gives the corresponding criterion for Bellman contraction.

Theorem 5.1. *Adopt the setting of Proposition 5.1, so that the return laws are supported on $E := \{x \in \mathbb{R}^d : \|x\| \le M\}$ with $M \ge r_{\max}/(1-\gamma)$. Let k be a CPD kernel. Fix $\tau \in [0,1)$. For a reward value q, the Bellman difference kernel $k^{q,\gamma}_{\mathrm{diff}}$ of Equation 5 is CPD on E if for every $n \ge 1$, every $x_1, \dots, x_n \in E$,*

$$\sum_{i,j=1}^{n} w_i w_j \, k^{q,\gamma}_{\mathrm{diff}}(x_i, x_j) \ge 0 \qquad \forall\, w_1, \dots, w_n \in \mathbb{R} \text{ with } \sum_{i=1}^{n} w_i = 0.$$

Then the following two statements are equivalent. (i) $\mathcal{T}^\pi$ is a τ-contraction in $\overline{\mathbf{MMD}}^2_k$ when restricted to return laws supported on E, for every fixed policy and every MDP with discount γ whose rewards are bounded by $r_{\max}$. (ii) The kernel $k^{q,\gamma}_{\mathrm{diff}}$ is CPD on E for every reward value q with $\|q\| \le r_{\max}$.

Under this weaker notion, kernels that fail the global condition may still be contractive on bounded sets. The proof of Theorem 5.1 is given in Appendix C.2. We now turn to testing the CPD condition on the ball E numerically, using the RBF kernel as a practical example.

Reduction to a generalized eigenvalue problem. Assuming bounded rewards, return laws are supported on a fixed bounded set E, so contraction reduces to checking that the associated Bellman difference kernel is CPD on E rather than on all of $\mathbb{R}^d$. While this condition cannot be verified numerically on all of E, we can test it on finite subsets of E. As shown in Lemma C.5, CPD on E implies CPD on every finite subset of E, so if the CPD condition fails on some finite set $S \subseteq E$, contraction on E cannot hold. Moreover, a larger finite subset gives a stricter test, because the condition on a set implies the condition on each of its subsets. On a finite set of points, computing the smallest τ for which the CPD condition holds reduces to a generalized eigenvalue problem [26], which can be solved efficiently. The next proposition makes this reduction explicit. Further background and derivation are given in Appendix D.

Proposition 5.2. *Adopt the setting of Theorem 5.1, and assume that the kernel k is radial. Let $S = \{z_1, \dots, z_N\} \subseteq E$, let $Q \in \mathbb{R}^{N \times (N-1)}$ be a matrix whose columns form an orthonormal basis of $\mathbf{1}_N^\perp := \{w \in \mathbb{R}^N : \mathbf{1}_N^\top w = 0\}$, and define*

$$K_{ij} := k(z_i, z_j), \qquad (K_\gamma)_{ij} := k(\gamma z_i, \gamma z_j), \qquad A := Q^\top K Q, \qquad B := Q^\top K_\gamma Q.$$

Then the following two statements are equivalent. (i) The Bellman difference kernel $k^{q,\gamma}_{\mathrm{diff}}$ is CPD on S. (ii) The matrix $\tau A - B$ is positive semidefinite. If A is strictly positive definite, the smallest value of τ for which (ii) holds is the largest generalized eigenvalue $\lambda_{\max}(B, A)$ of the pair (B, A). We call it the contraction threshold of S:

$$\tau^\star_S(\gamma) = \lambda_{\max}(B, A) = \sup_{v \ne 0} \frac{v^\top B v}{v^\top A v}.$$

In particular, if $\tau < \tau^\star_S(\gamma)$ for some $S \subseteq E$, then, by Theorem 5.1, Bellman τ-contraction on return laws supported on E cannot hold.

Numerical experiments and contraction thresholds. We now use Proposition 5.2 to study contraction on a bounded set numerically. For each trial, we draw a set $S = \{z_1, \dots, z_N\}$ of N points uniformly at random in the ball $B(0, R)$, in dimension 1, and compute $\tau^\star_S(\gamma)$ by solving the generalized eigenvalue problem via Algorithm 1. Figure 2 reports the observed contraction threshold as a function of the discount γ and the ratio R/ℓ. For the RBF kernel, the threshold stays below 1 when R/ℓ is small enough, which suggests that the kernel induces contraction on a ball of radius R even though it does not in general. This is numerical evidence on finite sets of points, not a proof on the whole ball. Finding a condition on R/ℓ that guarantees contraction remains an open question. To show that the same method applies beyond the RBF case, we also run it on the logarithmic kernel $k_\ell(x, y) = -\log\big(1 + \|x - y\|^2/\ell^2\big)$, which is non-expansive by Proposition G.1(iii) and can become contractive on bounded supports. Further details are given in Appendix D.

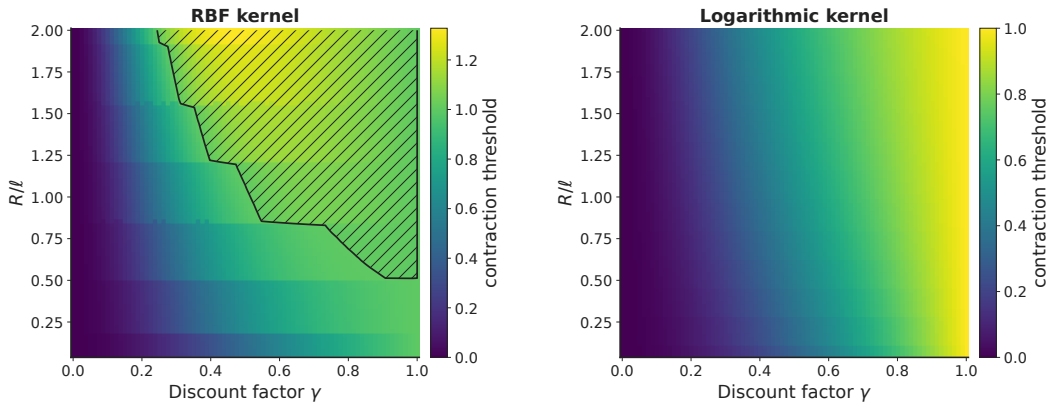

Figure 2: Contraction thresholds on a ball of radius R, for the RBF kernel $\exp(-\|x-y\|^2/\ell^2)$ (left) and the logarithmic kernel $-\log(1+\|x-y\|^2/\ell^2)$ (right). For each discount γ and each ratio R/ℓ, the color shows the largest contraction threshold $\tau_S^\star(\gamma)$ of Proposition 5.2 over 10 sets S of points drawn at random in $B(0,R)$, in dimension 1. In the hatched region, above the black line, this value exceeds 1, so contraction on the ball cannot hold. Where it stays below 1, the test does not rule out contraction. The RBF kernel passes the test when R/ℓ is small, even though it does not induce contraction on all of $\mathbb{R}^d$.

6 New kernels inducing contraction

The criterion of Theorem 4.3 opens the door to a systematic search for kernels that induce contraction, beyond the two families previously known. Since the kernel determines which aspects of the distributions the MMD objective is sensitive to [31], a broader catalogue allows practitioners to make that choice deliberately rather than by default. In this section we develop two complementary approaches.

Searching over complete Bernstein functions. In the complete Bernstein case, Theorem 4.3 reduces the contraction test to a one-dimensional supremum that depends only on the Lévy density $m(s)$. Standard references such as Schilling et al. [29] compile many complete Bernstein functions together with their Lévy densities. We extracted 43 entries from this reference, selected as described in Appendix G.1 and listed in Tables 3–6. Because the resulting list is too large to handle analytically case by case, we estimate the supremum numerically with the one-dimensional search summarized in Algorithm 2. Appendix G.1 validates this search against the exact contraction factors of Proposition 4.1. This search identifies 11 additional kernel families that induce contraction, collected in Table 8. Figure 8 shows their estimated contraction factors. Some of these families are parametric, and contraction holds only on part of the admissible parameter range, as Figure 9 in Appendix G.1.2 shows.

Chaining: a construction with prescribed contraction factor. In Section 4, we showed that the RBF kernel does not induce contraction. Surprisingly, a carefully constructed infinite sum of RBF kernels can. We call this construction *chaining* and show that given a discount factor γ, it can selectively induce contraction with any prescribed factor in $(\gamma^2, 1)$.

Theorem 6.1. *Let $\tau \in (\gamma^2, 1)$ be a target contraction factor, and fix a scale $a > 0$ and a weight $c > 0$. For a lengthscale $\ell > 0$, let $k_{\mathrm{rbf},\ell}(x,y) = \exp(-\|x-y\|^2/\ell^2)$ be the RBF kernel with lengthscale ℓ, and define the* chained RBF kernel

$$k_{\mathrm{chain}}(x,y) := \sum_{n=0}^{\infty} c\,\tau^{-n}\big(k_{\mathrm{rbf},\ell_n}(x,y) - 1\big), \qquad \ell_n := a^{-1/2}\gamma^{-n}.$$

Then $\mathcal{T}^\pi$ is a τ-contraction in $\overline{\mathrm{MMD}}^2_{k_{\mathrm{chain}}}$ for every MDP with discount γ and every fixed policy. Moreover, no smaller factor gives a contraction for every MDP with discount γ, every fixed policy and every ambient dimension $d \geq 1$.

Its proof combines Theorem 3.1 and Proposition G.3 in the appendix. Proposition G.2 applies the same construction to kernels other than the RBF. Figures 12–13 plot the Bernstein function and

the Lévy measure of the chained RBF kernel for three values of τ. The construction combines RBF kernels whose lengthscales and weights both grow. Since $\gamma < 1$, the lengthscales grow, $\ell_{n+1} = \gamma^{-1}\ell_n > \ell_n$, so each new term is sensitive to larger distances between returns. Since $\tau < 1$, the weights grow as well, $c\tau^{-(n+1)} = \tau^{-1}c\tau^{-n} > c\tau^{-n}$, so each new term counts more. In practice, the sum must be truncated, and a few hundred terms are enough to make the truncation error small, as shown numerically in Appendix G.2.

The chaining construction produces a kernel that induces contraction with any prescribed factor, making it a principled new option for distributional RL with MMD. However, a faster contraction rate does not automatically make a kernel more desirable, as we discuss in Section 7.

7 A tradeoff perspective

The preceding sections provided tools for characterizing the smallest contraction factor of a given kernel across dimensions. One might therefore be tempted to compare kernels through this quantity alone and to conclude that the kernel with the smallest contraction factor is the most desirable. This conclusion would be misleading. Changing the kernel also changes the discrepancy itself, so a smaller contraction factor need not translate into better practical behavior when the induced $\mathbf{MMD}$ is less sensitive to distributional differences that matter in practice. This section makes that tension precise and quantifies it empirically through the lens of statistical testing. Within the family of kernels induced by Bernstein functions, we show that the kernels attaining the smallest contraction factors are precisely those whose induced $\mathbf{MMD}$ is least informative. We begin by recalling what it means for $\mathbf{MMD}_k$ to be a metric.

Definition 7.1. A kernel k is said to induce a metric between probability distributions if $\mathbf{MMD}_k(P, Q) = 0$ implies $P = Q$ for all $P, Q \in \mathcal{P}(X)$.

For the radial kernels considered in this work, which are built from Bernstein functions, the next two propositions show that these kernels induce a metric except in one case, where the resulting discrepancy reduces to mean comparison. The proofs are given in Appendix H and Appendix I.

Proposition 7.1. *Let $k(x, y) = C - \phi(\|x - y\|^2)$, where ϕ is a Bernstein function with Lévy measure Λ. Then $\mathbf{MMD}_k$ is a metric if and only if $\Lambda \not\equiv 0$. Equivalently, $\mathbf{MMD}_k$ fails to be a metric exactly when $\phi(t) = bt$, that is, when $k(x, y) = C - b\|x - y\|^2$.*

Proposition 7.2. *Let $k(x, y) = C - b\|x - y\|^2$ with $b \geq 0$. Then $\mathbf{MMD}_k^2(P, Q) = 2b\|\mathbb{E}_P[X] - \mathbb{E}_Q[X]\|^2$. In particular, $\mathbf{MMD}_k^2(P, Q) = 0$ whenever P and Q have the same mean.*

In this degenerate case, distributional policy evaluation reduces to comparing expected returns, as in classical Q-learning [36]. The next proposition is the first illustration of the tension discussed in this section, as it shows that the smallest possible contraction factor γ^2 is attained only in this degenerate case. Its proof is given in the appendix as Proposition H.2.

Proposition 7.3. *Let $k(x, y) = C - \phi(\|x - y\|^2)$, where ϕ is a Bernstein function with Lévy measure Λ. If $\Lambda \not\equiv 0$, that is, if $\mathbf{MMD}_k$ is a metric (Proposition 7.1), then $\tau_k^\star(\mathcal{T}^\pi) > \gamma^2$, so the kernel cannot reach the smallest possible contraction factor γ^2. Equivalently, this value can only be reached by the kernel $k(x, y) = C - b\|x - y\|^2$.*

Empirical detectability. Beyond its role as a discrepancy, MMD can be used as a test statistic to decide whether two distributions differ [7, 15, 16]. Run with each kernel, such a test offers a principled way to compare kernels in practice, through the probability that it detects a difference. In distributional policy evaluation, this means detecting a statistically significant difference between predicted and target return laws from a limited number of samples.

Definition 7.2. A test $\Delta_k(X_n, Y_m) \in \{0, 1\}$ uses an empirical MMD statistic to test $H_0 \colon P = Q$ against $H_1 \colon P \neq Q$, with $\Delta_k = 1$ meaning rejection. It has *level* α if $\sup_P \mathbb{P}_{P\times P}(\Delta_k = 1) \leq \alpha$. For $P \neq Q$, its *power* is $\mathbb{P}_{P\times Q}(\Delta_k = 1)$.

Motivated by Proposition 7.3, we make the following empirical claim: *For a fixed discount γ and a parametric kernel family k_θ, a smaller contraction factor on a ball E containing the data, $\tau^\star_{k_{\theta_1},E}(\gamma) < \tau^\star_{k_{\theta_2},E}(\gamma)$, empirically comes with reduced power to distinguish return laws.*

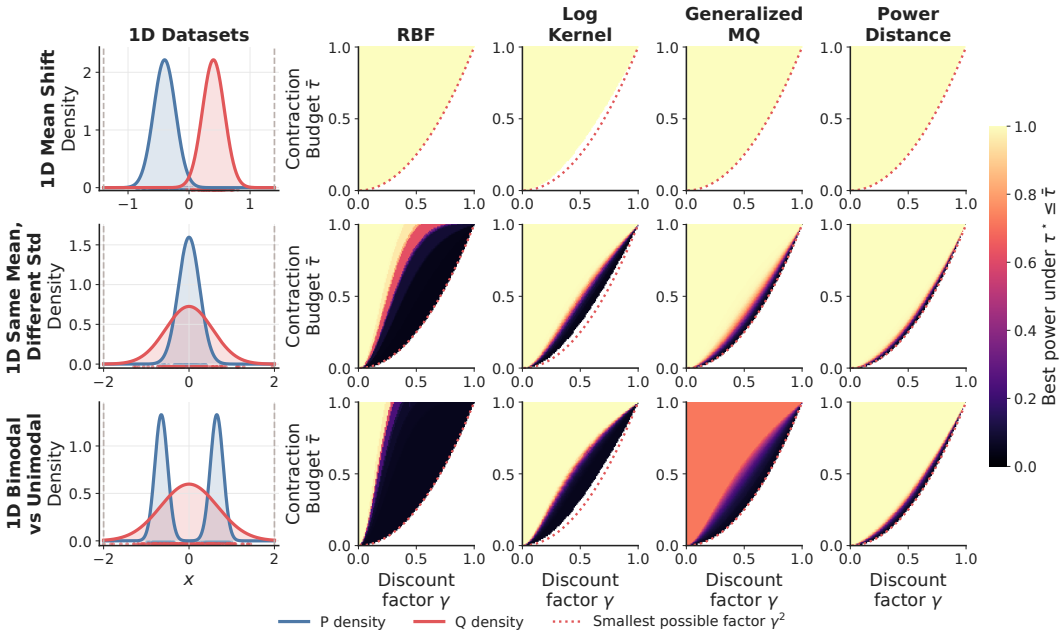

Figure 3: One-dimensional tradeoff experiments. The left column shows the clipped task distributions, with dashed lines marking the clipping radius R. The remaining columns show, for each kernel family, the best empirical test power achievable under the constraint $\tau^\star_{k_\theta,E}(\gamma) \leq \bar{\tau}$. The dotted curve is the smallest possible contraction factor γ^2.

To test this claim, we fix two distributions P and Q and measure how reliably a test of level α built from the empirical MMD can detect their difference. We use a permutation test [16] and estimate its power for each kernel configuration. The full protocol is described in Appendix H.1.

We consider several kernel families k_θ, with parameter $\theta \in \Theta$. We vary θ within Θ and, for each resulting kernel and each discount γ, use the procedure of Section 5 to estimate $\tau^\star_{k_\theta,E}(\gamma)$, the smallest contraction factor of k_θ on E. We also estimate the power of each kernel with the permutation test. For each discount γ and each bound $\bar{\tau}$ on the contraction factor, we then select, within each family, the kernel with the highest power among those with $\tau^\star_{k_\theta,E}(\gamma) \leq \bar{\tau}$. This shows how each kernel family trades contraction against empirical detectability.

Results are shown in Figure 3 for the RBF, logarithmic, generalized multiquadric, and power-distance kernels. The tradeoff depends both on the distributions being compared and on the kernel family. When the two distributions differ only in their mean, the tradeoff essentially disappears across all kernels. Comparing means is then enough, and each family has parameter regimes close to comparing means only. When they have the same mean but different spreads, or when one is bimodal and the other unimodal, the tradeoff becomes much stronger, especially for some kernel families, showing that a contraction factor close to γ^2 can come at a substantial loss of power. Details about the experiments and additional two-dimensional examples, shown in Figure 17, are provided in Appendix H. In Appendix H.3, we show that these observations carry over to dimensions up to 64, where the tradeoff stays negligible for a mean shift and grows for a checkerboard, until the dimension is too large for 100 samples per distribution to reveal the difference with any kernel.

8 Conclusion and Further Directions

This work develops a mathematical framework for studying contraction of the distributional Bellman operator under MMD. The central result is a general reduction showing that Bellman contraction is equivalent to the conditional positive definiteness of a kernel derived from the base kernel. For radial kernels, this criterion reduces further to a one-dimensional condition on the Lévy measure of the Bernstein function that generates the kernel, which can be verified analytically or tested numerically. The framework simultaneously explains and unifies existing results [12, 19, 25, 38]

and yields new ones. We recover contraction results for the power-distance and multiquadric kernels, identify numerically 11 additional kernel families that induce contraction, and introduce the chaining construction, which produces a kernel that induces contraction with any prescribed factor by aggregating RBF kernels.

We also study the RBF kernel, which does not induce contraction in general, yet performs well in practice [19, 25, 38]. When rewards are bounded, return laws remain supported on a ball, and contraction only requires the Bellman difference kernel to be conditionally positive definite on that ball. We test this condition on random finite sets of points in the ball, where it reduces to a generalized eigenvalue problem. The results suggest that the RBF kernel induces contraction on balls whose radius is small relative to the lengthscale, which may explain part of its success.

We further show that contraction rate and discriminative power are in tension. Among the radial kernels of this work, those with the smallest contraction factors are those whose MMD is least sensitive to differences beyond the mean. This tradeoff suggests that the contraction rate alone is not a sufficient criterion for kernel selection in distributional RL, and that practitioners should also consider discriminative power.

Finally, this work leaves several further directions open. We focused on contraction in distributional policy evaluation. The natural next step is to run experiments on RL environments and compare kernels on control tasks. The tradeoff between contraction rate and discriminative power depends on the distributions being compared. In distributional RL, these are the predicted and target return laws, which depend on the environment and on the policy. Therefore, measuring the tradeoff on the return laws that arise during training could guide kernel selection.

Acknowledgments and Disclosure of Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 101021347).

References

[1] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[2] Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva Tb, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*, 2018.

[3] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. Pmlr, 2017.

[4] Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

[5] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic analysis on semigroups: theory of positive definite and related functions*. Springer, 1984.

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

[7] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22 (14):e49–e57, 2006.

[8] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

[9] John B Conway. *A course in functional analysis*. Springer, 1990.

[10] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

[11] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[12] Baptiste Debes and Tinne Tuytelaars. Distributional value gradients for stochastic environments. In *International Conference on Learning Representations*, 2026.

[13] Baptiste Debes and Tinne Tuytelaars. Multivariate distributional reinforcement learning using sliced divergences. In *International Conference on Machine Learning*, volume 306, pages 23456–23526. PMLR, 2026.

[14] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 2681–2690. PMLR, 2019.

[15] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

[16] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL `http://jmlr.org/papers/v13/gretton12a.html`.

[17] Jean Carlo Guella. Generalization of the energy distance by Bernstein functions, 2021. URL `https://arxiv.org/abs/2102.00633`.

[18] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[19] Ludvig Killingberg and Helge Langseth. The multiquadric kernel for moment-matching distributional reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=z49eaB8kiH`.

[20] Erwin Kreyszig. *Introductory functional analysis with applications*. John Wiley & Sons, 1989.

[21] Clare Lyle, Marc G Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4504–4511, 2019.

[22] Charles A Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[23] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 799–806, 2010.

[24] The mpmath development team. *mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 1.4.0)*, 2026. `http://mpmath.org/`.

[25] Thanh Nguyen-Tang, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning via moment matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9144–9152, 2021.

[26] Beresford N Parlett. *The symmetric eigenvalue problem*. SIAM, 1998.

[27] Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 29–37. PMLR, 2018.

[28] Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pages 5528–5536. PMLR, 2019.

[29] René L Schilling, Renming Song, and Zoran Vondraček. *Bernstein functions: theory and applications*. Walter de Gruyter, 2009.

[30] Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[31] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

[32] Rahul Singh, Qinsheng Zhang, and Yongxin Chen. Improving robustness via risk averse distributional reinforcement learning. In *Learning for Dynamics and Control*, pages 958–968. PMLR, 2020.

[33] Rahul Singh, Keuntaek Lee, and Yongxin Chen. Sample-based distributional policy gradient. In *Learning for Dynamics and Control Conference*, pages 676–688. PMLR, 2022.

[34] Ke Sun, Yingnan Zhao, Wulong Liu, Bei Jiang, and Linglong Kong. Distributional reinforcement learning with regularized Wasserstein loss. *Advances in Neural Information Processing Systems*, 37:63184–63221, 2024.

[35] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[36] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

[37] Harley Wiltzer, Jesse Farebrother, Arthur Gretton, and Mark Rowland. Foundations of multivariate distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 37:101297–101336, 2024.

[38] Pushi Zhang, Xiaoyu Chen, Li Zhao, Wei Xiong, Tao Qin, and Tie-Yan Liu. Distributional reinforcement learning for multi-dimensional reward functions. *Advances in Neural Information Processing Systems*, 34: 1519–1529, 2021.

Appendix

A Background on Maximum Mean Discrepancy **14**

A.1 Reproducing kernel Hilbert spaces 14

A.2 Maximum mean discrepancy 15

A.3 Gram matrices and empirical MMD 16

B Basic identities for MMD^2 **18**

C Bellman contraction in Maximum Mean Discrepancy **20**

C.1 Contraction via difference kernel 20

C.2 Contraction on a bounded support 25

D Contraction on a bounded support: numerical experiments **28**

D.1 From a bounded to a discrete support condition 28

D.2 From a support condition to a matrix test 28

D.3 Experimental setup 31

D.4 Results 31

E Global contraction under radial kernels via Bernstein functions **34**

E.1 Motivations for radial kernels 34

E.2 Definitions and Schoenberg's theorem 35

E.3 Bernstein reduction of difference kernel 36

E.4 Results on measures and the measure domination condition 39

E.5 From Bernstein reduction to kernels inducing contraction 47

E.6 Contraction from complete Bernstein functions 49

E.7 Chaining as a general mechanism 50

F Explaining known results **57**

G Some new kernels inducing contraction **61**

G.1 From complete Bernstein functions 61

G.2 From geometric chaining 73

H Contraction rate and discriminability **78**

H.1 Permutation test and power estimation 80

H.2 Estimating the tradeoff between power and contraction 81

H.3 Higher dimensions 85

I Properties of radial kernels built from Bernstein functions **86**

A Background on Maximum Mean Discrepancy

A.1 Reproducing kernel Hilbert spaces

Definition A.1. A *Hilbert space* $\mathcal{H}$ is a vector space of functions equipped with an inner product $\langle u, v\rangle_{\mathcal{H}}$. The inner product measures angles between functions, and the norm it induces, $\|v\|_{\mathcal{H}} := \sqrt{\langle v, v\rangle_{\mathcal{H}}}$, measures their lengths.

For the formal definition, which also requires completeness, see [9, Ch. 1] or [20, Ch. 3].

Definition A.2 (Kernel)**.** Let X be a nonempty set. A *kernel* on X is any function $k : X \times X \to \mathbb{R}$. We say k is *symmetric* if $k(x, y) = k(y, x)$ for all $x, y \in X$.

Definition A.3 (Positive Semidefinite Kernel [6, Ch. 1])**.** Let X be a nonempty set and $k\colon X\times X \to \mathbb{R}$ be symmetric. We say k is *positive semidefinite (PSD)* if for every integer $n \geq 1$, any points $x_1, \ldots, x_n \in X$ and real coefficients $c_1, \ldots, c_n$,

$$\sum_{i,j=1}^{n} c_i\, c_j\, k(x_i, x_j) \;\geq\; 0.$$

Equivalently,

$$\iint_{X\times X} k(x, y)\, d\sigma(x)\, d\sigma(y) \geq 0$$

for every finite signed measure σ on X for which this integral is well defined.

Definition A.4 (Reproducing Kernel and RKHS [1, p. 343])**.** A function $k\colon X \times X \to \mathbb{R}$ is called a *reproducing kernel* if there exists a Hilbert space $\mathcal{H}$ of functions $f\colon X \to \mathbb{R}$ such that $k(x, \cdot) \in \mathcal{H}$ for each $x \in X$, and such that the *reproducing property* holds,

$$f(x) = \big\langle f,\, k(x, \cdot)\big\rangle_{\mathcal{H}} \qquad \text{for all } f \in \mathcal{H} \text{ and } x \in X.$$

The Hilbert space $\mathcal{H}$ is uniquely determined by k and is called the *reproducing kernel Hilbert space* (RKHS) of k.

Proposition A.1 (Reproducing kernels are the symmetric PSD kernels [1, p. 344])**.** A function $k\colon X \times X \to \mathbb{R}$ is a reproducing kernel if and only if it is symmetric and positive semidefinite.

Feature map and mean embeddings. Let $\mathcal{H}$ be the RKHS of k. Then there is a canonical *feature map*

$$\Phi : X \to \mathcal{H}, \qquad \Phi(x) = k(x, \cdot), \qquad k(x, x') = \langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}} \quad \forall x, x' \in X.$$

For any $P \in \mathcal{P}(X)$, the set of probability measures on X, the *mean embedding* of P is

$$\mu_P := \int_X \Phi(x)\, dP(x) \in \mathcal{H}.$$

Since $\mu_P \in \mathcal{H}$, we may apply the reproducing property. For any $x' \in X$,

$$\begin{aligned}
\mu_P(x') &= \langle \mu_P,\, k(x', \cdot)\rangle_{\mathcal{H}} \quad \text{(reproducing property)} \\
&= \Big\langle \int_X \Phi(x)\, dP(x),\; k(x', \cdot)\Big\rangle_{\mathcal{H}} \quad \text{(definition of } \mu_P\text{)} \\
&= \int_X \langle \Phi(x),\, k(x', \cdot)\rangle_{\mathcal{H}}\, dP(x) \quad \text{(linearity of inner product)} \\
&= \int_X k(x', x)\, dP(x) \quad \text{(reproducing property)} \\
&= \mathbb{E}_{x\sim P}\big[k(x', x)\big] \quad \text{(definition of expectation).}
\end{aligned}$$

Hence $\mu_P(x')$ can be evaluated without computing the feature map, using kernel values alone, by averaging $k(x', x)$ over points x drawn from P.

Expectation as an inner product. The RKHS turns integration under P into a single inner product against the mean embedding μ_P,

$$\begin{aligned}\mathbb{E}_{x\sim P}[f(x)] &= \int_X f(x)\,dP(x) \quad \text{(definition of expectation)}\\ &= \int_X \langle f,\, k(x,\cdot)\rangle_{\mathcal{H}}\,dP(x) \quad \text{(reproducing property)}\\ &= \Big\langle f,\, \int_X k(x,\cdot)\,dP(x)\Big\rangle_{\mathcal{H}} \quad \text{(linearity of inner product)}\\ &= \langle f,\, \mu_P\rangle_{\mathcal{H}} \quad \text{(definition of } \mu_P\text{)}.\end{aligned}$$

Distance between feature maps. Since $k(x,y) = \big\langle \Phi(x), \Phi(y)\big\rangle_{\mathcal{H}}$ is an inner product, the squared distance in $\mathcal{H}$ between feature maps is

$$\begin{aligned}\big\|\Phi(x) - \Phi(y)\big\|_{\mathcal{H}}^2 &= \big\langle \Phi(x) - \Phi(y),\, \Phi(x) - \Phi(y)\big\rangle_{\mathcal{H}}\\ &= \big\langle \Phi(x), \Phi(x)\big\rangle_{\mathcal{H}} - 2\,\big\langle \Phi(x), \Phi(y)\big\rangle_{\mathcal{H}} + \big\langle \Phi(y), \Phi(y)\big\rangle_{\mathcal{H}}\\ &= k(x,x) \;-\; 2\,k(x,y) \;+\; k(y,y).\end{aligned}$$

A.2 Maximum mean discrepancy

Definition A.5 (CPD and CSPD [5, Def. 3.1.1], [19, Def. 3.2])**.** Let X be a measurable space and let $k : X \times X \to \mathbb{R}$ be symmetric. We say that k is *conditionally positive definite* (CPD) if

$$\iint_{X\times X} k(x,y)\,d\sigma(x)\,d\sigma(y) \;\geq\; 0$$

for every finite signed measure σ on X with $\sigma(X) = 0$ for which this integral is well defined. If the inequality is strict for every nonzero such σ, then k is *conditionally strictly positive definite* (CSPD). For continuous kernels, CPD is equivalent to the condition with finite sums of Definition 3.1 [17, Lemma 3.4].

Definition A.6 (Maximum mean discrepancy)**.** Let X be a measurable space and let $k\colon X \times X \to \mathbb{R}$ be symmetric.

- If k is a reproducing (positive semidefinite) kernel with RKHS $\mathcal{H}_k$, then for $P, Q \in \mathcal{P}(X)$ the squared maximum mean discrepancy [16, Def. 2 and Lemma 4] is defined by

$$\mathbf{MMD}_k^2(P,Q) := \big\|\mu_P - \mu_Q\big\|_{\mathcal{H}_k}^2, \qquad \mu_P := \int_X k(x,\cdot)\,dP(x).$$

- More generally, if k is CPD (Definition A.5), we *define* the squared maximum mean discrepancy by the quadratic form [31, Eq. (2.6) and Thm. 22]

$$\mathbf{MMD}_k^2(P,Q) := \iint_{X\times X} k(x,y)\,d(P-Q)(x)\,d(P-Q)(y),$$

  whenever the integral is well defined. Equivalently, writing $x, x' \sim P$ and $y, y' \sim Q$, all independent, one has

$$\mathbf{MMD}_k^2(P,Q) = \mathbb{E}\,k(x,x') + \mathbb{E}\,k(y,y') - 2\,\mathbb{E}\,k(x,y),$$

  whenever the expectations are finite.

When k is a reproducing kernel, the expressions as a quadratic form and as expectations coincide with the RKHS definition [16, Lemma 6].

Proposition A.2 (Centering CPD kernels [31])**.** Let X be a measurable space and let $k : X \times X \to \mathbb{R}$ be symmetric and CPD. Fix $x_0 \in X$ and define the kernel centered at x_0,

$$k^\circ(x,y) := k(x,y) - k(x,x_0) - k(x_0,y) + k(x_0,x_0).$$

Then k° is positive semidefinite and hence admits an RKHS. Moreover, for any $P, Q \in \mathcal{P}(X)$ such that the expressions below are well defined,

$$\boxed{\mathbf{MMD}_k^2(P,Q) = \iint_{X\times X} k^\circ(x,y)\, d(P-Q)(x)\, d(P-Q)(y) = \mathbf{MMD}_{k^\circ}^2(P,Q).}$$

Definition A.7 (Characteristic kernel)**.** A reproducing kernel k is *characteristic* if

$$\mathbf{MMD}_k(P,Q) = 0 \quad \Longrightarrow \quad P = Q \qquad \text{for all } P, Q \in \mathcal{P}(X).$$

Proposition A.3 (MMD is a metric under CSPD)**.** Let X be a measurable space and let $k : X \times X \to \mathbb{R}$ be symmetric and CPD. Assume $\mathbf{MMD}_k^2(P,Q)$ is defined by Definition A.6 for the pairs considered, and set $\mathbf{MMD}_k(P,Q) := \sqrt{\mathbf{MMD}_k^2(P,Q)}$. Then $\mathbf{MMD}_k$ satisfies the metric axioms, except possibly identity of indiscernibles. More precisely, for any $P, Q, R \in \mathcal{P}(X)$ such that all the quantities below are finite, the following hold.

1. **Nonnegativity:** $\mathbf{MMD}_k(P,Q) \geq 0$.
2. **Symmetry:** $\mathbf{MMD}_k(P,Q) = \mathbf{MMD}_k(Q,P)$.
3. **Triangle inequality:** $\mathbf{MMD}_k(P,Q) \leq \mathbf{MMD}_k(P,R) + \mathbf{MMD}_k(R,Q)$.
4. **Identity of indiscernibles:**

$$\mathbf{MMD}_k(P,Q) = 0 \Rightarrow P = Q \quad \text{holds if and only if} \quad k \text{ is CSPD.}$$

Proof. When k is a reproducing kernel, $\mathbf{MMD}_k$ is a distance in the RKHS norm [16, Lemma 4], hence it satisfies nonnegativity, symmetry, and the triangle inequality. For a CPD kernel k, Proposition A.2 yields a centered positive semidefinite kernel k° such that $\mathbf{MMD}_k^2 = \mathbf{MMD}_{k^\circ}^2$, so $\mathbf{MMD}_k = \mathbf{MMD}_{k^\circ}$ is again an RKHS norm distance. Moreover, identity of indiscernibles holds exactly when k is CSPD [31]. □

A.3 Gram matrices and empirical MMD

General formulation for discrete signed measures. Fix points $z_1, \ldots, z_N \in X$ and coefficients $w \in \mathbb{R}^N$, and define the finite signed measure

$$\sigma := \sum_{i=1}^N w_i\, \delta_{z_i}.$$

Let $K \in \mathbb{R}^{N\times N}$ be the Gram matrix $K_{ij} := k(z_i, z_j)$. Then

$$\iint_{X\times X} k(x,y)\, d\sigma(x)\, d\sigma(y) = \sum_{i=1}^N \sum_{j=1}^N w_i w_j\, k(z_i, z_j) = w^\top K w.$$

In particular, if $\sum_{i=1}^N w_i = 0$ and k is CPD in the sense of Definition A.5, then $w^\top K w \geq 0$.

Two-sample empirical estimator. Given samples $x_{1:n} = (x_1, \ldots, x_n)$ from P and $y_{1:m} = (y_1, \ldots, y_m)$ from Q, define the Gram matrices

$$\begin{aligned} K_{xx} &\in \mathbb{R}^{n\times n}, & (K_{xx})_{ij} &:= k(x_i, x_j), \\ K_{yy} &\in \mathbb{R}^{m\times m}, & (K_{yy})_{ij} &:= k(y_i, y_j), \\ K_{xy} &\in \mathbb{R}^{n\times m}, & (K_{xy})_{ij} &:= k(x_i, y_j). \end{aligned}$$

Let $\mathbf{1}_n \in \mathbb{R}^n$ denote the vector of ones, and similarly $\mathbf{1}_m \in \mathbb{R}^m$.

The standard biased empirical estimator of $\mathbf{MMD}_k^2(P,Q)$ is

$$\widehat{\mathbf{MMD}}_{k,\mathrm{b}}^2(x_{1:n}, y_{1:m}) := \frac{1}{n^2}\sum_{i,j=1}^n k(x_i, x_j) + \frac{1}{m^2}\sum_{i,j=1}^m k(y_i, y_j) - \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m k(x_i, y_j),$$

which can be written with Gram matrices as

$$\widehat{\mathbf{MMD}}^2_{k,\mathrm{b}}(x_{1:n}, y_{1:m}) = \frac{1}{n^2}\,\mathbf{1}_n^\top K_{xx}\,\mathbf{1}_n + \frac{1}{m^2}\,\mathbf{1}_m^\top K_{yy}\,\mathbf{1}_m - \frac{2}{nm}\,\mathbf{1}_n^\top K_{xy}\,\mathbf{1}_m.$$

The estimator can also be written with a single Gram matrix, built from all $n+m$ points. Let $N := n+m$ and define the aggregate sample $z_{1:N} = (z_1, \dots, z_N)$ by

$$z_i := x_i \quad (1 \le i \le n), \qquad z_{n+j} := y_j \quad (1 \le j \le m).$$

Define the two-sample weight vector $\ell \in \mathbb{R}^N$ by

$$\ell := \begin{pmatrix} \frac{1}{n}\mathbf{1}_n \\ -\frac{1}{m}\mathbf{1}_m \end{pmatrix}, \qquad \text{so that } \mathbf{1}_N^\top \ell = 0,$$

and let $K \in \mathbb{R}^{N\times N}$ be the Gram matrix of the aggregate sample, $K_{ij} := k(z_i, z_j)$, which has the blocks

$$K = \begin{pmatrix} K_{xx} & K_{xy} \\ K_{xy}^\top & K_{yy} \end{pmatrix}.$$

Multiplying out $\ell^\top K \ell$ block by block gives the expression with Gram matrices above, so

$$\widehat{\mathbf{MMD}}^2_{k,\mathrm{b}}(x_{1:n}, y_{1:m}) = \ell^\top K \ell.$$

Discrete measures on a common support. Let $S = \{z_1, \dots, z_N\} \subset X$ and consider two probability measures supported on S,

$$\mu = \sum_{i=1}^N a_i\,\delta_{z_i}, \qquad \nu = \sum_{i=1}^N b_i\,\delta_{z_i}, \qquad \sum_{i=1}^N a_i = \sum_{i=1}^N b_i = 1.$$

Define $w \in \mathbb{R}^N$ by $w_i := a_i - b_i$, so that $\mathbf{1}_N^\top w = 0$, and set $\sigma := \mu - \nu = \sum_{i=1}^N w_i \delta_{z_i}$. Then the same identity for discrete measures yields

$$\mathbf{MMD}_k^2(\mu, \nu) = \iint_{X\times X} k(x,y)\,d(\mu-\nu)(x)\,d(\mu-\nu)(y) = w^\top K w, \qquad K_{ij} := k(z_i, z_j).$$

The constraint $\mathbf{1}_N^\top w = 0$ is the discrete counterpart of the condition $(\mu - \nu)(X) = 0$.

Centering and restriction to vectors whose entries sum to zero. In several arguments in this paper, we must control quadratic forms $w^\top K w$ for all vectors w whose entries sum to zero. It is convenient to encode the constraint $\mathbf{1}_N^\top w = 0$ using the centering projector

$$H_N := I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top,$$

which satisfies $H_N w = w$ whenever $\mathbf{1}_N^\top w = 0$. Consequently, for any such w,

$$w^\top K w = (H_N w)^\top K (H_N w) = w^\top (H_N K H_N) w.$$

Moreover, the condition

$$w^\top K w \ge 0 \qquad \text{for all } w \in \mathbb{R}^N \text{ with } \mathbf{1}_N^\top w = 0$$

is equivalent to the matrix inequality

$$H_N K H_N \succeq 0.$$

Indeed, assume first that $H_N K H_N \succeq 0$ and let $w \in \mathbb{R}^N$ satisfy $\mathbf{1}_N^\top w = 0$. Since $H_N w = w$, we have

$$w^\top K w = (H_N w)^\top K (H_N w) = w^\top (H_N K H_N) w \ge 0.$$

Conversely, assume that $w^\top K w \ge 0$ for every $w \in \mathbb{R}^N$ with $\mathbf{1}_N^\top w = 0$. Fix $v \in \mathbb{R}^N$ and set $w := H_N v$, so that $\mathbf{1}_N^\top w = 0$. Then

$$v^\top (H_N K H_N) v = (H_N v)^\top K (H_N v) = w^\top K w \ge 0.$$

Since this holds for all $v \in \mathbb{R}^N$, it follows that $H_N K H_N \succeq 0$.

B Basic identities for MMD^2

Lemma B.1 (Equivalence of nonnegativity of $\mathbf{MMD}^2$ on a set and CPD on that set)**.** Let $k\colon X \times X \to \mathbb{R}$ be a symmetric kernel on a measurable space X, and let $E \subseteq X$ be measurable. Then

$$\boxed{\begin{array}{c}\mathbf{MMD}_k^2(P,Q) \geq 0 \quad \forall\, P, Q \in \mathcal{P}(X) \text{ with } \operatorname{supp} P, \operatorname{supp} Q \subseteq E \\ \text{if and only if} \\ k \text{ is CPD on } E\end{array}}$$

Here, "CPD on E" means that

$$\iint_{E\times E} k(x,y)\, d\mu(x)\, d\mu(y) \geq 0$$

for every finite signed measure μ supported on E with $\mu(E) = 0$ for which this integral is well defined.

Proof.

CPD on E implies nonnegativity. Take any two probability measures P, Q supported on E, and set

$$\sigma := P - Q.$$

Then σ is supported on E and satisfies $\sigma(E) = 1 - 1 = 0$. Hence, by the assumption that k is CPD on E,

$$\mathbf{MMD}_k^2(P,Q) = \iint k(x,y)\, d\sigma(x)\, d\sigma(y) \geq 0.$$

Nonnegativity implies CPD on E. Assume that $\mathbf{MMD}_k^2(P,Q) \geq 0$ for all probability measures P, Q supported on E for which $\mathbf{MMD}_k^2(P,Q)$ is well defined. Let μ be any finite signed measure supported on E with $\mu(E) = 0$ for which $\iint k\, d\mu\, d\mu$ is well defined. By the Hahn–Jordan decomposition,

$$\mu = \mu^+ - \mu^-, \qquad \mu^+(E) = \mu^-(E) =: A,$$

where the equality follows from $\mu(E) = 0$. If $A = 0$, then $\mu^+(E) = \mu^-(E) = 0$, hence $\mu^+ = \mu^- = 0$, so that $\mu = 0$. In that case,

$$\iint k(x,y)\, d\mu(x)\, d\mu(y) = 0,$$

and the conclusion is immediate. Assume therefore that $A > 0$. Define

$$P = \frac{1}{A}\mu^+, \qquad Q = \frac{1}{A}\mu^-.$$

These are probability measures supported on E, and they satisfy

$$P - Q = \frac{1}{A}\mu.$$

By bilinearity of the quadratic form and the identity $\mu = A(P - Q)$,

$$\begin{aligned}\iint k(x,y)\, d\mu(x)\, d\mu(y) &= A^2 \iint k(x,y)\, d(P-Q)(x)\, d(P-Q)(y) \\ &= A^2\, \mathbf{MMD}_k^2(P,Q) \\ &\geq 0.\end{aligned}$$

Hence k is CPD on E. □

Lemma B.2 ($\mathbf{MMD}^2$ of pushforward measures)**.** Let $f\colon X \to Y$ be a measurable map, let ν, μ be probability measures on X, let $k\colon Y \times Y \to \mathbb{R}$ be any kernel, and define $k_f(x,x') := k\big(f(x), f(x')\big)$. Then

$$\boxed{\mathbf{MMD}_k^2\big(f\#\nu,\ f\#\mu\big) \ =\ \mathbf{MMD}_{k_f}^2\big(\nu,\ \mu\big).}$$

Proof. Define the pushforward measures

$$(f\#\nu)(B) \;=\; \nu\big(f^{-1}(B)\big), \quad (f\#\mu)(B) \;=\; \mu\big(f^{-1}(B)\big).$$

By the definition of $\mathbf{MMD}^2$,

$$\begin{aligned}
\mathbf{MMD}_k^2\big(f\#\nu, f\#\mu\big) &= \mathop{\mathbb{E}}_{u,u'\sim f\#\nu}\big[k(u,u')\big] + \mathop{\mathbb{E}}_{v,v'\sim f\#\mu}\big[k(v,v')\big] \\
&\quad - 2\mathop{\mathbb{E}}_{u\sim f\#\nu,\, v\sim f\#\mu}\big[k(u,v)\big] \\
&= \mathop{\mathbb{E}}_{x,x'\sim\nu}\big[k\big(f(x),f(x')\big)\big] + \mathop{\mathbb{E}}_{y,y'\sim\mu}\big[k\big(f(y),f(y')\big)\big] \\
&\quad - 2\mathop{\mathbb{E}}_{x\sim\nu,\, y\sim\mu}\big[k\big(f(x),f(y)\big)\big] \\
&= \mathop{\mathbb{E}}_{x,x'\sim\nu}\big[k_f(x,x')\big] + \mathop{\mathbb{E}}_{y,y'\sim\mu}\big[k_f(y,y')\big] - 2\mathop{\mathbb{E}}_{x\sim\nu,\, y\sim\mu}\big[k_f(x,y)\big] \\
&= \mathbf{MMD}_{k_f}^2\left(\nu,\mu\right).
\end{aligned}$$

□

Lemma B.3 ($\mathbf{MMD}^2$ is linear in the kernel)**.** Let ν, μ be probability measures on a measurable space X, let $k_1, k_2 : X \times X \to \mathbb{R}$ be kernels, and let $\alpha_1, \alpha_2 \in \mathbb{R}$. Then

$$\boxed{\mathbf{MMD}^2_{\alpha_1 k_1 + \alpha_2 k_2}\big(\nu,\mu\big) = \alpha_1 \mathbf{MMD}^2_{k_1}\big(\nu,\mu\big) + \alpha_2 \mathbf{MMD}^2_{k_2}\big(\nu,\mu\big).}$$

In particular, setting $k_2 \equiv 0$ gives

$$\mathbf{MMD}^2_{\alpha_1 k_1}\big(\nu,\mu\big) = \alpha_1 \mathbf{MMD}^2_{k_1}\big(\nu,\mu\big).$$

Proof. This follows from

$$\mathbf{MMD}^2_k\big(\nu,\mu\big) = \iint_{X\times X} k(x,y)\,\big(\nu-\mu\big)(dx)\,\big(\nu-\mu\big)(dy)$$

by linearity of the integral. □

C Bellman contraction in Maximum Mean Discrepancy

Preamble. This section provides the contraction theory underlying all subsequent arguments. By Theorem C.1, for a fixed map f, an MMD inequality of the form

$$\mathbf{MMD}_k^2\big(f\#\nu, f\#\mu\big) \le \tau\, \mathbf{MMD}_k^2(\nu,\mu)$$

holds for all ν, μ if and only if the *difference kernel* $\tau k - k_f$ is CPD, where $k_f(x,y) = k(f(x), f(y))$. This criterion is then extended to the full distributional Bellman operator by representing $\mathcal{T}^\pi$ as a mixture of affine pushforwards and exploiting convexity of the CPD quadratic form under averaging, yielding a condition that is both sufficient and necessary for τ-contraction in $\overline{\mathbf{MMD}}_k^2$ across all MDPs with discount γ (Theorem C.2). Section C.2 then extends the analysis to return laws supported on a bounded set, where the CPD condition is required only on that set. It is therefore weaker, and holds for some kernels that fail it on all of $\mathbb{R}^d$. For radial kernels on $\mathbb{R}^d$ with d fixed, as in Proposition E.2, the same criterion can be equivalently expressed as a conditionally negative definite (CND) property for an associated radial kernel. This connects Bellman contraction in MMD to CND kernels and to the characterizations of Schoenberg and Bernstein used in the next section.

C.1 Contraction via difference kernel

Theorem C.1 (Contraction of MMD on a set under CPD kernels)**.** Let $f\colon X \to X$ be measurable, let $k\colon X \times X \to \mathbb{R}$ be a symmetric, conditionally positive definite (CPD) kernel, fix a measurable set $E \subseteq X$, and define

$$k_f(x,y) = k\big(f(x), f(y)\big), \qquad x, y \in X.$$

Fix $\tau \in \mathbb{R}$. Then

$$\boxed{\begin{array}{c} \mathbf{MMD}_k^2\big(f\#\nu, f\#\mu\big) \;\le\; \tau\, \mathbf{MMD}_k^2(\nu,\mu) \qquad \forall\, \mu,\nu \in \mathcal{P}(X) \text{ with } \operatorname{supp}\mu, \operatorname{supp}\nu \subseteq E \\ \text{if and only if} \\ \tau\, k \;-\; k_f \text{ is CPD on } E \end{array}}$$

Here, "CPD on E" means that

$$\iint_{E\times E} \big(\tau k - k_f\big)(x,y)\, d\sigma(x)\, d\sigma(y) \ge 0$$

for every finite signed measure σ supported on E with $\sigma(E) = 0$ for which this integral is well defined.

Accordingly, for laws supported on E, the smallest such τ is

$$\tau_E^\star := \inf\{\tau > 0 :\ \tau k - k_f \text{ is CPD on } E\}.$$

Contraction on laws supported on E holds exactly when $\tau_E^\star < 1$. In that case, $\tau_E^\star$ is the smallest contraction factor on E in $\mathbf{MMD}^2$, and the corresponding smallest contraction factor in $\mathbf{MMD}$ is $\sqrt{\tau_E^\star}$. The choice $E = X$ yields contraction on all probability laws on X, while proper subsets E restrict the comparison to laws supported on a prescribed region, such as the bounded sets used later.

Proof. Fix $\mu, \nu \in \mathcal{P}(X)$ with

$$\operatorname{supp}\mu, \operatorname{supp}\nu \subseteq E.$$

By Lemma B.2, we have

$$\mathbf{MMD}_k^2\big(f\#\nu, f\#\mu\big) = \mathbf{MMD}_{k_f}^2(\nu,\mu).$$

By Lemma B.3,

$$\tau\, \mathbf{MMD}_k^2(\nu,\mu) = \mathbf{MMD}_{\tau k}^2(\nu,\mu).$$

Hence

$$\tau\, \mathbf{MMD}_k^2(\nu,\mu) - \mathbf{MMD}_k^2\big(f\#\nu, f\#\mu\big) = \mathbf{MMD}_{\tau k}^2(\nu,\mu) - \mathbf{MMD}_{k_f}^2(\nu,\mu).$$

By Lemma B.3, this difference is

$$\mathbf{MMD}_{\tau k - k_f}^2(\nu,\mu).$$

Therefore,

$$\mathbf{MMD}_k^2\big(f\#\nu, f\#\mu\big) \le \tau\,\mathbf{MMD}_k^2(\nu,\mu) \quad \Longleftrightarrow \quad \mathbf{MMD}_{\tau k - k_f}^2(\nu,\mu) \ge 0.$$

Thus

$$\mathbf{MMD}_k^2\big(f\#\nu, f\#\mu\big) \le \tau\,\mathbf{MMD}_k^2(\nu,\mu) \qquad \forall\,\mu,\nu \in \mathcal{P}(X) \text{ with } \operatorname{supp}\mu, \operatorname{supp}\nu \subseteq E$$

if and only if

$$\mathbf{MMD}_{\tau k - k_f}^2(\nu,\mu) \ge 0 \qquad \forall\,\mu,\nu \in \mathcal{P}(X) \text{ with } \operatorname{supp}\mu, \operatorname{supp}\nu \subseteq E.$$

Finally, by Lemma B.1, this is equivalent to $\tau k - k_f$ being CPD on E. □

Lemma C.1 (The Bellman operator averages pushforwards under affine maps)**.** Fix $(s,a) \in \mathcal{S}\times\mathcal{A}$, and let

$$C := \big(R(s,a), S', A'\big)$$

with law $\rho_{s,a}$. For each realization $C = (q, s', a')$, define

$$f_q : \mathbb{R}^d \to \mathbb{R}^d, \qquad f_q(x) := q + \gamma x.$$

Let $\eta : \mathcal{S}\times\mathcal{A} \to \mathcal{P}(\mathbb{R}^d)$, and let $X' \mid C \sim \eta(S', A')$. Given C (the reward, next state and next action), the Bellman update maps the next return law to its pushforward under the affine map f_q,

$$\operatorname{Law}(q + \gamma X' \mid C) = f_q\#\eta(s', a').$$

Consequently,

$$\boxed{(\mathcal{T}^\pi\eta)(s,a) = \int f_q\#\eta(s',a')\,\rho_{s,a}(dC).}$$

Proof. Fix a realization $C = (q, s', a')$. Conditionally on C, the random variable X' has law $\eta(s', a')$. Hence, for every Borel set $B \subseteq \mathbb{R}^d$,

$$\Pr(q + \gamma X' \in B \mid C) = \Pr(X' \in f_q^{-1}(B) \mid C) = \eta(s',a')\big(f_q^{-1}(B)\big) = (f_q\#\eta(s',a'))(B).$$

Therefore,

$$\operatorname{Law}(q + \gamma X' \mid C) = f_q\#\eta(s', a').$$

Integrating this conditional law with respect to $\rho_{s,a}$ yields

$$(\mathcal{T}^\pi\eta)(s,a) = \int f_q\#\eta(s',a')\,\rho_{s,a}(dC),$$

which is the claimed representation. □

Lemma C.2 (A Bellman operator can reduce to an affine map)**.** Fix $q \in \mathbb{R}^d$ and $\gamma \in (0,1)$. Then there exists a discounted MDP with a single state s, a single action a whose transition always returns to s, and deterministic reward q, together with its unique fixed policy π, such that

$$\boxed{(\mathcal{T}^\pi\eta)(s,a) = f_q\#\eta(s,a), \qquad f_q(x) = q + \gamma x,}$$

for every law $\eta(s,a) \in \mathcal{P}(\mathbb{R}^d)$. Moreover, since $\mathcal{S}\times\mathcal{A} = \{(s,a)\}$,

$$\overline{\mathbf{MMD}}_k^2(\eta_1,\eta_2) = \mathbf{MMD}_k^2\big(\eta_1(s,a), \eta_2(s,a)\big)$$

for all laws $\eta_1(s,a), \eta_2(s,a) \in \mathcal{P}(\mathbb{R}^d)$.

Proof. Consider the MDP with one state s, one action a, transition kernel

$$P(\{s\} \mid s, a) = 1,$$

and deterministic reward equal to q. Let π be the unique policy on this MDP. For any law $\eta(s,a) \in \mathcal{P}(\mathbb{R}^d)$, let $X' \sim \eta(s,a)$. By definition of the distributional Bellman operator,

$$(\mathcal{T}^\pi\eta)(s,a) = \operatorname{Law}(q + \gamma X') = f_q\#\eta(s,a).$$

The identity for $\overline{\mathbf{MMD}}_k^2$ is immediate because there is only one state-action pair. □

Lemma C.3 (Averaging does not increase the CPD quadratic form)**.** Let $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a symmetric CPD kernel, let $(\Omega, \mathcal{F}, \rho)$ be a probability space, and let $(\sigma_c)_{c\in\Omega}$ be a measurable family of finite signed measures on $\mathbb{R}^d$, each with total mass 0. Define

$$\bar{\sigma} := \int_\Omega \sigma_c \, \rho(dc).$$

Then

$$\boxed{\iint k(x,y)\, d\bar{\sigma}(x)\, d\bar{\sigma}(y) \le \int_\Omega \left(\iint k(x,y)\, d\sigma_c(x)\, d\sigma_c(y) \right) \rho(dc)}$$

whenever the quadratic forms are well defined.

Proof. Define

$$Q(\sigma) := \iint k(x,y)\, d\sigma(x)\, d\sigma(y).$$

Since k is CPD and each $\sigma_c - \bar{\sigma}$ has total mass 0, we have

$$Q(\sigma_c - \bar{\sigma}) \ge 0 \qquad \text{for every } c \in \Omega.$$

Integrating with respect to ρ yields

$$0 \le \int_\Omega Q(\sigma_c - \bar{\sigma})\, \rho(dc).$$

Expanding the quadratic form and using bilinearity,

$$\begin{aligned}\int_\Omega Q(\sigma_c - \bar{\sigma})\, \rho(dc) = \int_\Omega Q(\sigma_c)\, \rho(dc) - 2\int_\Omega \iint k(x,y)\, d\sigma_c(x)\, d\bar{\sigma}(y)\, \rho(dc) \\ + Q(\bar{\sigma}).\end{aligned}$$

To simplify the mixed term, write

$$\bar{\sigma} = \int_\Omega \sigma_{c'}\, \rho(dc'),$$

where c' is a dummy variable. Then, using bilinearity to expand $\bar{\sigma}$ and Fubini to interchange the c and c' integrals, we obtain

$$\begin{aligned}\int_\Omega \iint k(x,y)\, d\sigma_c(x)\, d\bar{\sigma}(y)\, \rho(dc) &= \int_\Omega \int_\Omega \left(\iint k(x,y)\, d\sigma_c(x)\, d\sigma_{c'}(y) \right) \rho(dc')\, \rho(dc) \\ &= \iint k(x,y)\, d\left(\int_\Omega \sigma_c\, \rho(dc) \right)(x)\, d\left(\int_\Omega \sigma_{c'}\, \rho(dc') \right)(y) \\ &= \iint k(x,y)\, d\bar{\sigma}(x)\, d\bar{\sigma}(y) \\ &= Q(\bar{\sigma}).\end{aligned}$$

Therefore

$$0 \le \int_\Omega Q(\sigma_c)\, \rho(dc) - Q(\bar{\sigma}),$$

which is exactly the claimed inequality. □

Lemma C.4 ($\mathbf{MMD}^2$ contraction from affine maps to the Bellman operator)**.** Let $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a symmetric CPD kernel as in Definition A.5, let $E \subseteq \mathbb{R}^d$, and let $\eta_1, \eta_2 : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R}^d)$ satisfy

$$\operatorname{supp} \eta_i(s,a) \subseteq E \qquad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}, \quad i = 1, 2.$$

Assume that there exists $\tau \ge 0$ such that, for every $q \in \mathbb{R}^d$, the affine map

$$f_q : \mathbb{R}^d \to \mathbb{R}^d, \qquad f_q(x) = q + \gamma x,$$

satisfies

$$\mathbf{MMD}_k^2\big(f_q \# \nu, f_q \# \mu\big) \le \tau\, \mathbf{MMD}_k^2(\nu, \mu) \qquad \forall\, \nu, \mu \in \mathcal{P}(\mathbb{R}^d) \text{ with } \operatorname{supp}\nu, \operatorname{supp}\mu \subseteq E,$$

for which the quantities are well defined. Then, for every $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\boxed{\mathbf{MMD}_k^2\big((\mathcal{T}^\pi \eta_1)(s,a), (\mathcal{T}^\pi \eta_2)(s,a)\big) \le \tau\, \overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2).}$$

Proof. Fix $(s,a) \in \mathcal{S} \times \mathcal{A}$, and let

$$C := \big(R(s,a), S', A'\big)$$

denote the reward, next state and next action drawn from (s,a), with law

$$\rho_{s,a} := \mathrm{Law}(C).$$

By Lemma C.1, for each realization $C = (q, s', a')$, define

$$\mu_C := f_q \# \eta_1(s', a'), \qquad \nu_C := f_q \# \eta_2(s', a').$$

Then

$$(\mathcal{T}^\pi \eta_1)(s,a) = \int \mu_C \, \rho_{s,a}(dC), \qquad (\mathcal{T}^\pi \eta_2)(s,a) = \int \nu_C \, \rho_{s,a}(dC).$$

Comparing quadratic forms for mixtures.

Write

$$\sigma_C := \mu_C - \nu_C, \qquad \bar{\sigma} := \int \sigma_C \, \rho_{s,a}(dC).$$

Since

$$\bar{\sigma} = (\mathcal{T}^\pi \eta_1)(s,a) - (\mathcal{T}^\pi \eta_2)(s,a),$$

Lemma C.3 gives

$$\iint k(x,y) \, d\bar{\sigma}(x) \, d\bar{\sigma}(y) \le \int \left(\iint k(x,y) \, d\sigma_C(x) \, d\sigma_C(y) \right) \rho_{s,a}(dC).$$

By the definition of $\mathbf{MMD}_k^2$ as a quadratic form for CPD kernels in Definition A.6, this yields

$$\mathbf{MMD}_k^2\big((\mathcal{T}^\pi \eta_1)(s,a), (\mathcal{T}^\pi \eta_2)(s,a)\big) \le \int \mathbf{MMD}_k^2(\mu_C, \nu_C) \, \rho_{s,a}(dC).$$

Applying the assumed contraction on E.

Since $\eta_1(s', a')$ and $\eta_2(s', a')$ are supported on E, the inequality assumed for f_q applies and gives

$$\mathbf{MMD}_k^2(\mu_C, \nu_C) \le \tau \, \mathbf{MMD}_k^2\big(\eta_1(s', a'), \eta_2(s', a')\big).$$

Therefore,

$$\begin{aligned} \mathbf{MMD}_k^2\big((\mathcal{T}^\pi \eta_1)(s,a), (\mathcal{T}^\pi \eta_2)(s,a)\big) &\le \int \mathbf{MMD}_k^2(\mu_C, \nu_C) \, \rho_{s,a}(dC) \\ &\le \tau \int \mathbf{MMD}_k^2\big(\eta_1(s', a'), \eta_2(s', a')\big) \, \rho_{s,a}(dC) \\ &\le \tau \, \overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2), \end{aligned}$$

which proves the claim. □

Theorem C.2 (CPD criterion for Bellman contraction in $\overline{\mathbf{MMD}}_k^2$)**.** Let $k \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a symmetric CPD kernel as in Definition A.5, and fix $\tau \in [0, 1)$. For each $q \in \mathbb{R}^d$, define the Bellman difference kernel

$$k_{\mathrm{diff}}^{q,\gamma}(x,y) := \tau k(x,y) - k(q + \gamma x, q + \gamma y).$$

If $k_{\mathrm{diff}}^{q,\gamma}$ is CPD for every $q \in \mathbb{R}^d$, then $\mathcal{T}^\pi$ is a τ-contraction in $\overline{\mathbf{MMD}}_k^2$, that is,

$$\boxed{\overline{\mathbf{MMD}}_k^2(\mathcal{T}^\pi \eta_1, \mathcal{T}^\pi \eta_2) \le \tau \, \overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2) \qquad \forall \, \eta_1, \eta_2 : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R}^d).}$$

Furthermore, this condition is necessary for τ-contraction to hold for the Bellman operator of every MDP with discount γ under every fixed policy.

Proof.

Sufficiency. Let $\eta_1, \eta_2 : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R}^d)$. By Theorem C.1, applied with $X = \mathbb{R}^d$, $E = \mathbb{R}^d$, and

$$f_q(x) = q + \gamma x,$$

the assumption that $k_{\text{diff}}^{q,\gamma}$ is CPD for every $q \in \mathbb{R}^d$ implies that

$$\mathbf{MMD}_k^2\big(f_q \# \nu, f_q \# \mu\big) \le \tau\, \mathbf{MMD}_k^2(\nu, \mu) \qquad \forall\, \nu, \mu \in \mathcal{P}(\mathbb{R}^d),$$

for every $q \in \mathbb{R}^d$ and for all quantities that are well defined. Applying Lemma C.4 with $E = \mathbb{R}^d$, we obtain, for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathbf{MMD}_k^2\big((\mathcal{T}^\pi \eta_1)(s,a), (\mathcal{T}^\pi \eta_2)(s,a)\big) \le \tau\, \overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2).$$

Taking the supremum over (s, a) yields

$$\overline{\mathbf{MMD}}_k^2(\mathcal{T}^\pi \eta_1, \mathcal{T}^\pi \eta_2) \le \tau\, \overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2),$$

so $\mathcal{T}^\pi$ is a τ-contraction in $\overline{\mathbf{MMD}}_k^2$.

Necessity. Assume now that τ-contraction holds for the Bellman operator of every MDP with discount γ under every fixed policy. Fix $q \in \mathbb{R}^d$. By Lemma C.2, there exists a discounted MDP and a fixed policy for which

$$(\mathcal{T}^\pi \eta)(s,a) = f_q \# \eta(s,a)$$

at the unique state-action pair (s, a). Since in this problem

$$\overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2) = \mathbf{MMD}_k^2\big(\eta_1(s,a), \eta_2(s,a)\big),$$

the assumed Bellman τ-contraction implies

$$\mathbf{MMD}_k^2\big(f_q \# \nu, f_q \# \mu\big) \le \tau\, \mathbf{MMD}_k^2(\nu, \mu) \qquad \forall\, \nu, \mu \in \mathcal{P}(\mathbb{R}^d).$$

Applying Theorem C.1 with $X = \mathbb{R}^d$ and $E = \mathbb{R}^d$, we conclude that $k_{\text{diff}}^{q,\gamma}$ is CPD. Since q was arbitrary, this holds for every $q \in \mathbb{R}^d$. □

C.2 Contraction on a bounded support

Proposition C.1 (The ball as a natural set under bounded rewards)**.** Assume that there exists $r_{\max} \geq 0$ such that

$$\|R(s,a)\| \leq r_{\max} \qquad \text{for every } (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Let

$$E := B_M = \{x \in \mathbb{R}^d : \|x\| \leq M\}, \qquad M \geq \frac{r_{\max}}{1-\gamma}.$$

Then the following hold.

1. For every $(s,a) \in \mathcal{S} \times \mathcal{A}$, the return law $\eta^\pi(s,a)$ is supported on E.
2. If $\eta : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R}^d)$ satisfies

   $$\operatorname{supp} \eta(s,a) \subseteq E \qquad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A},$$

   then also

   $$\operatorname{supp} (\mathcal{T}^\pi \eta)(s,a) \subseteq E \qquad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}.$$

   In other words, the class of laws supported on E is invariant under $\mathcal{T}^\pi$.

Proof.

1. Bounded rewards imply bounded returns.

Fix $(s,a) \in \mathcal{S} \times \mathcal{A}$, and consider the random return

$$Z^\pi(s,a) = \sum_{t \geq 0} \gamma^t R_t.$$

For almost every realization of this random variable, the triangle inequality and the reward bound give

$$\|Z^\pi(s,a)\| \leq \sum_{t \geq 0} \gamma^t \|R_t\| \leq \sum_{t \geq 0} \gamma^t r_{\max} = \frac{r_{\max}}{1-\gamma} \leq M.$$

Thus almost every realization of $Z^\pi(s,a)$ belongs to $E = B_M$. Therefore,

$$\operatorname{supp} \eta^\pi(s,a) \subseteq E.$$

2. Bellman invariance of the ball.

Assume now that

$$\operatorname{supp} \eta(s,a) \subseteq E \qquad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Fix (s,a), and let $X' \sim \eta(S', A')$ in the Bellman update. Since $\eta(S', A')$ is supported on E, almost every realization of X' satisfies

$$\|X'\| \leq M.$$

Hence

$$\|R(s,a) + \gamma X'\| \leq \|R(s,a)\| + \gamma \|X'\| \leq r_{\max} + \gamma M \leq M,$$

where the last inequality follows from $M \geq r_{\max}/(1-\gamma)$. Thus almost every realization of $R(s,a) + \gamma X'$ lies in E, so

$$\operatorname{supp} (\mathcal{T}^\pi \eta)(s,a) \subseteq E.$$

Since (s,a) was arbitrary, the conclusion holds for all state-action pairs. □

Theorem C.3 (CPD criterion for Bellman contraction on a set)**.** Let $k \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a symmetric CPD kernel as in Definition A.5, let $E \subseteq \mathbb{R}^d$ be measurable, and fix $\tau \in [0,1)$. For each $q \in \mathbb{R}^d$, define the Bellman difference kernel

$$k^{q,\gamma}_{\text{diff}}(x,y) := \tau k(x,y) - k(q + \gamma x, q + \gamma y).$$

Assume that $k^{q,\gamma}_{\text{diff}}$ is CPD on E for every $q \in \mathbb{R}^d$. Let

$$\mathcal{C}_E := \big\{\eta : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R}^d) \ : \ \operatorname{supp} \eta(s,a) \subseteq E \ \ \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}\big\}.$$

Then, for all $\eta_1, \eta_2 \in \mathcal{C}_E$,

$$\boxed{\overline{\mathbf{MMD}}_k^2(\mathcal{T}^\pi \eta_1, \mathcal{T}^\pi \eta_2) \le \tau\, \overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2).}$$

If, in addition, $\mathcal{T}^\pi(\mathcal{C}_E) \subseteq \mathcal{C}_E$, then $\mathcal{T}^\pi$ is a τ-contraction on $\mathcal{C}_E$. Conversely, assume that for every MDP with discount γ under every fixed policy, whenever $\mathcal{T}^\pi(\mathcal{C}_E) \subseteq \mathcal{C}_E$, the Bellman operator is a τ-contraction on $\mathcal{C}_E$. Then, for every $q \in \mathbb{R}^d$ such that $q + \gamma E \subseteq E$, the kernel $k_{\mathrm{diff}}^{q,\gamma}$ is CPD on E. Proposition C.1 gives one natural situation in which such an invariant set E exists.

Proof.

Sufficiency. Fix $\eta_1, \eta_2 \in \mathcal{C}_E$. By Theorem C.1, applied with $X = \mathbb{R}^d$, the set E, and

$$f_q(x) = q + \gamma x,$$

the assumption that $k_{\mathrm{diff}}^{q,\gamma}$ is CPD on E implies that, for every $q \in \mathbb{R}^d$,

$$\mathbf{MMD}_k^2\big(f_q \# \nu, f_q \# \mu\big) \le \tau\, \mathbf{MMD}_k^2(\nu, \mu) \qquad \forall\, \mu, \nu \in \mathcal{P}(\mathbb{R}^d) \text{ with } \operatorname{supp}\mu, \operatorname{supp}\nu \subseteq E.$$

Since $\eta_1, \eta_2 \in \mathcal{C}_E$, the assumptions of Lemma C.4 are satisfied with the same set E. Applying that lemma, we obtain, for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathbf{MMD}_k^2\big((\mathcal{T}^\pi \eta_1)(s,a), (\mathcal{T}^\pi \eta_2)(s,a)\big) \le \tau\, \overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2).$$

Taking the supremum over (s, a) yields

$$\overline{\mathbf{MMD}}_k^2(\mathcal{T}^\pi \eta_1, \mathcal{T}^\pi \eta_2) \le \tau\, \overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2),$$

which proves the boxed inequality.

If, in addition, $\mathcal{T}^\pi(\mathcal{C}_E) \subseteq \mathcal{C}_E$, then $\mathcal{T}^\pi$ maps $\mathcal{C}_E$ into itself and satisfies the above inequality on $\mathcal{C}_E$. Hence $\mathcal{T}^\pi$ is a τ-contraction on $\mathcal{C}_E$.

Necessity. Suppose that, for every MDP with discount γ under every fixed policy, the Bellman operator is a τ-contraction on $\mathcal{C}_E$ whenever it maps $\mathcal{C}_E$ into itself. We show that this forces $k_{\mathrm{diff}}^{q,\gamma}$ to be CPD on E for every $q \in \mathbb{R}^d$ satisfying

$$q + \gamma E \subseteq E.$$

Fix such a q. By Lemma C.2, there exists a discounted MDP with a single state s, a single action a whose transition always returns to s, deterministic reward q, and its unique fixed policy π, such that

$$(\mathcal{T}^\pi \eta)(s,a) = f_q \# \eta(s,a), \qquad f_q(x) = q + \gamma x.$$

Moreover, this Bellman operator preserves laws supported on E. Indeed, if $\eta \in \mathcal{C}_E$, then $\eta(s,a)$ is supported on E, and since $q + \gamma E \subseteq E$, the pushforward $f_q \# \eta(s,a)$ is also supported on E. Hence

$$\mathcal{T}^\pi(\mathcal{C}_E) \subseteq \mathcal{C}_E.$$

Therefore, the assumption applies to this MDP. Since $\mathcal{S} \times \mathcal{A} = \{(s,a)\}$, one has

$$\overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2) = \mathbf{MMD}_k^2\big(\eta_1(s,a), \eta_2(s,a)\big)$$

for all $\eta_1, \eta_2 \in \mathcal{C}_E$. The assumed τ-contraction on $\mathcal{C}_E$ therefore implies

$$\mathbf{MMD}_k^2\big(f_q \# \nu, f_q \# \mu\big) \le \tau\, \mathbf{MMD}_k^2(\nu, \mu) \qquad \forall\, \mu, \nu \in \mathcal{P}(\mathbb{R}^d) \text{ with } \operatorname{supp}\mu, \operatorname{supp}\nu \subseteq E.$$

Applying Theorem C.1 with the set E, we conclude that $k_{\mathrm{diff}}^{q,\gamma} = \tau k - k_{f_q}$ is CPD on E. Since q was arbitrary among those satisfying $q + \gamma E \subseteq E$, this proves the necessity statement. □

Lemma C.5 (CPD on a set implies CPD on all subsets)**.** Let X be a measurable space, let $k : X \times X \to \mathbb{R}$ be symmetric, and let $F \subseteq E \subseteq X$ be measurable. If k is CPD on E, then k is CPD on F.

Proof. Assume that k is CPD on E. Let μ be any finite signed measure supported on F with $\mu(F) = 0$. Since $F \subseteq E$, the measure μ is also supported on E. Moreover,

$$\mu(E) = \mu(F) = 0,$$

because μ vanishes outside F. Since k is CPD on E, it follows that

$$\iint_{E\times E} k(x, y)\, d\mu(x)\, d\mu(y) \geq 0.$$

As μ is supported on F, this is the same as

$$\iint_{F\times F} k(x, y)\, d\mu(x)\, d\mu(y) \geq 0.$$

Hence k is CPD on F. □

D Contraction on a bounded support: numerical experiments

Preamble. This section turns the criterion of Theorem C.3, CPD on the bounded set E, into a numerical procedure for estimating $\tau_E^\star(\lambda)$. Section D.1 replaces $E = B(0, R)$ by random finite sets of points in E, which by Lemma C.5 give lower bounds on $\tau_E^\star(\lambda)$ that can only improve as more sets are drawn. Section D.2 reduces the CPD condition on a finite support to a generalized eigenvalue problem, solved by the symmetric reduction of Algorithm 1. Section D.4 validates the procedure against the power-distance and generalized multiquadric kernels of [19, 25], whose contraction factors are known. Section D.3 gives the settings behind Figure 2, which applies the procedure to the RBF and logarithmic kernels.

D.1 From a bounded to a discrete support condition

By Theorem C.1, for a fixed map, $\mathbf{MMD}^2$ contraction on a set $E \subseteq \mathbb{R}^d$ is equivalent to the corresponding difference kernel being CPD on E. Theorem C.3 gives Bellman contraction on E when this holds for the map $x \mapsto q + \gamma x$ of every reward value q.

For a translation-invariant kernel, $k(q + \lambda x, q + \lambda y) = k(\lambda x, \lambda y)$, so the difference kernel is the same for every reward value q. We therefore take $q = 0$ and $f(x) = \lambda x$, and the relevant difference kernel is

$$k_{\mathrm{diff},\tau}^{0,\lambda}(x, y) := \tau k(x, y) - k(\lambda x, \lambda y), \qquad \lambda \in (0, 1).$$

For each measurable set $E \subseteq \mathbb{R}^d$, define

$$\tau_E^\star(\lambda) := \inf\Big\{\tau > 0 :\ k_{\mathrm{diff},\tau}^{0,\lambda} \text{ is CPD on } E\Big\}.$$

By Lemma C.5, CPD on a set implies CPD on every subset, so $\tau_E^\star(\lambda)$ can only increase when E grows. In particular,

$$\tau_{\mathbb{R}^d}^\star(\lambda) \geq \tau_E^\star(\lambda) \qquad \text{for every measurable } E \subseteq \mathbb{R}^d.$$

By Theorem 3.1, for a translation-invariant kernel and $\lambda = \gamma$, $\tau_{\mathbb{R}^d}^\star(\gamma)$ is the smallest contraction factor $\tau_k^\star(\mathcal{T}^\pi)$ of the main text. Thus, requiring CPD only on a bounded set E is less restrictive than requiring it on all of $\mathbb{R}^d$, so the condition can hold on E for a kernel that fails it on all of $\mathbb{R}^d$. In reinforcement learning, Proposition C.1 shows that under bounded rewards, return laws are supported in a ball $E = B_M$ with $M \geq r_{\max}/(1 - \gamma)$.

The CPD condition on E concerns all signed measures supported on E, so it cannot be checked numerically. We test it instead on J finite sets of points drawn at random in E,

$$S_j = \{z_1^{(j)}, \ldots, z_N^{(j)}\} \subseteq E, \qquad j = 1, \ldots, J,$$

and define the contraction threshold of each set,

$$\tau_{S_j}^\star(\lambda) := \inf\Big\{\tau > 0 :\ k_{\mathrm{diff},\tau}^{0,\lambda} \text{ is CPD on } S_j\Big\}.$$

Since each S_j is a subset of E, Lemma C.5 gives

$$\tau_{\mathbb{R}^d}^\star(\lambda) \geq \tau_E^\star(\lambda) \geq \max_{1 \leq j \leq J} \tau_{S_j}^\star(\lambda).$$

So the largest of these thresholds is a lower bound on $\tau_E^\star(\lambda)$. It can only increase with J, because the maximum is taken over more sets.

D.2 From a support condition to a matrix test

Fix a finite set of points $S = \{z_1, \ldots, z_N\} \subseteq E$. We call it a finite support, because the CPD condition on S is the contraction inequality for laws whose mass sits on these points. The threshold $\tau_S^\star(\lambda)$ of S is determined by whether the difference kernel

$$k_{\mathrm{diff},\tau}^{0,\lambda}(x, y) = \tau k(x, y) - k(\lambda x, \lambda y)$$

is CPD on S. We now express this condition in matrix form.

Checking the CPD condition on S. Define the Gram matrices

$$K \in \mathbb{R}^{N\times N}, \qquad K_{ij} := k(z_i, z_j), \qquad K_\lambda \in \mathbb{R}^{N\times N}, \qquad (K_\lambda)_{ij} := k(\lambda z_i, \lambda z_j),$$

and the corresponding *difference matrix*

$$D_\tau^{(\lambda)} := \tau K - K_\lambda.$$

A signed measure on S is $\sigma = \sum_{i=1}^N w_i\,\delta_{z_i}$, with $\sigma(S) = \mathbf{1}_N^\top w$ and $\iint k_{\mathrm{diff},\tau}^{0,\lambda}\, d\sigma\, d\sigma = w^\top D_\tau^{(\lambda)} w$. The CPD condition on S is therefore

$$w^\top D_\tau^{(\lambda)} w \geq 0 \qquad \text{for all } w \in \mathbf{1}_N^\perp, \qquad \mathbf{1}_N^\perp := \{w \in \mathbb{R}^N : \mathbf{1}_N^\top w = 0\}.$$

Using the centering projector $H_N = I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$ introduced in Section A.3, this is equivalent to

$$H_N D_\tau^{(\lambda)} H_N \succeq 0.$$

A symmetric matrix is positive semidefinite if and only if its smallest eigenvalue is nonnegative. Since $H_N D_\tau^{(\lambda)} H_N$ is symmetric, the CPD condition on S therefore holds if and only if

$$\lambda_{\min}\big(H_N D_\tau^{(\lambda)} H_N\big) \geq 0.$$

Monotonicity in τ. Fix $\tau' \geq \tau$ and any $w \in \mathbf{1}_N^\perp$. Since $D_\tau^{(\lambda)} = \tau K - K_\lambda$, we have

$$w^\top D_{\tau'}^{(\lambda)} w = w^\top D_\tau^{(\lambda)} w + (\tau' - \tau)\, w^\top K w.$$

Thus, if $\tau' > \tau$ and $w^\top K w > 0$, increasing τ strictly increases $w^\top D_\tau^{(\lambda)} w$. More generally, if the CPD condition on S holds for τ, it also holds for every $\tau' \geq \tau$ whenever

$$w^\top K w \geq 0 \qquad \forall\, w \in \mathbf{1}_N^\perp,$$

that is, whenever k is CPD on S. This is the case for every sampled support when k is CPD on $\mathbb{R}^d$.

Working directly with vectors whose entries sum to zero. The CPD condition on S concerns the quadratic form $w^\top D_\tau^{(\lambda)} w$ only for weights w satisfying $\mathbf{1}_N^\top w = 0$, that is, $w \in \mathbf{1}_N^\perp$. Although the centering projector H_N encodes this constraint, we do not form it explicitly. Instead, we choose an orthonormal basis $Q \in \mathbb{R}^{N\times(N-1)}$ of $\mathbf{1}_N^\perp$, so that

$$Q^\top Q = I_{N-1}, \qquad \mathrm{Range}(Q) = \mathbf{1}_N^\perp.$$

Since every $w \in \mathbf{1}_N^\perp$ is $w = Qv$ for some $v \in \mathbb{R}^{N-1}$, and $(Qv)^\top D_\tau^{(\lambda)}(Qv) = v^\top\big(Q^\top D_\tau^{(\lambda)} Q\big)v$, the CPD condition on S holds if and only if $Q^\top D_\tau^{(\lambda)} Q \succeq 0$.

Generalized eigenvalues for vectors whose entries sum to zero. Write the two Gram matrices restricted to $\mathbf{1}_N^\perp$ in the basis Q,

$$A := Q^\top K Q, \qquad B := Q^\top K_\lambda Q,$$

so that

$$Q^\top D_\tau^{(\lambda)} Q = \tau A - B.$$

Hence the CPD condition on S is equivalent to

$$\tau A - B \succeq 0.$$

If A is strictly positive definite, meaning

$$v^\top A v > 0 \qquad \text{for every } v \neq 0,$$

or equivalently, as shown in the next paragraph,

$$w^\top K w > 0 \qquad \text{for every nonzero } w \in \mathbf{1}_N^\perp,$$

then, for every nonzero v,

$$v^\top(\tau A - B)v \geq 0 \qquad \Longleftrightarrow \qquad \tau \geq \frac{v^\top B v}{v^\top A v}.$$

Therefore,

$$\tau_S^\star(\lambda) = \sup_{v\neq 0} \frac{v^\top B v}{v^\top A v} = \lambda_{\max}(B, A),$$

that is, the smallest τ for which the CPD condition holds on S is exactly the largest generalized eigenvalue of (B, A).

CSPD and strict positivity of A. The assumption $A \succ 0$ is equivalent to requiring that k be CSPD on the finite support S in the sense of Definition A.5. Indeed, since the columns of Q form an orthonormal basis of $\mathbf{1}_N^\perp$, every $w \in \mathbf{1}_N^\perp$ can be written uniquely as $w = Qv$ with $v \in \mathbb{R}^{N-1}$, and $w \neq 0$ if and only if $v \neq 0$. Therefore,

$$v^\top A v = v^\top Q^\top K Q v = (Qv)^\top K (Qv) = w^\top K w.$$

Hence A is strictly positive definite if and only if

$$w^\top K w > 0 \qquad \text{for every nonzero } w \in \mathbf{1}_N^\perp,$$

that is, if and only if k is CSPD on S. In particular, this condition is automatically satisfied if k is CSPD globally. By Proposition A.3, this is equivalent to $\mathbf{MMD}_k$ being a metric, and we assume it in the rest of this section.

Reduction to an ordinary symmetric eigenvalue problem. Under the assumption that k is CSPD, A is strictly positive definite. We can therefore reduce the generalized eigenvalue problem to an ordinary symmetric one, using the eigendecomposition of A [26, Theorem 15.3.2, Chapter 15.4, method 2],

$$A = P \Delta^2 P^\top, \qquad \Delta := \operatorname{diag}(\mu_1^{1/2}, \dots, \mu_{N-1}^{1/2}), \qquad \mu_i > 0.$$

We then introduce the change of coordinates $u := \Delta P^\top v$, or equivalently $v = P \Delta^{-1} u$. With $C := \Delta^{-1} P^\top B P \Delta^{-1}$, the generalized eigenvalue problem $Bv = \tau A v$ is transformed into the ordinary symmetric eigenvalue problem $Cu = \tau u$. Indeed,

$$\begin{aligned} Bv &= BP\Delta^{-1}u \\ &= \tau A v \\ &= \tau A P \Delta^{-1} u \\ &= \tau P \Delta^2 P^\top P \Delta^{-1} u \\ &= \tau P \Delta u. \end{aligned}$$

Applying $\Delta^{-1} P^\top$ to both sides of the identity

$$BP\Delta^{-1}u = \tau P \Delta u$$

yields

$$\Delta^{-1} P^\top B P \Delta^{-1} u = \tau u,$$

that is,

$$Cu = \tau u.$$

So the threshold of S is the largest eigenvalue of C,

$$\tau_S^\star(\lambda) = \lambda_{\max}(C).$$

Truncation of very small eigenvalues. Although A is strictly positive definite, some of its eigenvalues can be extremely small in practice. The factors $\mu_i^{-1/2}$ in the reduction above then become very large and make the reduced problem numerically unstable, a difficulty also discussed in [26, Chapter 15.4.1, Remark 5]. We therefore discard very small eigenvalues before forming the reduced matrix, following the idea of [26, Chapter 15.5].

Concretely, we compute

$$A = P \operatorname{diag}(\mu_1, \dots, \mu_{N-1})\, P^\top,$$

choose a relative tolerance $\varepsilon_{\text{eig}} > 0$, and retain only the eigenvalues satisfying

$$\mu_i > \varepsilon_{\text{eig}}\, \mu_{\max}, \qquad \mu_{\max} := \max_i \mu_i.$$

Setting $\mu_i^{-1/2} = 0$ for discarded eigenvalues, we define

$$\Delta_\varepsilon^{-1} := \operatorname{diag}(\nu_1, \dots, \nu_{N-1}), \qquad \nu_i = \begin{cases} \mu_i^{-1/2}, & \mu_i > \varepsilon_{\text{eig}}\, \mu_{\max}, \\ 0, & \text{otherwise.} \end{cases}$$

Table 2: Numerical settings for the heatmaps of Figure 2.

Setting	Value
Radius	$R = 1$
Dimension	$d = 1$
Support sampler	Uniform on the ball of radius R
Support size	$N = 512$
Number of sampled supports	$J = 10$
Parameter grid	$R/\ell \in [0.05, 2.00]$, 81 linearly spaced values
Scaling grid	$\lambda \in [0, 1]$, 81 linearly spaced values
Eigenvalue tolerance ε_{eig}	10^{-9}
Hatched cells (contraction ruled out)	$\widehat{\tau}_E^\star(\lambda) > 1 + 10^{-6}$

We then form

$$C_\varepsilon := \Delta_\varepsilon^{-1} P^\top B P \Delta_\varepsilon^{-1}.$$

This gives the truncated reduced problem, and we define the corresponding numerical estimate by

$$\widehat{\tau}_S^\star(\lambda) := \lambda_{\max}(C_\varepsilon).$$

If no eigenvalues are discarded, then $\widehat{\tau}_S^\star(\lambda) = \tau_S^\star(\lambda)$. Otherwise, let V be the space spanned by the eigenvectors of A that are kept. Then

$$\widehat{\tau}_S^\star(\lambda) = \max_{v \in V,\, v \neq 0} \frac{v^\top B v}{v^\top A v} \leq \max_{v \neq 0} \frac{v^\top B v}{v^\top A v} = \tau_S^\star(\lambda),$$

since the first maximum is taken over fewer vectors. So the estimate can only underestimate the threshold.

D.3 Experimental setup

Motivated by Proposition C.1, we take

$$E = B(0, R) := \{x \in \mathbb{R}^d : \|x\| \leq R\}.$$

Fix a support size N and a trial count J. For each $j \in \{1, \dots, J\}$, we draw an independent random support

$$S_j = \{z_1^{(j)}, \dots, z_N^{(j)}\} \subseteq B(0, R), \qquad z_i^{(j)} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\big(B(0, R)\big).$$

Each sampled support S_j induces a threshold $\tau_{S_j}^\star(\lambda)$. Our implementation computes instead the numerical approximation

$$\widehat{\tau}_{S_j}^\star(\lambda) \approx \tau_{S_j}^\star(\lambda)$$

obtained from the truncated symmetric problem described above. We report the largest of these values,

$$\widehat{\tau}_E^\star(\lambda) := \max_{1 \leq j \leq J} \widehat{\tau}_{S_j}^\star(\lambda),$$

which estimates the lower bound $\max_j \tau_{S_j}^\star(\lambda)$ on $\tau_E^\star(\lambda)$ of Section D.1. Algorithm 1 summarizes this process, and Table 2 gives the numerical settings used for the heatmaps of Figure 2.

Compute resources. The experiments of this section were implemented in JAX [8] and run on CPU using float64 precision. For a fixed kernel configuration, computing one contraction curve took at most one minute.

D.4 Results

Verification on known kernels. The method introduced in this section was designed to estimate a lower bound on $\tau_E^\star(\lambda)$, where $E = B(0, R)$. As R increases, the set E approaches the ambient space, so the estimate on $B(0, R)$ should approach the smallest contraction factor on all of $\mathbb{R}^d$,

$$\lim_{R \to \infty} \tau_{B(0,R)}^\star(\lambda) = \tau_{\mathbb{R}^d}^\star(\lambda).$$

Algorithm 1 Monte Carlo sampling of $\widehat{\tau}^{\star}_{S_j}(\lambda)$ from CPD tests on finite supports

Require: Dimension $d \geq 1$, support size $N \geq 2$, trial count $J \geq 1$
Require: Radius $R > 0$ defining $E = B(0, R) \subset \mathbb{R}^d$
Require: Scaling parameter $\lambda \in (0, 1)$
Require: Symmetric kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$
Require: Relative tolerance $\varepsilon_{\text{eig}} > 0$
Output: Samples $\{\widehat{\tau}^{\star}_{S_j}(\lambda)\}_{j=1}^{J}$
Compute an orthonormal basis $Q \in \mathbb{R}^{N \times (N-1)}$ of $\mathbf{1}_N^{\perp}$
for $j = 1, \ldots, J$ **do**
    Sample a support and build the corresponding restricted matrices.
    Draw $S_j = \{z_1^{(j)}, \ldots, z_N^{(j)}\}$ with $z_i^{(j)} \overset{\text{i.i.d.}}{\sim} \text{Unif}(B(0, R))$
    Form the Gram matrices

$$K_{ab} \leftarrow k\Big(z_a^{(j)}, z_b^{(j)}\Big), \qquad (K_\lambda)_{ab} \leftarrow k\Big(\lambda z_a^{(j)}, \lambda z_b^{(j)}\Big)$$

    Restrict to $\mathbf{1}_N^{\perp}$

$$A \leftarrow Q^\top K Q, \qquad B \leftarrow Q^\top K_\lambda Q$$

    Reduce the generalized eigenvalue problem to a truncated ordinary symmetric one.
    Compute an eigendecomposition

$$A = P \,\text{diag}(\mu_1, \ldots, \mu_{N-1})\, P^\top$$

    Set

$$\mu_{\max} \leftarrow \max_{1 \leq i \leq N-1} \mu_i$$

    Set the truncated diagonal matrix

$$\Delta_\varepsilon^{-1} \leftarrow \text{diag}(\nu_1, \ldots, \nu_{N-1}), \qquad \nu_i = \begin{cases} \mu_i^{-1/2}, & \mu_i > \varepsilon_{\text{eig}}\, \mu_{\max}, \\ 0, & \text{otherwise} \end{cases}$$

    Form the reduced symmetric matrix

$$C_\varepsilon \leftarrow \Delta_\varepsilon^{-1} P^\top B P \Delta_\varepsilon^{-1}$$

    Extract the threshold estimate from the largest eigenvalue.
    Compute

$$\widehat{\tau}^{\star}_{S_j}(\lambda) \leftarrow \lambda_{\max}(C_\varepsilon)$$

end for
Return: $\{\widehat{\tau}^{\star}_{S_j}(\lambda)\}_{j=1}^{J}$

To validate Algorithm 1, we apply it to the power-distance and generalized multiquadric kernels, whose smallest contraction factor across dimensions $\tau^\star(\lambda)$ is known (Propositions F.1 and F.2) and bounds $\tau^\star_{\mathbb{R}^d}(\lambda)$ from above. Figure 4 shows the estimates obtained on $B(0, R)$ for increasing values of R. For the power-distance kernels, the estimates already match $\tau^\star(\lambda)$. For the generalized multiquadric kernels, the estimates are smaller on small balls and approach $\tau^\star(\lambda)$ as R increases. This is consistent with the fact that CPD on $B(0, R)$ is weaker than CPD on the whole ambient space.

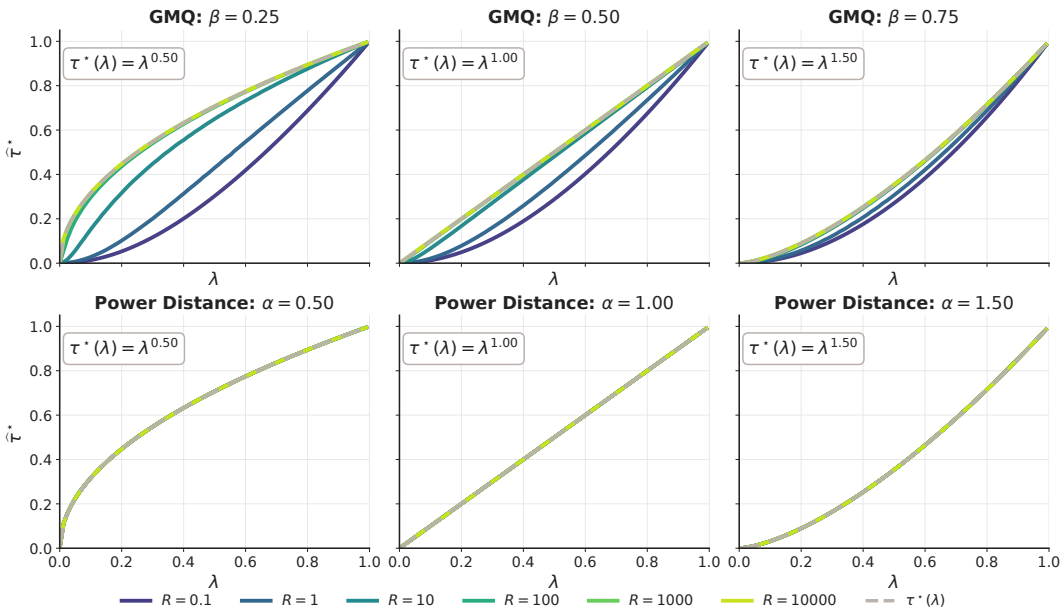

Figure 4: Estimates on balls $B(0, R)$ for kernels with known smallest contraction factors across dimensions. As R increases, the estimates approach the theoretical factors.

E Global contraction under radial kernels via Bernstein functions

Preamble. This section studies contraction under radial kernels. It turns the CPD criterion of Appendix C into conditions on the Bernstein function that generates the kernel.

Section E.1 explains why we focus on radial kernels and shows that contraction under an affine map then reduces to a CND condition on the difference profile (Proposition E.2). Section E.2 collects the necessary background, namely the definitions of Bernstein and complete Bernstein functions and Schoenberg's theorem (Theorem E.1), by which a radial kernel is CND exactly when its profile is a Bernstein function. In Section E.3, Theorem E.2 shows that contraction under affine maps, in every dimension, holds exactly when the difference profile $\psi_\tau = \tau\phi - \phi(\lambda^2 \cdot)$ is a Bernstein function, and Proposition E.3 shows that saturating profiles cannot yield strict contraction.

Section E.4 proves the results on Lévy measures needed to make the Bernstein criterion explicit. Section E.5 uses them to express the criterion as pointwise inequalities on the drift and the Lévy measure (Theorem E.4). Section E.6 treats complete Bernstein functions, for which the smallest contraction factor across dimensions reduces to a single supremum of the Lévy density ratio (Theorem E.5). Section E.7 introduces geometric chaining, which builds Bernstein profiles that induce contraction with a chosen contraction factor (Theorem E.6).

E.1 Motivations for radial kernels

For scalar returns ($d = 1$), the following proposition shows that a symmetric kernel that is invariant under translations is radial. Hence, in one dimension, focusing on radial kernels covers every translation-invariant kernel.

Proposition E.1 (A symmetric, translation-invariant univariate kernel is radial)**.** Let $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

(i) **Translation invariance:** $k(x + t, y + t) = k(x, y)$ for all $x, y, t \in \mathbb{R}$.

(ii) **Symmetry:** $k(x, y) = k(y, x)$ for all $x, y \in \mathbb{R}$.

Then (i) and (ii) hold if and only if there exists a function $g : [0, \infty) \to \mathbb{R}$ such that

$$k(x, y) = g(|x - y|) \qquad \forall\, x, y \in \mathbb{R}.$$

Proof.

(⇒) Necessity. Assume translation invariance. Define $\kappa : \mathbb{R} \to \mathbb{R}$ by $\kappa(r) := k(r, 0)$. Then for any $x, y \in \mathbb{R}$, applying translation invariance with $t = -y$ gives

$$k(x, y) = k(x - y, 0) = \kappa(x - y).$$

Assume also that k is symmetric. Then for all $r \in \mathbb{R}$,

$$\kappa(r) = k(r, 0) = k(0, r) = k(-r, 0) = \kappa(-r),$$

where the step $k(0, r) = k(-r, 0)$ uses translation invariance with $t = -r$. Thus κ is even. Define $g : [0, \infty) \to \mathbb{R}$ by $g(u) := \kappa(u)$ for $u \geq 0$. Since κ is even, $\kappa(r) = g(|r|)$ for all $r \in \mathbb{R}$, and therefore

$$k(x, y) = \kappa(x - y) = g(|x - y|).$$

(⇐) Sufficiency. Conversely, if $k(x, y) = g(|x - y|)$, then for any $t \in \mathbb{R}$,

$$k(x + t, y + t) = g(|(x + t) - (y + t)|) = g(|x - y|) = k(x, y),$$

so k is translation invariant. Also,

$$k(y, x) = g(|y - x|) = g(|x - y|) = k(x, y),$$

so k is symmetric. □

Proposition E.2 (For radial kernels, contraction under an affine map reduces to a CND condition)**.** Fix $d \geq 1$. Assume $f\colon \mathbb{R}^d \to \mathbb{R}^d$ is an affine map of the form

$$f(x) = q + \lambda x, \qquad q \in \mathbb{R}^d,\ \lambda \in \mathbb{R}.$$

Let $k(u, v) = g(\|u - v\|)$ be a radial kernel on $\mathbb{R}^d$, and define

$$h(r) = g(0) - g(r), \quad r \geq 0.$$

As in Theorem C.1, assume k is CPD on $\mathbb{R}^d$, so that $(u, v) \mapsto h(\|u - v\|)$ is CND on $\mathbb{R}^d$.

Then, for any $\tau \in \mathbb{R}$, the contraction inequality

$$\mathbf{MMD}_k^2\big(f\#\nu, f\#\mu\big) \ \leq\ \tau\, \mathbf{MMD}_k^2(\nu, \mu)$$

holds for every $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ if and only if

$$\Gamma_\tau(x, y) = \tau\, h\big(\|x - y\|\big) - \ h\big(|\lambda|\, \|x - y\|\big) \text{ is CND on } \mathbb{R}^d.$$

Moreover, $\Gamma_\tau(x, x) = 0$.

Proof. Since $f(x) = q + \lambda x$, we have

$$\|f(x) - f(y)\| = |\lambda|\, \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^d.$$

By Theorem C.1, applied with $E = \mathbb{R}^d$, the contraction statement

$$\mathbf{MMD}_k^2\big(f\#\nu, f\#\mu\big) \ \leq\ \tau\, \mathbf{MMD}_k^2(\nu, \mu) \qquad \forall\, \mu, \nu \in \mathcal{P}(\mathbb{R}^d)$$

is equivalent to Γ_τ being CND on $\mathbb{R}^d$, where

$$\Gamma_\tau(x, y) = \tau\, h\big(\|x - y\|\big) - \ h\big(\|f(x) - f(y)\|\big).$$

Substituting $\|f(x) - f(y)\| = |\lambda|\, \|x - y\|$ yields

$$\Gamma_\tau(x, y) = \tau\, h\big(\|x - y\|\big) - \ h\big(|\lambda|\, \|x - y\|\big).$$

Finally, $h(0) = g(0) - g(0) = 0$, hence

$$\Gamma_\tau(x, x) = \tau\, h(0) - h(0) = 0.$$

□

E.2 Definitions and Schoenberg's theorem

Definition E.1 (Bernstein function)**.** A function $\phi\colon [0, \infty) \to [0, \infty)$ is called a *Bernstein function* [29, Def. 3.1] if it is continuous at 0, infinitely differentiable on $(0, \infty)$, and its derivative is completely monotone,

$$\phi \in C^\infty((0, \infty)), \quad (-1)^{n-1}\, \frac{d^n}{dt^n}\phi(t) \ \geq\ 0 \quad \forall\, n \geq 1,\ t > 0.$$

Equivalently, it admits the Lévy–Khintchine representation [29, Thm. 3.2]

$$\phi(t) = a + b\, t + \int_{(0,\infty)} \big(1 - e^{-t\, s}\big)\, \Lambda(ds), \qquad a \geq 0,\ b \geq 0,$$

together with the integrability condition

$$\int_{(0,\infty)} \min\{1, s\}\, \Lambda(ds) < \infty,$$

and moreover, by [29, Rem. 3.3(iv)], one has

$$a = \phi(0+), \qquad b = \lim_{t \to \infty} \frac{\phi(t)}{t}.$$

Thus b, called the *drift* of ϕ, determines whether ϕ has linear or sublinear growth.

Definition E.2 (Complete Bernstein function)**.** A Bernstein function ϕ (Definition E.1) is called a *complete Bernstein function* [29, Def. 6.1] if its Lévy measure Λ is absolutely continuous with a completely monotone density m,

$$\Lambda(ds) = m(s)\, ds,$$

where

$$m \in C^\infty((0,\infty)), \quad (-1)^n\, m^{(n)}(s) \;\geq\; 0 \quad \forall\, n \geq 0,\ s > 0.$$

Theorem E.1 (Schoenberg's characterization of radial CND kernels, restated via Bernstein functions [30])**.** Let $h\colon [0,\infty) \to [0,\infty)$ be continuous with $h(0) = 0$, and define the radial kernel

$$k(x,y) = h\big(\|x-y\|\big), \quad x, y \in \mathbb{R}^d.$$

Then k is conditionally negative definite on $\mathbb{R}^d$ for *every* $d \geq 1$ if and only if there exists a Bernstein function ϕ such that

$$h(r) = \phi(r^2), \quad r \geq 0.$$

In particular, ϕ admits a Lévy–Khintchine representation

$$\phi(t) = a + b\,t + \int_{(0,\infty)} \big(1 - e^{-s\,t}\big)\, \Lambda(ds), \quad a, b \geq 0.$$

Moreover, since $h(0) = 0$, evaluating at $r = 0$ gives $0 = h(0) = \phi(0) = a$, which in turn yields

$$\boxed{\phi(t) = b\,t + \int_{(0,\infty)} \big(1 - e^{-s\,t}\big)\, \Lambda(ds), \quad h(r) = b\,r^2 + \int_{(0,\infty)} \big(1 - e^{-s\,r^2}\big)\, \Lambda(ds).}$$

E.3 Bernstein reduction of difference kernel

Preamble. In this subsection, the contraction condition on the difference kernel is translated into a condition on a Bernstein function. The key point is that, in order for Schoenberg's theorem to apply, the relevant radial kernel must be conditionally negative definite on $\mathbb{R}^d$ for every $d \geq 1$, not only in one fixed dimension. Theorem E.2 shows that, when this holds in every dimension, the contraction condition is equivalent to the difference profile being a Bernstein function. Theorem E.3 then rewrites this criterion directly in terms of the drift term and the Lévy measure in the Lévy–Khintchine representation. Finally, Proposition E.3 shows that saturating profiles cannot yield strict MMD contraction under affine maps.

Theorem E.2 (For radial kernels, contraction in every dimension reduces to a Bernstein condition)**.** Fix $\lambda \in \mathbb{R}$. Let $g : [0,\infty) \to \mathbb{R}$ be continuous, and assume that, for every $d \geq 1$, the radial kernel

$$k(u,v) = g(\|u-v\|), \qquad u, v \in \mathbb{R}^d,$$

is CPD on $\mathbb{R}^d$. Define

$$h(r) = g(0) - g(r), \qquad r \geq 0.$$

Then, for every $d \geq 1$, the kernel $(u,v) \mapsto h(\|u-v\|)$ is CND on $\mathbb{R}^d$. Hence, by Theorem E.1, there exists a Bernstein function ϕ such that

$$h(r) = \phi(r^2), \qquad r \geq 0.$$

For each $d \geq 1$ and $q \in \mathbb{R}^d$, consider the affine map

$$f(x) = q + \lambda x, \qquad x \in \mathbb{R}^d.$$

In the first statement below, the kernel k is understood as the above radial kernel on the ambient space $\mathbb{R}^d$. Then, for any $\tau \in \mathbb{R}$, the following are equivalent.

$$\boxed{\begin{array}{c} \mathbf{MMD}_k^2\big(f\#\nu, f\#\mu\big) \;\leq\; \tau\, \mathbf{MMD}_k^2(\nu,\mu) \\ \text{for every } d \geq 1 \text{ and every } \mu,\nu \in \mathcal{P}(\mathbb{R}^d) \\ \text{if and only if} \\ \Gamma_\tau(x,y) = \tau\, h(\|x-y\|) - \; h(|\lambda|\,\|x-y\|) \text{ is CND} \\ \text{on } \mathbb{R}^d \text{ for every } d \geq 1 \\ \text{if and only if} \\ \psi_\tau(t) = \tau\,\phi(t) - \phi(\lambda^2 t) \text{ is a Bernstein function.} \end{array}}$$

Proof. Since, for every $d \geq 1$, the kernel

$$k(u, v) = g(\|u - v\|)$$

is CPD on $\mathbb{R}^d$, the kernel

$$(u, v) \mapsto h(\|u - v\|) = g(0) - g(\|u - v\|)$$

is CND on $\mathbb{R}^d$ for every $d \geq 1$. Theorem E.1 therefore yields a Bernstein function ϕ such that

$$h(r) = \phi(r^2), \qquad r \geq 0.$$

Fix $\tau \in \mathbb{R}$, and let $d \geq 1$ and $q \in \mathbb{R}^d$. By Proposition E.2, applied in the fixed ambient dimension d to the affine map $f(x) = q + \lambda x$, the first statement is equivalent to

$$\Gamma_\tau(x, y) = \tau\, h(\|x - y\|) - \ h(|\lambda|\, \|x - y\|)$$

being CND on $\mathbb{R}^d$. As $d \geq 1$ was arbitrary, the first two statements are equivalent.

Finally, using $h(r) = \phi(r^2)$, we may rewrite

$$\Gamma_\tau(x, y) = \tau\, \phi(\|x - y\|^2) - \phi(\lambda^2 \|x - y\|^2).$$

Thus $\Gamma_\tau(x, y)$ depends on x and y only through $\|x - y\|^2$. We therefore define the difference profile $\psi_\tau : [0, \infty) \to \mathbb{R}$ by

$$\psi_\tau(t) = \tau\, \phi(t) - \phi(\lambda^2 t).$$

With this notation,

$$\Gamma_\tau(x, y) = \psi_\tau(\|x - y\|^2).$$

Applying Theorem E.1 once more, Γ_τ is CND on $\mathbb{R}^d$ for every $d \geq 1$ if and only if ψ_τ is a Bernstein function. □

Theorem E.3 (For the difference kernel, CND in every dimension reduces to a Lévy–Khintchine criterion)**.** Let $\phi : [0, \infty) \to [0, \infty)$ be a Bernstein function with $\phi(0) = 0$, admitting the Lévy–Khintchine representation

$$\phi(t) = \ b\, t \ + \ \int_0^\infty \big(1 - e^{-s\, t}\big)\, \Lambda(ds),$$

where $b \geq 0$ and Λ is a nonnegative Borel measure on $(0, \infty)$. Fix $\lambda \in \mathbb{R} \setminus \{0\}$ and $\tau \in \mathbb{R}$, and define the difference profile ψ_τ and the kernel Γ_τ built from it by

$$\psi_\tau(t) := \tau\, \phi(t) - \phi(\lambda^2 t), \qquad \Gamma_\tau(x, y) = \psi_\tau\big(\|x - y\|^2\big), \qquad x, y \in \mathbb{R}^d.$$

Then

(i) Γ_τ is CND on $\mathbb{R}^d$ for every $d \geq 1$,
if and only if
(ii) $b(\tau - \lambda^2) \ \geq \ 0$ and $\tau\, \Lambda(B) \ \geq \ \Lambda\big(B/\lambda^2\big)$ for every Borel set $B \subset (0, \infty)$.

Proof.

(i)⇒(ii). Assume Γ_τ is CND on $\mathbb{R}^d$ for every $d \geq 1$. Using the Lévy–Khintchine representation of ϕ,

$$\psi_\tau(t) = \underbrace{b(\tau - \lambda^2)}_{\text{drift}}\, t + \int_{(0,\infty)} \big(1 - e^{-st}\big)\ \underbrace{[\tau\, \Lambda - \Lambda_\lambda]}_{\text{Lévy measure}}(ds), \qquad \Lambda_\lambda(B) := \Lambda(B/\lambda^2).$$

Since Γ_τ is CND on $\mathbb{R}^d$ for every $d \geq 1$, Theorem E.1 implies that ψ_τ is Bernstein. Its drift and its Lévy measure are therefore nonnegative, hence

$$b(\tau - \lambda^2) \geq 0 \quad \text{and} \quad \tau\, \Lambda(B) \geq \Lambda(B/\lambda^2) \ \text{ for every Borel } B \subset (0, \infty).$$

(ii)⇒(i). Assume

$$b(\tau - \lambda^2) \geq 0 \quad \text{and} \quad \tau\,\Lambda(B) \geq \Lambda(B/\lambda^2) \ \text{ for every Borel } B \subset (0,\infty).$$

Then the same Lévy–Khintchine representation shows that

$$\psi_\tau(t) = \tau\phi(t) - \phi(\lambda^2 t)$$

is a Bernstein function. Applying Schoenberg's theorem again shows that

$$\Gamma_\tau(x,y) = \psi_\tau\big(\|x-y\|^2\big)$$

is CND on $\mathbb{R}^d$ for every $d \geq 1$. □

Proposition E.3 (A saturating radial profile cannot induce contraction)**.** Adopt the setting of Theorem E.2. Fix $d \geq 1$ and $\lambda \in (0,1)$, and define the affine map

$$f(x) = q + \lambda x, \qquad x \in \mathbb{R}^d.$$

Let $\phi : [0,\infty) \to [0,\infty)$ satisfy $\phi(0) = 0$ and define the radial kernel

$$k(x,y) = \phi(0) - \phi(\|x-y\|^2) = -\phi(\|x-y\|^2).$$

Assume that ϕ is *saturating*, meaning that $\lim_{t\to\infty} \phi(t) < \infty$. Then, for every $\tau \in (0,1)$, no strict $\mathbf{MMD}^2$ contraction under f can hold unless $\phi \equiv 0$. More precisely, for any $\tau \in (0,1)$, define the difference profile ψ_τ and the kernel Γ_τ built from it by

$$\psi_\tau(t) = \tau\,\phi(t) - \phi(\lambda^2 t), \qquad \Gamma_\tau(x,y) = \psi_\tau(\|x-y\|^2).$$

If $\phi \not\equiv 0$ and ϕ is saturating, then Γ_τ is *not* conditionally negative definite on $\mathbb{R}^d$. As a consequence, there is no $\tau \in (0,1)$ such that

$$\mathbf{MMD}_k^2(f_\#\nu, f_\#\mu) \ \leq \ \tau\,\mathbf{MMD}_k^2(\nu,\mu)$$

holds for all $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$.

Proof. Fix $\tau \in (0,1)$. If Γ_τ were conditionally negative definite on $\mathbb{R}^d$, then for every choice of two points $x_1, x_2 \in \mathbb{R}^d$ and weights $(w_1, w_2) = (1,-1)$ with $w_1 + w_2 = 0$, we would have

$$\sum_{i,j=1}^{2} w_i w_j\,\Gamma_\tau(x_i,x_j) \ \leq \ 0.$$

Let $r = \|x_1 - x_2\|$. Since $\Gamma_\tau(x_1,x_2) = \psi_\tau(r^2)$ and $\Gamma_\tau(x_1,x_1) = \Gamma_\tau(x_2,x_2) = \psi_\tau(0)$, a direct expansion gives

$$\sum_{i,j=1}^{2} w_i w_j\,\Gamma_\tau(x_i,x_j) = 2\psi_\tau(0) - 2\psi_\tau(r^2).$$

Thus CND would imply

$$\psi_\tau(r^2) \ \geq \ \psi_\tau(0) \qquad \text{for all } r \geq 0.$$

Since $\phi(0) = 0$, we have $\psi_\tau(0) = \tau\phi(0) - \phi(0) = 0$, hence

$$\psi_\tau(t) \ \geq \ 0 \qquad \text{for all } t \geq 0,$$

that is,

$$\tau\,\phi(t) \ \geq \ \phi(\lambda^2 t) \qquad \text{for all } t \geq 0.$$

Now assume ϕ is saturating and $\phi \not\equiv 0$. Then the limit

$$L := \lim_{t\to\infty} \phi(t)$$

exists and satisfies $L \in (0,\infty)$. Letting $t \to \infty$ in $\tau\phi(t) \geq \phi(\lambda^2 t)$ yields

$$\tau L \ \geq \ L.$$

Since $L > 0$, this forces $\tau \geq 1$, contradicting $\tau \in (0,1)$. Therefore, if $\tau \in (0,1)$ and ϕ is saturating with $\phi \not\equiv 0$, the kernel Γ_τ cannot be CND on $\mathbb{R}^d$.

Finally, by Proposition E.2, strict $\mathbf{MMD}^2$ contraction under the map $f(x) = q + \lambda x$ with factor $\tau \in (0,1)$ would require Γ_τ to be CND on $\mathbb{R}^d$. This is impossible when ϕ is saturating and $\phi \not\equiv 0$, so no such strict contraction can hold unless $\phi \equiv 0$. □

E.4 Results on measures and the measure domination condition

Preamble. This subsection studies the condition

$$\tau\,\Lambda(B) \;\geq\; \Lambda(B/\rho) \qquad \text{for all Borel sets } B \subset (0,\infty),$$

with $\rho = \lambda^2$. Lemma E.2 reduces it to two pointwise inequalities, one on the density $m(s)$ and one on the mass $\Lambda(\{x\})$ at every point $x > 0$. Lemma E.1 shows that a Lévy measure has at most countably many atoms. The remaining results describe the atoms of a measure that satisfies this condition. Lemma E.3 shows that a measure with at least one but finitely many atoms cannot satisfy the condition, so a measure that satisfies it has either no atoms or infinitely many. Lemma E.4 shows that a single geometric chain can occur, and Lemma E.5 partitions $(0,\infty)$ into disjoint geometric chains $\{\rho^n r : n \in \mathbb{Z}\}$, indexed by their representative $r \in (\rho, 1]$. Proposition E.4 combines these results. When $\tau \leq \rho$, Λ has no atoms, and when $\tau > \rho$, atoms can exist but the set of atoms must be a union of disjoint tails of geometric chains.

Lemma E.1 (Atoms of a Lévy measure are countable)**.** Let Λ be a Lévy measure on $(0,\infty)$, that is,

$$\int_{(0,\infty)} \min\{1, s\}\,\Lambda(ds) < \infty.$$

Then the set of atoms $\{x > 0 : \Lambda(\{x\}) > 0\}$ is at most countable.

Proof. For each $n \in \mathbb{N}$ let $E_n = (1/n, \infty)$. By the Lévy condition,

$$\int_{(1,\infty)} 1\,\Lambda(ds) \;=\; \int_{(1,\infty)} \min\{1, s\}\,\Lambda(ds) \;<\; \infty,$$

so $\Lambda((1,\infty)) < \infty$. Moreover,

$$\int_{(1/n,1]} 1\,\Lambda(ds) \;\leq\; n\int_{(1/n,1]} s\,\Lambda(ds) \;\leq\; n\int_{(0,1]} s\,\Lambda(ds) \;\leq\; n\int_{(0,\infty)} \min\{1, s\}\,\Lambda(ds) \;<\; \infty,$$

hence $\Lambda((1/n, 1]) < \infty$. Therefore $\Lambda(E_n) < \infty$ for every n.

Fix n. We show that the set of atoms in E_n is countable by grouping atoms according to a lower bound on their mass. For each $\ell \in \mathbb{N}$, define

$$A_{n,\ell} \;:=\; \{\, x \in E_n : \Lambda(\{x\}) \geq 1/\ell \,\}.$$

Each $A_{n,\ell}$ contains the atoms in E_n whose mass is at least $1/\ell$.

We claim that $A_{n,\ell}$ is finite for every ℓ. Indeed, if $A_{n,\ell}$ had infinitely many points, then for any $M \in \mathbb{N}$ we could choose distinct $x_1, \ldots, x_M \in A_{n,\ell}$. Since the singletons $\{x_i\}$ are disjoint and contained in E_n,

$$\Lambda(E_n) \;\geq\; \sum_{i=1}^{M} \Lambda(\{x_i\}) \;\geq\; M \cdot \tfrac{1}{\ell}.$$

Since M is arbitrary, letting $M \to \infty$ forces $\Lambda(E_n) = \infty$, contradicting $\Lambda(E_n) < \infty$. Hence each $A_{n,\ell}$ is finite.

Now let $x \in E_n$ be an atom with $\Lambda(\{x\}) = \alpha > 0$. Choose $\ell > \frac{1}{\alpha}$. Then $\Lambda(\{x\}) \geq 1/\ell$, so $x \in A_{n,\ell}$. Therefore every atom in E_n belongs to some $A_{n,\ell}$, and

$$\{x \in E_n : \Lambda(\{x\}) > 0\} \;=\; \bigcup_{\ell=1}^{\infty} A_{n,\ell}.$$

Since this is a countable union of finite sets, the set of atoms in E_n is countable. Finally,

$$\{x > 0 : \Lambda(\{x\}) > 0\} \;=\; \bigcup_{n=1}^{\infty} \{x \in E_n : \Lambda(\{x\}) > 0\}$$

is a countable union of countable sets, hence at most countable.

□

Lemma E.2 (The measure domination condition reduces to an inequality on the density and one on the atoms)**.** Let Λ be a nonnegative Borel measure on $(0,\infty)$, *locally finite* in the sense that $\Lambda([a,b]) < \infty$ for all $0 < a < b$, admitting the decomposition

$$\Lambda \;=\; \underbrace{m(s)\,ds}_{\text{absolutely continuous part}} \;+\; \underbrace{\sum_{j\ge 1} c_j\,\delta_{s_j}}_{\text{atomic part}} \quad \text{(no singular continuous part)},$$

Here $m(s) \ge 0$ for almost every $s > 0$, and each $c_j > 0$. Since Λ is locally finite, m is locally integrable on $(0,\infty)$, that is, $\int_a^b m(s)\,ds \le \Lambda([a,b]) < \infty$ for all $0 < a < b$.

Fix $\lambda \in (0,1]$, set $\rho = \lambda^2$, and define, for every Borel set $B \subset (0,\infty)$,

$$\Lambda_\lambda(B) \;:=\; \Lambda\big(B/\rho\big), \qquad B/\rho = \{\, s > 0 : \rho s \in B \,\}.$$

Let $\tau \in (0,1)$. Then

$$\boxed{\begin{array}{c} (i)\quad \tau\,\Lambda(B) \;\ge\; \Lambda_\lambda(B) \quad \text{for every Borel set } B \subset (0,\infty), \\ \text{if and only if} \\ (ii)\quad \begin{cases} (1)\quad \tau\,\rho\,m(s) \;\ge\; m\big(s/\rho\big) & \text{for almost every } s > 0, \\ (2)\quad \tau\,\Lambda(\{x\}) \;\ge\; \Lambda\big(\{x/\rho\}\big) & \text{for every } x > 0. \end{cases} \end{array}}$$

Proof.

Necessity (i) ⇒ (ii).

1. The inequality on the atoms. For any $x > 0$, take $B = \{x\}$. Since $\{x\}/\rho = \{x/\rho\}$,

$$\Lambda_\lambda(\{x\}) = \Lambda(\{x/\rho\}).$$

Thus the hypothesis $\tau\,\Lambda(B) \ge \Lambda_\lambda(B)$ gives

$$\tau\,\Lambda(\{x\}) \;\ge\; \Lambda(\{x/\rho\}) \qquad \text{for all } x > 0.$$

2. The inequality on the density. Let $S = \{x > 0 : \Lambda(\{x\}) > 0\} = \{s_j : j \ge 1\}$ be the set of atoms, which is countable. Take a Borel set $B \subset (0,\infty)$ that contains no point of S and no point of $\rho S = \{\rho x : x \in S\}$. Then neither B nor B/ρ contains an atom, so only the absolutely continuous part contributes, and we have

$$\Lambda(B) = \int_B m(s)\,ds, \qquad \Lambda_\lambda(B) = \Lambda(B/\rho) = \int_{B/\rho} m(s)\,ds.$$

Hence the assumption gives

$$\tau \int_B m(s)\,ds \;\ge\; \int_{B/\rho} m(s)\,ds.$$

Perform the change of variables $u = \rho s$, so $s = u/\rho$ and $ds = \frac{1}{\rho}\,du$. Then

$$\int_{B/\rho} m(s)\,ds = \int_{\{s:\rho s\in B\}} m(s)\,ds = \frac{1}{\rho}\int_B m(u/\rho)\,du.$$

Thus for every Borel set B that contains no point of $S \cup \rho S$,

$$\tau \int_B m(s)\,ds \;\ge\; \frac{1}{\rho}\int_B m(s/\rho)\,ds.$$

Finally, choose B to be a small interval around a point $s > 0$, minus the countably many points of $S \cup \rho S$. A countable set has length zero, so this changes neither $|B|$ nor the integrals of m. Divide

both sides by $|B|$, and let the interval shrink to $\{s\}$. By the Lebesgue differentiation theorem, in the limit one obtains

$$\tau\, m(s) \;\geq\; \frac{1}{\rho}\, m(s/\rho) \quad \text{for almost every } s > 0,$$

or equivalently

$$\tau\, \rho\, m(s) \;\geq\; m(s/\rho).$$

Sufficiency (ii) $\Rightarrow$ (i). Assume

$$\tau\, \rho\, m(s) \;\geq\; m(s/\rho) \quad \text{almost everywhere,} \qquad \tau\, \Lambda(\{x\}) \;\geq\; \Lambda(\{x/\rho\}) \quad \forall\, x > 0.$$

Set

$$\Delta\Lambda \;=\; \tau\, \Lambda \;-\; \Lambda_\lambda \;=\; \big[\tau\, m(s) - \tfrac{1}{\rho}\, m(s/\rho)\big]\, ds \;+ \sum_{\substack{x>0:\\ \Lambda(\{x\})+\Lambda(\{x/\rho\})>0}} \big[\tau\, \Lambda(\{x\}) - \Lambda(\{x/\rho\})\big]\, \delta_x.$$

Here the sum runs over the countably many points x such that x or x/ρ is an atom of Λ, and each such x carries the mass $\tau\, \Lambda(\{x\}) - \Lambda(\{x/\rho\})$.

1. Absolutely continuous part. By assumption (ii.1), $\tau m(s) - \frac{1}{\rho} m(s/\rho) \geq 0$ almost everywhere, so the absolutely continuous part of $\Delta\Lambda$ is nonnegative.

2. Atomic part. By assumption (ii.2), each mass $\tau\, \Lambda(\{x\}) - \Lambda(\{x/\rho\})$ is nonnegative, so all atoms of $\Delta\Lambda$ are nonnegative.

Hence $\Delta\Lambda$ is a nonnegative measure and $\tau\, \Lambda = \Lambda_\lambda + \Delta\Lambda \geq \Lambda_\lambda$. □

Lemma E.3 (A measure with at least one but finitely many atoms cannot satisfy the measure domination condition)**.** Fix $\lambda \in (0, 1)$ and $\tau \in (0, 1)$, and let $\rho = \lambda^2$. Assume that a Borel measure Λ on $(0, \infty)$ satisfies the measure domination condition

$$\tau\, \Lambda(B) \;\geq\; \Lambda_\lambda(B) \quad \text{for every Borel set } B \subset (0, \infty), \qquad \Lambda_\lambda(B) := \Lambda(B/\rho).$$

Assume in addition that Λ has finitely many atoms, that is, $\Lambda \;=\; m(s)\, ds + \sum_{j=1}^{N} c_j\, \delta_{s_j}$ with $m(s) \geq 0$ and $c_j \geq 0$. Then Λ has no atoms.

Proof. Suppose for contradiction that Λ has an atom at some $s^\star > 0$, and let $c^\star := \Lambda(\{s^\star\}) > 0$. To use the measure domination condition, we choose the singleton $B = \{\rho s^\star\}$, since $B/\rho = \{s^\star\}$, so the right-hand side is exactly the mass of the atom at $s^\star$. Indeed, by definition of Λ_λ,

$$\Lambda_\lambda(\{\rho s^\star\}) = \Lambda(\{s^\star\}) = c^\star.$$

Applying the measure domination condition with $B = \{\rho s^\star\}$ gives

$$\tau\, \Lambda(\{\rho s^\star\}) \;\geq\; \Lambda_\lambda(\{\rho s^\star\}) \;=\; c^\star,$$

so $\Lambda(\{\rho s^\star\}) > 0$. Thus, for the measure domination condition to hold, $\rho s^\star$ must also be an atom of Λ. Iterating, we obtain

$$\Lambda(\{\rho^n s^\star\}) > 0 \qquad \text{for every } n \geq 1.$$

The locations $\rho^n s^\star$ are all distinct because $\rho \in (0, 1)$, so Λ has infinitely many atoms. This contradicts the assumption that Λ has finitely many atoms. Hence Λ has no atoms. □

Lemma E.4 (Atoms on a single geometric chain satisfy the measure domination condition)**.** Fix $\lambda \in (0, 1)$, set $\rho = \lambda^2$, and let $\tau \in (0, 1)$. Fix $s_0 > 0$ and $c > 0$, and define the measure

$$\Lambda \;=\; \sum_{n=0}^{\infty} a_n\, \delta_{s_n}, \qquad s_n = \rho^n s_0, \qquad a_n = c\, \tau^{-n}.$$

Define, for every Borel set $B \subset (0, \infty)$,

$$\Lambda_\lambda(B) \;:=\; \Lambda\big(B/\rho\big), \qquad B/\rho = \{\, s > 0 : \rho s \in B \,\}.$$

Then the measure domination condition

$$\tau\, \Lambda(B) \;\geq\; \Lambda_\lambda(B) \qquad \text{for every Borel set } B \subset (0, \infty)$$

holds.

Moreover, Λ is a Lévy measure if and only if $\tau > \rho$. In particular, when $\tau > \rho$, a single geometric chain gives an example satisfying the measure domination condition.

Proof.

Step 1: Verification of the measure domination condition. We first compute Λ_λ on the singletons that support the chain. For $n \geq 1$, the previous point of the chain is $s_{n-1} = s_n/\rho$, so

$$\Lambda_\lambda(\{s_n\}) = \Lambda(\{s_n\}/\rho) = \Lambda(\{s_n/\rho\}) = \Lambda(\{s_{n-1}\}) = a_{n-1} = \tau\, a_n = \tau\, \Lambda(\{s_n\}).$$

For $n = 0$, we instead have

$$\Lambda_\lambda(\{s_0\}) = \Lambda(\{s_0\}/\rho) = \Lambda(\{s_0/\rho\}) = 0,$$

since s_0/ρ does not belong to the support of Λ, whereas $\tau\, \Lambda(\{s_0\}) = \tau a_0 > 0$.

We now pass from these singleton identities to an arbitrary Borel set $B \subset (0, \infty)$. Since Λ is a sum of atoms,

$$\begin{aligned}
\Lambda_\lambda(B) &= \Lambda(B/\rho) \qquad \text{(definition of } \Lambda_\lambda) \\
&= \sum_{n=0}^{\infty} a_n\, \mathbf{1}_{B/\rho}(s_n) \qquad \big(\Lambda = \sum_{n=0}^{\infty} a_n\, \delta_{s_n}\big) \\
&= \sum_{n=0}^{\infty} a_n\, \mathbf{1}_B(\rho s_n) \qquad \big(\mathbf{1}_{B/\rho}(s_n) = \mathbf{1}_B(\rho s_n)\big) \\
&= \sum_{n=0}^{\infty} a_n\, \mathbf{1}_B(s_{n+1}) \qquad (\rho s_n = s_{n+1}) \\
&= \sum_{j=1}^{\infty} a_{j-1}\, \mathbf{1}_B(s_j) \qquad \text{(reindex } j = n+1) \\
&= \sum_{j=1}^{\infty} \tau\, a_j\, \mathbf{1}_B(s_j) \qquad (a_{j-1} = \tau a_j) \\
&\leq \tau \sum_{j=0}^{\infty} a_j\, \mathbf{1}_B(s_j) \qquad \text{(add the nonnegative } j = 0 \text{ term)} \\
&= \tau\, \Lambda(B).
\end{aligned}$$

Thus the measure domination condition holds for all Borel $B \subset (0, \infty)$. Moreover, equality holds whenever $s_0 \notin B$, since the only extra term added in the inequality above is the $j = 0$ contribution.

Step 2: Verification of the Lévy condition. By definition, Λ is a Lévy measure when it satisfies the Lévy integrability condition

$$\int_{(0,\infty)} \min\{1, s\}\, \Lambda(ds) < \infty.$$

Since Λ puts the mass a_n at s_n, this integral equals

$$\sum_{n \geq 0} a_n \min\{1, s_n\}.$$

Since $\rho < 1$, the atoms $s_n = \rho^n s_0$ decrease to 0, so there is an index K after which they are all at most 1, with $K = 0$ when $s_0 \leq 1$. For $n \geq K$, $\min\{1, s_n\} = s_n$. The sum above therefore differs from

$$\sum_{n \geq 0} a_n s_n = s_0 c \sum_{n \geq 0} \Big(\frac{\rho}{\tau}\Big)^n, \qquad \text{using } a_n s_n = c\, \tau^{-n} \rho^n s_0,$$

only in its first K terms, each of them finite, so one is finite exactly when the other is. This geometric series converges if and only if $\rho/\tau < 1$, that is, if and only if $\tau > \rho$. Therefore Λ is a Lévy measure if and only if $\tau > \rho$. □

Lemma E.5 (Every $x > 0$ admits a unique decomposition $x = \rho^n r$ with $r \in (\rho, 1]$)**.** Fix $\rho \in (0, 1)$. Then every $x > 0$ admits a unique decomposition

$$x = \rho^n r, \qquad n \in \mathbb{Z}, \quad r \in (\rho, 1],$$

given by

$$n = \left\lfloor \frac{\log x}{\log \rho} \right\rfloor, \qquad r = \rho^{-n} x.$$

Equivalently, each geometric chain $\{\rho^n x : n \in \mathbb{Z}\}$ meets the interval $(\rho, 1]$ in exactly one point. We write $n(x)$ and $r(x)$ for these n and r, and call $r(x)$ the *chain representative* of x and $n(x)$ its *chain index*.

Proof. Fix $x > 0$. For an integer n, set $r = \rho^{-n} x$, so that $x = \rho^n r$. Then $r \in (\rho, 1]$ means

$$\rho \;<\; \rho^{-n} x \;\le\; 1,$$

and multiplying by $\rho^n > 0$, this is equivalent to

$$\rho^{n+1} \;<\; x \;\le\; \rho^n,$$

that is, to x lying in the interval $(\rho^{n+1}, \rho^n]$. For $n \in \mathbb{Z}$, these intervals cover $(0, \infty)$ without overlapping,

$$\cdots, \qquad (\rho^2, \rho], \qquad (\rho, 1], \qquad (1, \rho^{-1}], \qquad (\rho^{-1}, \rho^{-2}], \qquad \cdots$$

To find the one that contains x, we take logarithms and divide by $\log \rho < 0$, which reverses the inequalities. The condition becomes

$$n \;\le\; \frac{\log x}{\log \rho} \;<\; n + 1.$$

Exactly one integer satisfies it, namely $n = \lfloor \log x / \log \rho \rfloor$. Hence x has exactly one decomposition $x = \rho^n r$ with $n \in \mathbb{Z}$ and $r \in (\rho, 1]$, given by $n(x) = \lfloor \log x / \log \rho \rfloor$ and $r(x) = \rho^{-n(x)} x$.

Finally, multiplying by ρ moves a point from one interval to the next one on its left, so the chain $\{\rho^n x : n \in \mathbb{Z}\}$ has exactly one point in each interval, and in particular exactly one in $(\rho, 1]$, namely $r(x)$. □

Proposition E.4 (The measure domination condition rules out atoms of the Lévy measure or forces them onto infinite geometric chains)**.** Let Λ be a Lévy measure on $(0, \infty)$, that is,

$$\int_{(0,\infty)} \min\{1, s\}\, \Lambda(ds) < \infty.$$

Assume Λ has no singular continuous part and admits the decomposition

$$\Lambda(ds) = m(s)\, ds + \sum_{j \ge 1} c_j\, \delta_{s_j}, \qquad m(s) \ge 0 \text{ for almost every } s > 0, \quad c_j \ge 0.$$

Fix $\lambda \in (0, 1)$, set $\rho = \lambda^2$, and let $\tau \in (0, 1)$. Define, for every Borel set $B \subset (0, \infty)$,

$$\Lambda_\lambda(B) \;:=\; \Lambda\big(B/\rho\big), \qquad B/\rho = \{\, s > 0 : \rho s \in B \,\}.$$

Then the measure domination condition

$$\tau\, \Lambda(B) \;\ge\; \Lambda_\lambda(B) \quad \text{for every Borel set } B \subset (0, \infty)$$

holds if and only if

$$\tau \rho\, m(s) \;\ge\; m(s/\rho) \text{ for almost every } s > 0, \qquad \tau\, \Lambda(\{x\}) \;\ge\; \Lambda(\{x/\rho\}) \text{ for every } x > 0.$$

Moreover, two cases occur, depending on whether $\tau \le \rho$ or $\tau > \rho$.

1. If $\tau \le \rho$ and the measure domination condition holds, then $c_j = 0$ for all j, so Λ has no atoms.
2. If $\tau > \rho$, then Λ can have atoms and still satisfy the measure domination condition, but its atoms must then follow the chains of Lemma E.5. Write every $x > 0$ as

$$x = \rho^n r, \qquad n \in \mathbb{Z}, \quad r \in (\rho, 1],$$

as in Lemma E.5, so that x lies on the chain $\{\rho^n r : n \in \mathbb{Z}\}$ of its representative r. If the condition holds and $\rho^n r$ is an atom, then every point of that chain closer to 0, namely $\rho^{n+1}r, \rho^{n+2}r, \ldots$, is also an atom. Hence, on a chain that carries an atom, the atoms are exactly the points $\rho^n r$ with $n \geq n_r$, where n_r is the index of the first atom, or $n_r = -\infty$ if every point of the chain is an atom. The set of atoms of Λ is therefore a union of disjoint tails of geometric chains, of the form

$$\{\rho^n r : n \geq n_r\}, \qquad r \in (\rho, 1],$$

one for each chain that carries an atom. In particular, Λ has either no atoms or infinitely many.

Such atoms can occur on the tail of a single geometric chain, by Lemma E.4, and also on the tails of finitely or countably many geometric chains at once, provided $\int_{(0,\infty)} \min\{1, s\}\, \Lambda(ds) < \infty$.

Proof.

From the measure domination condition to the inequalities on the density and on the atoms. By assumption, Λ has no singular continuous part. Hence Lemma E.2 gives the equivalence

$$\tau\, \Lambda(B) \;\geq\; \Lambda_\lambda(B) \quad \text{for every Borel } B \subset (0,\infty)$$
$$\Longleftrightarrow$$
$$\begin{cases} \tau\rho\, m(s) \geq m(s/\rho) \text{ for almost every } s > 0, \\ \tau\, \Lambda(\{x\}) \geq \Lambda(\{x/\rho\}) \;\; \forall x > 0. \end{cases}$$

In addition, since $\min\{1, s\} = s$ for $s \leq 1$ and $\min\{1, s\} = 1$ for $s > 1$, the decomposition of Λ splits the Lévy condition into two parts,

$$\int_{(0,\infty)} \min\{1, s\}\, \Lambda(ds) = \Big(\int_0^1 s\, m(s)\, ds + \sum_{j:\, s_j \leq 1} s_j c_j\Big) + \Big(\int_1^\infty m(s)\, ds + \sum_{j:\, s_j > 1} c_j\Big) < \infty.$$

Both parts are nonnegative, so each of them is finite,

$$\int_0^1 s\, m(s)\, ds + \sum_{j:\, s_j \leq 1} s_j c_j < \infty, \qquad \int_1^\infty m(s)\, ds + \sum_{j:\, s_j > 1} c_j < \infty.$$

We now turn to the two regimes for τ.

(1) If $\tau \leq \rho$, then Λ has no atoms. Assume for contradiction that there is an atom at $s_0 > 0$ with $\Lambda(\{s_0\}) =: c_0 > 0$. To use the measure domination condition, as in Lemma E.3, we evaluate it on the singleton $B = \{\rho s_0\}$.

$$\tau\, \Lambda(\{\rho s_0\}) \;\geq\; \Lambda_\lambda(\{\rho s_0\}) \;=\; \Lambda(\{s_0\}) \;=\; c_0.$$

Since the absolutely continuous part gives no mass to singletons, only a possible atom at ρs_0 can contribute there. Let

$$c_1 \;:=\; \Lambda(\{\rho s_0\}),$$

which is the mass carried by Λ at ρs_0 and is 0 if there is no atom at that point. The inequality above therefore gives

$$c_1 \;\geq\; c_0/\tau \;>\; 0.$$

There are two cases.

Case 1 (no atom at ρs_0). Then $\Lambda(\{\rho s_0\}) = 0$, contradicting $c_1 \geq c_0/\tau > 0$. Hence this case cannot occur.

Case 2 (an atom at ρs_0). Then $c_1 = \Lambda(\{\rho s_0\}) > 0$. Testing again on $B = \{\rho^2 s_0\}$ yields

$$\tau\, \Lambda(\{\rho^2 s_0\}) \;\geq\; \Lambda(\{\rho s_0\}) = c_1 \quad \Rightarrow \quad \Lambda(\{\rho^2 s_0\}) =: c_2 \;\geq\; \tau^{-1} c_1 \;\geq\; \tau^{-2} c_0 > 0.$$

Iterating, for all $n \geq 1$,

$$c_n \;:=\; \Lambda(\{\rho^n s_0\}) \;\geq\; \tau^{-n} c_0 \;>\; 0.$$

Since $\rho \in (0,1)$, we have $\rho^n s_0 \leq 1$ for all sufficiently large n. Therefore the contribution of these atoms to $\int_{(0,1]} s\,\Lambda(ds)$ satisfies

$$\int_{(0,1]} s\,\Lambda(ds) \;\geq\; \sum_{n:\,\rho^n s_0 \leq 1} (\rho^n s_0)\, c_n \;\geq\; s_0 c_0 \sum_{n:\,\rho^n s_0 \leq 1} \Big(\frac{\rho}{\tau}\Big)^n.$$

If $\tau \leq \rho$, then $\rho/\tau \geq 1$ and the series on the right diverges, contradicting the Lévy integrability $\int_{(0,1]} s\,\Lambda(ds) < \infty$. Thus Λ has no atoms.

(2) If $\tau > \rho$, then the set of atoms is a union of disjoint tails of geometric chains. By Lemma E.5, every $x > 0$ admits a unique decomposition

$$x = \rho^n r, \qquad n \in \mathbb{Z}, \quad r \in (\rho, 1].$$

Fix $r \in (\rho, 1]$, and let

$$A_r := \{\, n \in \mathbb{Z} : \Lambda(\{\rho^n r\}) > 0 \,\}$$

denote the set of indices along that chain at which Λ has an atom. Then, once a level n on that chain carries an atom, the next level $n+1$ must also carry an atom. Indeed, if $n \in A_r$, then $\Lambda(\{\rho^n r\}) > 0$, and the inequality on the atoms, (ii.2) in Lemma E.2, applied with $x = \rho^{n+1} r$, gives

$$\tau\,\Lambda(\{\rho^{n+1} r\}) \;\geq\; \Lambda(\{\rho^n r\}) \;>\; 0,$$

hence $\Lambda(\{\rho^{n+1} r\}) > 0$, that is,

$$n \in A_r \quad\Longrightarrow\quad n+1 \in A_r.$$

Equivalently, for each chain representative $r \in (\rho, 1]$, there is a threshold index $n_r \in \mathbb{Z} \cup \{-\infty, +\infty\}$ that describes where the atoms begin on that chain, in the sense that

$$A_r = \{\, n \in \mathbb{Z} :\; n \geq n_r \,\}.$$

Thus, if $n_r \in \mathbb{Z}$, then the first atom on the chain represented by r occurs at level n_r, and all later levels $n \geq n_r$ also carry atoms. The conventions $n_r = +\infty$ and $n_r = -\infty$ correspond respectively to $A_r = \varnothing$ and $A_r = \mathbb{Z}$.

Therefore, for each chain representative $r \in (\rho, 1]$, the atoms on that chain are exactly

$$\{\rho^n r :\; n \in A_r\} = \{\rho^n r :\; n \geq n_r\}.$$

Since every atom belongs to a unique chain represented by some $r \in (\rho, 1]$, the set of all atoms is obtained by taking the union over all such chains. Hence the set of atoms is a union of disjoint tails of geometric chains, of the form

$$\{\rho^n r :\; n \geq n_r\}, \qquad r \in (\rho, 1].$$

In particular, Λ has either no atoms or infinitely many, in agreement with Lemma E.3.

Conversely, when $\tau > \rho$, a single geometric chain can occur by Lemma E.4. More generally, let I be finite or countably infinite, let $(r_i)_{i \in I} \subset (\rho, 1]$ be distinct chain representatives, and for each $i \in I$ let

$$\Lambda^{(i)} = \sum_{n=0}^{\infty} a_{i,n}\, \delta_{\rho^n r_i}$$

be a chain measure satisfying the measure domination condition. Assume that

$$\sum_{i \in I} \int_{(0,\infty)} \min\{1, s\}\, \Lambda^{(i)}(ds) < \infty.$$

In particular, every term of this sum is finite, so each $\Lambda^{(i)}$ is a Lévy measure. Then

$$\Lambda := \sum_{i \in I} \Lambda^{(i)}$$

is a Lévy measure whose set of atoms is a union of disjoint tails of these chains, and it still satisfies the measure domination condition. Indeed,

$$\int_{(0,\infty)} \min\{1,s\}\,\Lambda(ds) = \sum_{i\in I} \int_{(0,\infty)} \min\{1,s\}\,\Lambda^{(i)}(ds) < \infty,$$

so Λ is a Lévy measure. Moreover, for every Borel set $B \subset (0,\infty)$,

$$\tau\,\Lambda(B) = \sum_{i\in I} \tau\,\Lambda^{(i)}(B) \;\geq\; \sum_{i\in I} \Lambda_\lambda^{(i)}(B) = \Lambda_\lambda(B),$$

where the last identity follows from

$$\Lambda_\lambda(B) = \Lambda(B/\rho) = \sum_{i\in I} \Lambda^{(i)}(B/\rho) = \sum_{i\in I} \Lambda_\lambda^{(i)}(B).$$

Thus, when $\tau > \rho$, finitely or countably many disjoint geometric chains can indeed occur, provided the resulting measure is still a Lévy measure (that is, $\int_{(0,\infty)} \min\{1,s\}\,\Lambda(ds) < \infty$). □

E.5 From Bernstein reduction to kernels inducing contraction

Theorem E.4 (Contraction and smallest contraction factor across dimensions reduce to a Lévy–Khintchine criterion)**.** Adopt the setting of Theorem E.2. For each $d \geq 1$ and $q \in \mathbb{R}^d$, consider the affine map

$$f(x) = q + \lambda x, \qquad x \in \mathbb{R}^d,$$

together with the radial kernel

$$k(u, v) = g(\|u - v\|), \qquad u, v \in \mathbb{R}^d.$$

Let $h(r) = g(0) - g(r)$, and let ϕ be the Bernstein function with $h(r) = \phi(r^2)$. Fix $\lambda \in (0, 1)$, $\tau \in (0, 1)$, and define the difference profile ψ_τ and the kernel Γ_τ built from it by

$$\psi_\tau(t) = \tau\, \phi(t) - \phi(\lambda^2 t), \qquad \Gamma_\tau(x, y) = \psi_\tau(\|x - y\|^2).$$

The contraction inequality

$$\mathbf{MMD}_k^2(f\#\nu, f\#\mu) \leq \tau\, \mathbf{MMD}_k^2(\nu, \mu)$$

holds for every $d \geq 1$ and every $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ if and only if

$$\Gamma_\tau \text{ is CND on } \mathbb{R}^d \text{ for every } d \geq 1.$$

Write the Lévy–Khintchine representation of ϕ as

$$\phi(t) = b\, t + \int_{(0,\infty)} (1 - e^{-st})\, \Lambda(ds), \qquad b \geq 0.$$

Assume Λ has no singular continuous part and decompose

$$\Lambda(ds) = m(s)\, ds + \sum_{j \geq 1} c_j\, \delta_{s_j}.$$

Then Γ_τ is CND on $\mathbb{R}^d$ for every $d \geq 1$ if and only if

1. $b(\tau - \lambda^2) \geq 0$,
2. $\tau\, \lambda^2\, m(s) \geq m(s/\lambda^2)$ for almost every $s > 0$, and
3. $\tau\, \Lambda(\{x\}) \geq \Lambda(\{x/\lambda^2\})$ for every $x > 0$.

Assume these conditions hold. If $\Lambda(\{s_0\}) > 0$, then $\Lambda(\{\lambda^{2n} s_0\}) \geq \tau^{-n} \Lambda(\{s_0\}) > 0$ for every $n \geq 0$. If $\tau \leq \lambda^2$, then Λ has no atoms. If $\tau > \lambda^2$, the set of atoms of Λ is a union of disjoint tails of geometric chains, so Λ has either no atoms or infinitely many.

To summarize the contraction side, define

$$\begin{aligned} \tau^\star(\lambda) \;&:=\; \inf\{\tau > 0 :\ \Gamma_\tau \text{ is CND on } \mathbb{R}^d \text{ for every } d \geq 1\} \\ &=\; \inf\{\tau > 0 :\ \psi_\tau \text{ is Bernstein}\}. \end{aligned}$$

Contraction across dimensions holds if and only if $\tau^\star(\lambda) < 1$.

Smallest contraction factor across dimensions: $\tau^\star(\lambda)$ for $\mathbf{MMD}_k^2$, $\sqrt{\tau^\star(\lambda)}$ for $\mathbf{MMD}_k$.

Drift cases. The drift is $b = \lim_{t \to \infty} \phi(t)/t$. If $b > 0$, then necessarily $\tau \geq \lambda^2$. If $b = 0$, no inequality on τ follows from the drift term.

The factor $\tau^\star(\lambda)$ asks for contraction in every dimension at once. For $\lambda = \gamma$, a factor that gives contraction in every dimension gives it in particular for returns in $\mathbb{R}^d$, so by Theorem 3.1 the smallest contraction factor of the main text satisfies $\tau_k^\star(\mathcal{T}^\pi) \leq \tau^\star(\gamma)$.

Proof.

Step 1: Reduction to a CND test. The first boxed equivalence is exactly Theorem E.2.

Step 2: Lévy–Khintchine drift and the measure domination condition. Write the Lévy–Khintchine representation

$$\phi(t) = b\,t + \int_{(0,\infty)} (1 - e^{-st})\,\Lambda(ds), \qquad b \geq 0,$$

with Λ a nonnegative Borel measure. By Theorem E.3,

$$\Gamma_\tau \text{ is CND on } \mathbb{R}^d \text{ for every } d \geq 1$$
$$\iff$$
$$b(\tau - \lambda^2) \geq 0 \;\text{ and }\; \tau\,\Lambda(B) \geq \Lambda(B/\lambda^2) \;\; \forall B \subset (0,\infty) \text{ Borel.}$$

Step 3: The inequalities on the density and on the atoms. By assumption, Λ has no singular continuous part and decomposes as

$$\Lambda(ds) = m(s)\,ds + \sum_{j \geq 1} c_j\,\delta_{s_j}.$$

Applying Lemma E.2 with $\rho = \lambda^2$ rewrites the measure domination condition as

$$\tau\,\Lambda(B) \geq \Lambda(B/\lambda^2) \;\; \forall \text{Borel } B \quad \iff \quad \begin{cases} \tau\lambda^2\,m(s) \geq m(s/\lambda^2) & \text{for almost every } s > 0, \\ \tau\,\Lambda(\{x\}) \geq \Lambda(\{x/\lambda^2\}) & \forall\, x > 0, \end{cases}$$

which, together with $b(\tau - \lambda^2) \geq 0$, gives the three conditions of the statement.

Step 4: Consequences for the atoms. Under condition 3 of the statement, if $\Lambda(\{s_0\}) > 0$, then

$$\Lambda(\{\lambda^{2n} s_0\}) \geq \tau^{-n} \Lambda(\{s_0\}) \qquad \text{for all } n \geq 0.$$

Thus, once an atom appears at a point, every smaller point of the same geometric chain must also carry an atom.

Moreover, Proposition E.4 with $\rho = \lambda^2$ shows that if $\tau \leq \lambda^2$, then

$$\Lambda(\{x\}) = 0 \qquad \forall x > 0.$$

If instead $\tau > \lambda^2$, then the set of atoms of Λ must be a union of disjoint tails of geometric chains, of the form

$$\{\lambda^{2n} r : \; n \geq n_r\}, \qquad r \in (\lambda^2, 1],$$

and in particular Λ has either no atoms or infinitely many. Conversely, such tails can indeed occur, by Proposition E.4.

Step 5: Smallest contraction factor. By Theorem E.2,

$$\Gamma_\tau \text{ is CND on } \mathbb{R}^d \text{ for every } d \geq 1 \quad \iff \quad \psi_\tau \text{ is Bernstein.}$$

Therefore

$$\tau^\star(\lambda) := \inf\{\tau > 0 : \; \Gamma_\tau \text{ is CND on } \mathbb{R}^d \text{ for every } d \geq 1\} = \inf\{\tau > 0 : \; \psi_\tau \text{ is Bernstein}\}.$$

Combining this with Step 1, contraction across dimensions holds if and only if $\tau^\star(\lambda) < 1$, and then $\tau^\star(\lambda)$ is the smallest contraction factor across dimensions in $\mathbf{MMD}_k^2$, and $\sqrt{\tau^\star(\lambda)}$ in $\mathbf{MMD}_k$.

Step 6: Drift cases. The drift condition $b(\tau - \lambda^2) \geq 0$ of Step 2 gives $\tau \geq \lambda^2$ if $b > 0$, and no restriction on τ if $b = 0$.

□

E.6 Contraction from complete Bernstein functions

Theorem E.5 (For complete Bernstein functions, the smallest contraction factor across dimensions reduces to a single supremum)**.** Adopt the setting of Theorem E.4. For each $d \geq 1$ and $q \in \mathbb{R}^d$, consider the affine map

$$f(x) = q + \lambda x, \qquad x \in \mathbb{R}^d.$$

Assume moreover that ϕ is a *complete* Bernstein function with $\phi(0) = 0$. Then ϕ admits a Lévy–Khintchine representation

$$\phi(t) = b\,t + \int_0^\infty \big(1 - e^{-st}\big)\, m(s)\, ds, \qquad t \geq 0,$$

with $b \geq 0$ and a Lévy density $m : (0, \infty) \to [0, \infty)$ such that, by Definition E.2,

$$m \in C^\infty((0,\infty)) \quad \text{and} \quad m \text{ is completely monotone.}$$

Fix $\lambda \in (0, 1)$ and $\tau > 0$, and define the difference profile

$$\psi_\tau(t) = \tau\,\phi(t) - \phi(\lambda^2 t), \qquad t \geq 0.$$

Then ψ_τ is a Bernstein function if and only if

$$b(\tau - \lambda^2) \geq 0 \qquad \text{and} \qquad \tau\,\lambda^2\, m(s) \;\geq\; m\Big(\frac{s}{\lambda^2}\Big) \;\text{ for every } s > 0.$$

If $m \equiv 0$, then

$$\tau^\star(\lambda) = \lambda^2\, \mathbf{1}_{\{b>0\}}.$$

If $m \not\equiv 0$, then the smallest τ for which ψ_τ is a Bernstein function is

$$\boxed{\tau^\star(\lambda) = \max\left\{\lambda^2\, \mathbf{1}_{\{b>0\}},\; \sup_{s>0} \frac{m\big(\frac{s}{\lambda^2}\big)}{\lambda^2\, m(s)}\right\}.}$$

In particular, for every $d \geq 1$ and all $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,

$$\mathbf{MMD}_k^2(f\#\nu, f\#\mu) \leq \tau\, \mathbf{MMD}_k^2(\nu, \mu)$$

holds if and only if $\tau \geq \tau^\star(\lambda)$. Consequently, contraction across dimensions holds if and only if

$$\tau^\star(\lambda) < 1,$$

in which case $\tau^\star(\lambda)$ is the smallest contraction factor across dimensions in $\mathbf{MMD}_k^2$.

Proof. By Theorem E.4, ψ_τ is Bernstein if and only if

$$b(\tau - \lambda^2) \geq 0 \qquad \text{and} \qquad \tau\,\lambda^2\, m(s) \geq m(s/\lambda^2) \quad \text{for almost every } s > 0.$$

By Definition E.2, one has $m \in C^\infty((0,\infty))$. Hence both sides of the inequality are continuous in s, so the inequality, which holds almost everywhere, extends to every $s > 0$. This proves the first claim.

If $m \equiv 0$, only the drift condition remains, so

$$\tau^\star(\lambda) = \lambda^2\, \mathbf{1}_{\{b>0\}}.$$

Assume now that $m \not\equiv 0$. Since m is completely monotone and nonzero, it is strictly positive on $(0, \infty)$. Therefore $\lambda^2 m(s) > 0$ for every $s > 0$, so we may divide both sides by $\lambda^2 m(s)$. Thus

$$\tau\,\lambda^2\, m(s) \geq m(s/\lambda^2) \qquad \forall s > 0$$

is equivalent to

$$\tau \geq \frac{m(s/\lambda^2)}{\lambda^2\, m(s)} \qquad \forall s > 0.$$

Combining this with $b(\tau - \lambda^2) \geq 0$, we obtain that ψ_τ is Bernstein if and only if

$$\tau \geq \lambda^2\, \mathbf{1}_{\{b>0\}} \qquad \text{and} \qquad \tau \geq \frac{m(s/\lambda^2)}{\lambda^2\, m(s)} \quad \forall s > 0.$$

Therefore

$$\tau^\star(\lambda) = \max\left\{\lambda^2\, \mathbf{1}_{\{b>0\}},\; \sup_{s>0} \frac{m(s/\lambda^2)}{\lambda^2\, m(s)}\right\}.$$

The final equivalence with the $\mathbf{MMD}_k^2$ contraction statement follows directly from Theorem E.4. □

Corollary E.5.1 (Limit at 0^+ and consequences for the metric property and contraction)**.** Adopt the setting of Theorem E.5. Assume there exists $\tau < 1$ such that

$$\psi_\tau(t) = \tau\phi(t) - \phi(\lambda^2 t)$$

is a Bernstein function. Since m is completely monotone, it is nonincreasing, so its limit at 0^+,

$$m(0^+) := \lim_{s\to 0^+} m(s) \in [0, \infty],$$

exists. Then $m(0^+)$ is either 0 or ∞.

1. If $m(0^+) = 0$, then $m \equiv 0$, so $\phi(t) = b\,t$ and $k(x, y) = g(0) - b\|x - y\|^2$. In this case, $\mathbf{MMD}_k$ is not a metric by Theorem I.1. By Theorem E.5, the smallest contraction factor across dimensions is $\tau^\star(\lambda) = \lambda^2$ if $b > 0$ and $\tau^\star(\lambda) = 0$ if $b = 0$.
2. If $\mathbf{MMD}_k$ is a metric, then $m(0^+) = \infty$.

Proof.

Step 1: Existence of $m(0^+)$. Since m is completely monotone, it is nonincreasing on $(0, \infty)$. Hence its boundary value at 0^+ is given by the supremum, and we may write

$$m(0^+) = \sup_{s>0} m(s) \ \in \ [0, \infty].$$

Step 2: A strict contraction excludes $m(0^+) \in (0, \infty)$. Assume for contradiction that $m(0^+) \in (0, \infty)$. By Theorem E.5, the assumption that ψ_τ is a Bernstein function implies that

$$\tau\,\lambda^2\, m(s) \ \geq \ m\Big(\frac{s}{\lambda^2}\Big) \qquad \text{for every } s > 0.$$

Letting $s \to 0^+$ and using $m(s) \to m(0^+)$ and $m(s/\lambda^2) \to m(0^+)$ gives

$$\tau\,\lambda^2\, m(0^+) \ \geq \ m(0^+).$$

Since $m(0^+) > 0$, we may divide by $m(0^+)$ to obtain $\tau\,\lambda^2 \geq 1$. This contradicts $\tau < 1$ and $\lambda \in (0, 1)$, which together imply $\tau\,\lambda^2 < 1$. Hence $m(0^+) \notin (0, \infty)$, and therefore $m(0^+) \in \{0, \infty\}$.

Step 3: If $m(0^+) = 0$ then $m \equiv 0$, MMD is not a metric, and contraction depends only on b. If $m(0^+) = 0$ and m is nonincreasing and nonnegative, then for all $s > 0$,

$$0 \ \leq \ m(s) \ \leq \ m(0^+) = 0,$$

hence $m \equiv 0$ and $\phi(t) = b\,t$. By Theorem I.1, $\mathbf{MMD}_k$ is a metric if and only if $\Lambda \not\equiv 0$, equivalently $m \not\equiv 0$. Thus $\mathbf{MMD}_k$ is not a metric in this case. Finally, when $m \equiv 0$, the supremum term in Theorem E.5 vanishes and $\tau^\star(\lambda) = \lambda^2\,\mathbf{1}_{\{b>0\}}$.

Step 4: If $\mathbf{MMD}_k$ is a metric then $m(0^+) = \infty$. If $\mathbf{MMD}_k$ is a metric, then Theorem I.1 implies $\Lambda \not\equiv 0$, hence $m \not\equiv 0$. Step 3 shows that $m(0^+) = 0$ would force $m \equiv 0$, a contradiction. Therefore $m(0^+) = \infty$. □

E.7 Chaining as a general mechanism

Preamble. Section E.4, and in particular Lemma E.4 and Proposition E.4, introduced a geometric chaining mechanism. When Λ has atoms and satisfies the measure domination condition, its atoms cannot lie anywhere. By Proposition E.4, the set of atoms is a union of disjoint tails of geometric chains.

In this subsection, we show that this mechanism is more general. The geometric chain of atoms is the special case of a broader construction in which the base Bernstein function is $t \mapsto 1 - e^{-st}$.

As a first step, we isolate the basic building block already present in Section E.4, namely a single geometric chain of atoms. This is the simplest instance of the general construction that follows.

Proposition E.5 (Single geometric chain of atoms)**.** Fix $\lambda \in (0,1)$ and set $\rho := \lambda^2 \in (0,1)$. Let $\tau \in (0,1)$, let $s_0 > 0$, and let $c > 0$. Define the measure

$$\Lambda(ds) := \sum_{n=0}^{\infty} c\,\tau^{-n}\,\delta_{\rho^n s_0}(ds).$$

If $\tau > \rho$, then

$$\int_{(0,\infty)} \min\{1,s\}\,\Lambda(ds) \le s_0 c \sum_{n=0}^{\infty} \Big(\frac{\rho}{\tau}\Big)^n < \infty,$$

so Λ is a Lévy measure. In addition, Lemma E.4 shows that this single geometric chain satisfies the measure domination condition. Its Bernstein profile is therefore

$$\phi_{\mathrm{chain}}(t) = bt + \sum_{n=0}^{\infty} c\,\tau^{-n}\big(1 - e^{-\rho^n s_0 t}\big), \qquad t \ge 0, \quad b \ge 0.$$

Since $\tau > \rho = \lambda^2$, the drift condition in Theorem E.4 is also satisfied. Hence, under the radial kernel generated by ϕ_{chain}, the affine map $x \mapsto q + \lambda x$ is a τ-contraction in $\mathbf{MMD}^2$.

See Figure 5 for a visualization of the atoms and the resulting ϕ_{chain}.

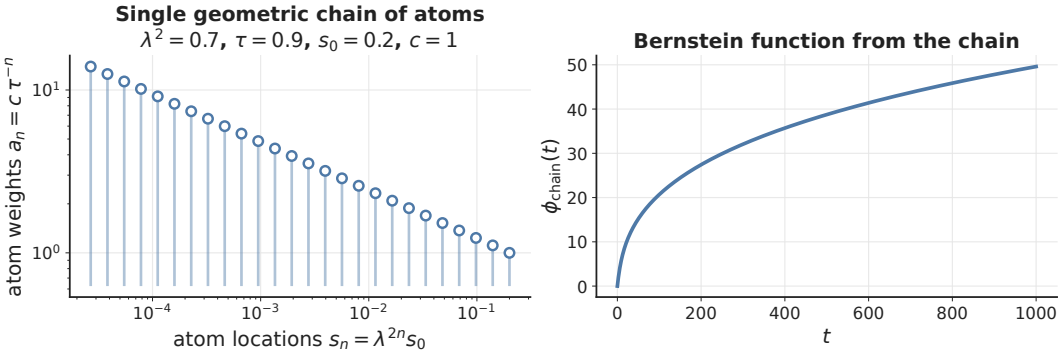

Figure 5: Single geometric chain of atoms. Left, atoms at $s_n = \lambda^{2n} s_0$ with weights $a_n = c\,\tau^{-n}$. The chain is geometric in both location and weight, since $s_{n+1}/s_n = \lambda^2$ and $a_{n+1}/a_n = \tau^{-1}$ are constant. The left panel is truncated after a finite number of atoms for readability. Right, the corresponding Bernstein function $\phi_{\mathrm{chain}}(t) = bt + \sum_{n\ge 0} c\,\tau^{-n}\big(1 - e^{-s_n t}\big)$, evaluated with truncation after 250 terms.

Proposition E.6 (Geometric chaining gives identities for the Lévy measure and the difference profile)**.** Fix $\rho \in (0,1)$ and $\tau \in (0,1)$. Let ϕ_0 be a *base* Bernstein function with $\phi_0(0) = 0$, and write its Lévy–Khintchine representation as

$$\phi_0(t) = b_0\,t + \int_{(0,\infty)} (1 - e^{-st})\,\Lambda_0(ds), \qquad t \ge 0,$$

where $b_0 \ge 0$ and Λ_0 is a Lévy measure on $(0,\infty)$.

Consider the chained profile

$$\phi(t) := \sum_{n=0}^{\infty} \tau^{-n} \phi_0(\rho^n t), \qquad t \ge 0.$$

For each $n \ge 0$, define the rescaled measure

$$\Lambda_0^{(n)}(B) := \Lambda_0(B/\rho^n), \qquad B/\rho^n := \{s > 0 : \rho^n s \in B\}.$$

Define

$$b := \sum_{n=0}^{\infty} \tau^{-n} \rho^n b_0, \qquad \Lambda(B) := \sum_{n=0}^{\infty} \tau^{-n} \Lambda_0(B/\rho^n),$$

and assume that

$$b < \infty \qquad \text{and} \qquad \int_{(0,\infty)} \min\{1, s\}\, \Lambda(ds) < \infty.$$

Then ϕ is a Bernstein function and admits the Lévy–Khintchine representation

$$\phi(t) = b\, t + \int_{(0,\infty)} (1 - e^{-st})\, \Lambda(ds), \qquad t \geq 0.$$

Moreover, if we define

$$\Lambda_\rho(B) := \Lambda(B/\rho),$$

then

$$\boxed{\tau\, \Lambda = \Lambda_\rho + \tau\, \Lambda_0}$$

as nonnegative Borel measures on $(0, \infty)$, and

$$\boxed{\tau\, \phi(t) - \phi(\rho t) = \tau\, \phi_0(t)}$$

for all $t \geq 0$.

Proof.

Lévy–Khintchine representation of ϕ. By definition,

$$\phi(t) = \sum_{n=0}^{\infty} \tau^{-n} \phi_0(\rho^n t).$$

Using the Lévy–Khintchine representation of ϕ_0, this becomes

$$\phi(t) = \sum_{n=0}^{\infty} \tau^{-n} \left[\rho^n b_0\, t + \int_{(0,\infty)} (1 - e^{-u\rho^n t})\, \Lambda_0(du) \right].$$

Since all terms are nonnegative, Tonelli's theorem yields

$$\phi(t) = \left(\sum_{n=0}^{\infty} \tau^{-n} \rho^n b_0 \right) t + \int_{(0,\infty)} \sum_{n=0}^{\infty} \tau^{-n} (1 - e^{-u\rho^n t})\, \Lambda_0(du).$$

By the definition of the rescaled measures $\Lambda_0^{(n)}(B) = \Lambda_0(B/\rho^n)$, this is exactly

$$\phi(t) = b\, t + \int_{(0,\infty)} (1 - e^{-st})\, \Lambda(ds).$$

Exact identities for Λ and ϕ. By definition,

$$\Lambda_\rho(B) = \Lambda(B/\rho) = \sum_{n=0}^{\infty} \tau^{-n} \Lambda_0(B/\rho^{n+1}).$$

Reindexing with $j = n + 1$ gives

$$\Lambda_\rho(B) = \sum_{j=1}^{\infty} \tau^{-(j-1)} \Lambda_0(B/\rho^j) = \tau \sum_{j=1}^{\infty} \tau^{-j} \Lambda_0(B/\rho^j).$$

On the other hand,

$$\tau\, \Lambda(B) = \tau \sum_{j=0}^{\infty} \tau^{-j} \Lambda_0(B/\rho^j) = \tau\, \Lambda_0(B) + \tau \sum_{j=1}^{\infty} \tau^{-j} \Lambda_0(B/\rho^j).$$

Comparing the two expressions, we obtain

$$\tau\, \Lambda(B) = \Lambda_\rho(B) + \tau\, \Lambda_0(B) \qquad \text{for every Borel set } B \subset (0, \infty).$$

That is,

$$\tau\,\Lambda = \Lambda_\rho + \tau\,\Lambda_0.$$

Recall that

$$\phi(t) = \sum_{n=0}^{\infty} \tau^{-n} \phi_0(\rho^n t).$$

Therefore

$$\tau\,\phi(t) = \tau \sum_{n=0}^{\infty} \tau^{-n} \phi_0(\rho^n t) = \tau\,\phi_0(t) + \tau \sum_{j=1}^{\infty} \tau^{-j} \phi_0(\rho^j t).$$

On the other hand,

$$\phi(\rho t) = \sum_{n=0}^{\infty} \tau^{-n} \phi_0(\rho^{n+1} t).$$

Reindexing with $j = n + 1$ gives

$$\phi(\rho t) = \sum_{j=1}^{\infty} \tau^{-(j-1)} \phi_0(\rho^j t) = \tau \sum_{j=1}^{\infty} \tau^{-j} \phi_0(\rho^j t).$$

Comparing the two expressions, we obtain

$$\tau\,\phi(t) - \phi(\rho t) = \tau\,\phi_0(t),$$

which is exactly the claimed identity for the difference profile. □

Proposition E.7 (When geometric chaining gives a Bernstein function)**.** Fix $\rho \in (0, 1)$ and $\tau \in (0, 1)$. Let ϕ_0 be a base Bernstein function with $\phi_0(0) = 0$, and write its Lévy–Khintchine representation as

$$\phi_0(t) = b_0\,t + \int_{(0,\infty)} (1 - e^{-st})\,\Lambda_0(ds), \qquad t \geq 0,$$

where $b_0 \geq 0$ and Λ_0 is a Lévy measure on $(0, \infty)$. Consider the chained profile

$$\phi(t) := \sum_{n=0}^{\infty} \tau^{-n} \phi_0(\rho^n t), \qquad t \geq 0.$$

For each $n \geq 0$, define the rescaled measure

$$\Lambda_0^{(n)}(A) := \Lambda_0(A/\rho^n), \qquad A/\rho^n := \{s > 0 : \rho^n s \in A\},$$

and define, as in Proposition E.6,

$$b := \sum_{n=0}^{\infty} \tau^{-n} \rho^n b_0, \qquad \Lambda(A) := \sum_{n=0}^{\infty} \tau^{-n} \Lambda_0(A/\rho^n).$$

Introduce also

$$W_{\rho,\tau}(s) := \sum_{n=0}^{\infty} \tau^{-n} \min\{1, \rho^n s\}, \qquad s > 0.$$

Then the conditions

$$b < \infty \quad \text{and} \quad \int_{(0,\infty)} \min\{1, s\}\,\Lambda(ds) < \infty$$

hold if and only if

$$\sum_{n=0}^{\infty} \tau^{-n} \rho^n b_0 < \infty \quad \text{and} \quad \int_{(0,\infty)} W_{\rho,\tau}(s)\,\Lambda_0(ds) < \infty.$$

Whenever these equivalent conditions hold, the chained profile is a Bernstein function and admits the representation

$$\phi(t) = b\,t + \int_{(0,\infty)} (1 - e^{-st})\,\Lambda(ds), \qquad t \geq 0.$$

In particular, if $\phi_0 \not\equiv 0$, these conditions can hold only if $\tau > \rho$.

Proof.

Equivalence of the integrability conditions. By definition,

$$b = \sum_{n=0}^{\infty} \tau^{-n} \rho^n b_0,$$

so the drift part is well defined if and only if

$$\sum_{n=0}^{\infty} \tau^{-n} \rho^n b_0 < \infty.$$

For the measure part, Tonelli's theorem applies because the measures $\tau^{-n}\Lambda_0^{(n)}$ are nonnegative and the integrand $\min\{1, u\}$ is nonnegative. Therefore

$$\begin{aligned}\int_{(0,\infty)} \min\{1,u\}\,\Lambda(du) &= \int_{(0,\infty)} \min\{1,u\} \sum_{n=0}^{\infty} \tau^{-n} \Lambda_0^{(n)}(du) \\ &= \sum_{n=0}^{\infty} \tau^{-n} \int_{(0,\infty)} \min\{1,u\}\,\Lambda_0^{(n)}(du).\end{aligned}$$

By the definition of $\Lambda_0^{(n)}$,

$$\int_{(0,\infty)} \min\{1,u\}\,\Lambda_0^{(n)}(du) = \int_{(0,\infty)} \min\{1,\rho^n s\}\,\Lambda_0(ds).$$

Hence

$$\begin{aligned}\int_{(0,\infty)} \min\{1,u\}\,\Lambda(du) &= \sum_{n=0}^{\infty} \tau^{-n} \int_{(0,\infty)} \min\{1,\rho^n s\}\,\Lambda_0(ds) \\ &= \int_{(0,\infty)} \sum_{n=0}^{\infty} \tau^{-n} \min\{1,\rho^n s\}\,\Lambda_0(ds) \\ &= \int_{(0,\infty)} W_{\rho,\tau}(s)\,\Lambda_0(ds).\end{aligned}$$

Thus Λ satisfies the Lévy integrability condition if and only if

$$\int_{(0,\infty)} W_{\rho,\tau}(s)\,\Lambda_0(ds) < \infty.$$

This proves the equivalence.

Under these equivalent conditions, Proposition E.6 applies to

$$\phi(t) = \sum_{n=0}^{\infty} \tau^{-n} \phi_0(\rho^n t)$$

and yields

$$\phi(t) = b\,t + \int_{(0,\infty)} (1 - e^{-st})\,\Lambda(ds).$$

In particular, ϕ is then a Bernstein function.

The regime $\tau > \rho$ is unavoidable. Assume $\phi_0 \not\equiv 0$. If $b_0 > 0$, then

$$b = b_0 \sum_{n=0}^{\infty} \left(\frac{\rho}{\tau}\right)^n,$$

and therefore $b < \infty$ implies $\rho/\tau < 1$, that is, $\tau > \rho$.

Assume now $b_0 = 0$. Since $\phi_0 \not\equiv 0$, the measure Λ_0 is nonzero, so there exists an interval $[u, v] \subset (0, \infty)$ such that

$$\int_{[u,v]} s\, \Lambda_0(ds) > 0.$$

Choose K such that $\rho^n v \le 1$ for every $n \ge K$. Then, for every $n \ge K$ and every $s \in [u, v]$, one has $\rho^n s \le \rho^n v \le 1$, hence $\min\{1, \rho^n s\} = \rho^n s$. Therefore,

$$\begin{aligned}\int_{(0,\infty)} \min\{1, \rho^n s\}\, \Lambda_0(ds) &\ge \int_{[u,v]} \min\{1, \rho^n s\}\, \Lambda_0(ds) \\ &= \int_{[u,v]} \rho^n s\, \Lambda_0(ds) \\ &= \rho^n \int_{[u,v]} s\, \Lambda_0(ds).\end{aligned}$$

It follows, by Tonelli's theorem as in the preceding part, that

$$\begin{aligned}\int_{(0,\infty)} W_{\rho,\tau}(s)\, \Lambda_0(ds) &= \int_{(0,\infty)} \sum_{n=0}^{\infty} \tau^{-n} \min\{1, \rho^n s\}\, \Lambda_0(ds) \\ &= \sum_{n=0}^{\infty} \tau^{-n} \int_{(0,\infty)} \min\{1, \rho^n s\}\, \Lambda_0(ds) \\ &\ge \sum_{n=K}^{\infty} \tau^{-n} \int_{(0,\infty)} \min\{1, \rho^n s\}\, \Lambda_0(ds) \\ &\ge \sum_{n=K}^{\infty} \tau^{-n} \rho^n \int_{[u,v]} s\, \Lambda_0(ds) \\ &= \left(\int_{[u,v]} s\, \Lambda_0(ds) \right) \sum_{n=K}^{\infty} \left(\frac{\rho}{\tau} \right)^n.\end{aligned}$$

If $\tau \le \rho$, then $\rho/\tau \ge 1$, so the series on the right diverges. Therefore

$$\int_{(0,\infty)} W_{\rho,\tau}(s)\, \Lambda_0(ds) = \infty,$$

which contradicts the integrability criterion proved above. Thus any such geometric chaining construction must satisfy $\tau > \rho$. □

Theorem E.6 (Geometric chaining gives contraction with factor τ)**.** Fix $\lambda \in (0, 1)$ and set $\rho := \lambda^2$. Let $\tau \in (0, 1)$, and let ϕ_0 be a base Bernstein function with $\phi_0(0) = 0$. Adopt the setting of Proposition E.6, and assume that $b < \infty$ and that Λ is a Lévy measure, as in Proposition E.7. Let

$$\phi(t) := b\, t + \int_{(0,\infty)} (1 - e^{-st})\, \Lambda(ds), \qquad t \ge 0.$$

For each $d \ge 1$ and $q \in \mathbb{R}^d$, consider the affine map

$$f(x) = q + \lambda x, \qquad x \in \mathbb{R}^d,$$

and define the radial profile and kernel by

$$h(r) := \phi(r^2), \qquad k(x, y) := C - \phi(\|x - y\|^2),$$

where $C \in \mathbb{R}$ is arbitrary. Define moreover the difference profile

$$\psi_\tau(t) := \tau\, \phi(t) - \phi(\rho t), \qquad t \ge 0.$$

Then

$$\boxed{\begin{array}{c} \psi_\tau(t) = \tau\, \phi_0(t) \qquad \text{for every } t \ge 0, \\ \text{hence } \psi_\tau \text{ is a Bernstein function,} \\ \text{and therefore } \mathbf{MMD}_k^2(f\#\nu, f\#\mu) \le \tau\, \mathbf{MMD}_k^2(\nu, \mu) \\ \text{for every } d \ge 1 \text{ and every } \mu, \nu \in \mathcal{P}(\mathbb{R}^d). \end{array}}$$

In particular, under the kernel generated by the geometrically chained profile ϕ, every such affine map f is a τ-contraction in $\mathbf{MMD}^2$.

Proof.

Step 1: Exact difference identity. By Proposition E.6,

$$\tau\,\phi(t) - \phi(\rho t) = \tau\,\phi_0(t) \qquad \text{for every } t \geq 0.$$

That is,

$$\psi_\tau(t) = \tau\,\phi_0(t).$$

Step 2: Bernstein property and contraction. Since ϕ_0 is a Bernstein function and the class of Bernstein functions is closed under multiplication by a nonnegative scalar [29, Corollary 3.7(i)], it follows from Step 1 that $\psi_\tau = \tau\,\phi_0$ is Bernstein. Applying Theorem E.2 with the present notation and $\rho = \lambda^2$, we obtain

$$\mathbf{MMD}_k^2(f\#\nu, f\#\mu) \leq \tau\,\mathbf{MMD}_k^2(\nu, \mu)$$

for every $d \geq 1$ and every $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. This is exactly the claimed τ-contraction in $\mathbf{MMD}^2$. □

F Explaining known results

Preamble. In this section, we show that our framework explains three known results on MMD contraction for radial kernels. As in Appendix E, we study contraction for a fixed affine map $f(x) = q + \lambda x$, where λ plays the role of the discount γ. These results can then be extended to the setting of distributional RL and its Bellman operators using the results from Appendix C.

The first positive result was established in [25] for power-distance kernels. Proposition F.1 shows that these kernels are induced by a complete Bernstein function and satisfy the criterion of Theorem E.5. Hence they induce contraction under a fixed map. The corresponding negative result was also established in [25], where it was shown that the RBF kernel does not induce contraction in general. Proposition F.3 shows that the RBF kernel is likewise induced by a Bernstein function, and that its non-contraction can be explained through its Lévy–Khintchine representation together with the obstruction for saturating profiles from Proposition E.3. The second positive result was established in [19] and [12, Lemma 10] for the multiquadric kernel. Proposition F.2 extends this result to the family of *generalized multiquadric kernels* and shows that these kernels are also induced by a complete Bernstein function.

Proposition F.1 (Contraction under power-distance kernels [25])**.** Let

$$k_\alpha(x, y) \;=\; -\|x - y\|^\alpha, \qquad 0 < \alpha \le 2,$$

and let

$$f(x) = q + \lambda x, \qquad \lambda \in (0, 1).$$

Then

$$h(r) = r^\alpha = \phi(r^2), \qquad \phi(t) = t^{\alpha/2},$$

and the smallest contraction factor across dimensions in $\mathbf{MMD}^2_{k_\alpha}$ is $\tau^\star(\lambda) = \lambda^\alpha$. For $\alpha = 2$, $\mathbf{MMD}_{k_2}$ is not a metric on probability measures.

Proof.

Lévy–Khintchine representation. By [29, Ch. 15.2, Identity 1],

- for $0 < \alpha < 2$ (so $\alpha/2 \in (0, 1)$),

$$\phi(t) = \frac{\alpha/2}{\Gamma(1 - \alpha/2)} \int_0^\infty (1 - e^{-st})\, s^{-1-\alpha/2}\, ds,$$

  so the drift is $b = 0$ and the Lévy measure is absolutely continuous,

$$\Lambda(ds) = c_\alpha\, s^{-1-\alpha/2}\, ds, \qquad c_\alpha = \frac{\alpha/2}{\Gamma(1 - \alpha/2)}.$$

- for $\alpha = 2$, $\phi(t) = t$, so the drift is $b = 1$ and $\Lambda = 0$.

Bernstein test for the difference profile. By Theorem E.2, in the present radial setting, contraction in $\mathbf{MMD}^2$ reduces to checking whether the difference profile

$$\psi_\tau(t) := \tau\, \phi(t) - \phi(\lambda^2 t)$$

is Bernstein. Here

$$\psi_\tau(t) = (\tau - \lambda^\alpha)\, t^{\alpha/2}.$$

Since $\phi(t) = t^{\alpha/2}$ is Bernstein and the class of Bernstein functions is closed under multiplication by a nonnegative scalar [29, Corollary 3.7(i)], it follows that ψ_τ is Bernstein if and only if $\tau - \lambda^\alpha \ge 0$. Therefore

$$\boxed{\begin{array}{c} \mathbf{MMD}^2_{k_\alpha}(f\#\nu, f\#\mu) \le \tau\, \mathbf{MMD}^2_{k_\alpha}(\nu, \mu) \\ \text{for every } d \ge 1 \text{ and every } \mu, \nu \in \mathcal{P}(\mathbb{R}^d) \\ \Longleftrightarrow \\ \tau \ge \lambda^\alpha. \end{array}}$$

Via Theorem E.5. For $0 < \alpha < 2$, the function $\phi(t) = t^{\alpha/2}$ is a complete Bernstein function by [29, Ch. 15.2, Identity 1], with $b = 0$ and

$$m(s) = c_\alpha \, s^{-1-\alpha/2}.$$

Hence Theorem E.5 applies directly and gives

$$\tau^\star(\lambda) = \sup_{s>0} \frac{m(s/\lambda^2)}{\lambda^2 \, m(s)} = \sup_{s>0} \frac{c_\alpha (s/\lambda^2)^{-1-\alpha/2}}{\lambda^2 c_\alpha s^{-1-\alpha/2}} = \lambda^\alpha.$$

For $\alpha = 2$, one has $\phi(t) = t$, hence $b = 1$ and $m \equiv 0$, so Theorem E.5 gives $\tau^\star(\lambda) = \lambda^2$. Thus in all cases $\tau^\star(\lambda) = \lambda^\alpha$, which agrees with the Bernstein test above.

Contraction factor. The smallest contraction factor across dimensions in $\mathbf{MMD}^2_{k_\alpha}$ is

$$\tau^\star(\lambda) = \lambda^\alpha \; (< 1),$$

so

$$\mathbf{MMD}^2_{k_\alpha}(f\#\nu, f\#\mu) \;\le\; \lambda^\alpha \, \mathbf{MMD}^2_{k_\alpha}(\nu, \mu).$$

Equivalently,

$$\mathbf{MMD}_{k_\alpha}(f\#\nu, f\#\mu) \;\le\; \lambda^{\alpha/2} \, \mathbf{MMD}_{k_\alpha}(\nu, \mu).$$

Special case $\alpha = 2$. Since $k_2(x, y) = -\|x - y\|^2$ can be written in the Bernstein form

$$k_2(x, y) = C - \phi(\|x - y\|^2) \quad \text{with} \quad C = 0, \; \phi(t) = t,$$

Theorem I.1 applies and gives the criterion

$$\mathbf{MMD}_{k_2} \text{ is a metric} \quad \Longleftrightarrow \quad \Lambda \not\equiv 0.$$

Here $\phi(t) = t$ is linear, so its Lévy measure satisfies $\Lambda \equiv 0$. Hence $\mathbf{MMD}_{k_2}$ is not a metric on probability measures. □

Proposition F.2 (Contraction under generalized multiquadric kernels, generalizing [19] and [12, Lemma 10])**.** Fix $h > 0$ and $\beta \in (0, 1]$. Consider

$$k_{\beta,h}(x, y) \;=\; -\big(1 + h^2\|x - y\|^2\big)^\beta, \qquad \phi_\beta(t) \;=\; \big(1 + h^2 t\big)^\beta - 1.$$

Assume $f(x) = q + \lambda x$ with $\lambda \in (0, 1)$. Then the smallest contraction factor across dimensions in $\mathbf{MMD}^2_{k_{\beta,h}}$ is $\tau^\star(\lambda) = \lambda^{2\beta}$.

Proof. Define

$$\psi_\tau(t) = \tau \, \phi_\beta(t) - \phi_\beta(\lambda^2 t), \qquad \Gamma_\tau(x, y) = \psi_\tau(\|x - y\|^2).$$

Lévy–Khintchine representation. By [29, Ch. 15.2, Identity 2], for $0 < \beta < 1$,

$$\phi_\beta(t) \;=\; \frac{\beta \, h^{2\beta}}{\Gamma(1-\beta)} \int_0^\infty \big(1 - e^{-st}\big) \, e^{-s/h^2} \, s^{-\beta-1} \, ds,$$

so $b = 0$ and the Lévy measure is absolutely continuous with density

$$m_{\beta,h}(s) \;=\; \frac{\beta \, h^{2\beta}}{\Gamma(1-\beta)} \, e^{-s/h^2} \, s^{-\beta-1}, \qquad s > 0.$$

For $\beta = 1$, $\phi_1(t) = h^2 t$ is linear, hence $b = h^2$ and $\Lambda \equiv 0$. In particular, ϕ_β is a complete Bernstein function.

Bernstein test for the difference profile. By Theorem E.2, in the present radial setting, contraction in $\mathbf{MMD}^2$ reduces to checking whether the difference profile

$$\psi_\tau(t) = \tau \, \phi_\beta(t) - \phi_\beta(\lambda^2 t)$$

is Bernstein. Thus

$$\mathbf{MMD}^2_{k_{\beta,h}}(f\#\nu, f\#\mu) \le \tau \, \mathbf{MMD}^2_{k_{\beta,h}}(\nu, \mu) \qquad \text{for every } d \ge 1 \text{ and every } \mu, \nu \in \mathcal{P}(\mathbb{R}^d)$$

holds if and only if ψ_τ is Bernstein.

Via Theorem E.5. For $0 < \beta < 1$, the representation above shows that ϕ_β is a complete Bernstein function with $b = 0$ and

$$m_{\beta,h}(s) = \frac{\beta\, h^{2\beta}}{\Gamma(1-\beta)}\, e^{-s/h^2}\, s^{-\beta-1}.$$

Hence Theorem E.5 applies directly, so the smallest contraction factor across dimensions is

$$\tau^\star(\lambda) = \sup_{s>0} \frac{m_{\beta,h}(s/\lambda^2)}{\lambda^2\, m_{\beta,h}(s)}.$$

Using the formula for $m_{\beta,h}$, we obtain for all $s > 0$,

$$\begin{aligned}
\frac{m_{\beta,h}(s/\lambda^2)}{\lambda^2\, m_{\beta,h}(s)} &= \frac{1}{\lambda^2}\, \frac{\frac{\beta\, h^{2\beta}}{\Gamma(1-\beta)}\, e^{-\frac{s}{\lambda^2 h^2}}\, \left(\frac{s}{\lambda^2}\right)^{-\beta-1}}{\frac{\beta\, h^{2\beta}}{\Gamma(1-\beta)}\, e^{-\frac{s}{h^2}}\, s^{-\beta-1}} \\
&= \frac{1}{\lambda^2}\, \exp\Big(-\frac{s}{\lambda^2 h^2} + \frac{s}{h^2}\Big)\, \left(\frac{s}{\lambda^2}\right)^{-\beta-1}\, s^{\beta+1} \\
&= \frac{1}{\lambda^2}\, \exp\Big(-\frac{1-\lambda^2}{h^2\lambda^2}\, s\Big)\, \lambda^{2(\beta+1)} \\
&= \lambda^{2\beta}\, \exp\Big(-\frac{1-\lambda^2}{h^2\lambda^2}\, s\Big) \;\le\; \lambda^{2\beta}.
\end{aligned}$$

Since $\lambda \in (0,1)$, the coefficient $\frac{1-\lambda^2}{h^2\lambda^2}$ is positive, so the exponential factor converges to 1 as $s \to 0^+$. Therefore the ratio converges to $\lambda^{2\beta}$ as $s \to 0^+$, and the supremum is exactly $\lambda^{2\beta}$. Thus $\tau^\star(\lambda) = \lambda^{2\beta}$ for $0 < \beta < 1$. For $\beta = 1$, one has $\phi_1(t) = h^2 t$, hence $b = h^2$ and $m \equiv 0$, so Theorem E.5 gives $\tau^\star(\lambda) = \lambda^2$. Therefore $\tau^\star(\lambda) = \lambda^{2\beta}$ for all $\beta \in (0,1]$. Equivalently,

$$\Gamma_\tau \text{ is CND on } \mathbb{R}^d \text{ for every } d \ge 1 \quad\Longleftrightarrow\quad \tau \ge \lambda^{2\beta},$$

or, by Theorem E.2,

$$\mathbf{MMD}^2_{k_{\beta,h}}(f\#\nu, f\#\mu) \le \tau\, \mathbf{MMD}^2_{k_{\beta,h}}(\nu,\mu) \quad\Longleftrightarrow\quad \tau \ge \lambda^{2\beta}.$$

Contraction factor. The smallest contraction factor across dimensions in $\mathbf{MMD}^2$ is $\tau^\star(\lambda) = \lambda^{2\beta}$, and the corresponding smallest contraction factor in $\mathbf{MMD}$ is $\sqrt{\tau^\star(\lambda)} = \lambda^\beta$, so

$$\mathbf{MMD}_{k_{\beta,h}}(f\#\nu, f\#\mu) \;\le\; \lambda^\beta\, \mathbf{MMD}_{k_{\beta,h}}(\nu,\mu).$$

In particular, for $\lambda \in (0,1)$ one has $\tau^\star(\lambda) < 1$, so f is a contraction in $\mathbf{MMD}^2$ with factor $\tau^\star(\lambda)$. □

Proposition F.3 (No contraction under RBF kernels [25])**.** Fix $h > 0$ and consider the RBF kernel

$$k_{\mathrm{rbf}}(x,y) = \exp\big(-h^2\|x-y\|^2\big), \qquad g(r) = e^{-h^2 r^2}.$$

Then

$$g(0) - g(r) = 1 - e^{-h^2 r^2} = \phi(r^2), \qquad \phi(t) = 1 - e^{-h^2 t}.$$

Assume $f(x) = q + \lambda x$ with $\lambda \in (0,1)$. Then there is no contraction factor in $\mathbf{MMD}^2_{k_{\mathrm{rbf}}}$, that is, $\tau^\star(\lambda) \ge 1$.

Proof. Define

$$\psi_\tau(t) = \tau\,\phi(t) - \phi(\lambda^2 t), \qquad \Gamma_\tau(x,y) = \psi_\tau(\|x-y\|^2).$$

Lévy–Khintchine representation. The function ϕ is a Bernstein function with

$$\phi(t) = \int_{(0,\infty)} (1 - e^{-st})\, \Lambda(ds), \qquad \Lambda = \delta_{h^2}, \qquad b = 0.$$

Thus the Lévy measure is a single atom at $s = h^2$.

Bernstein test for the difference profile. By Theorem E.2, in the present radial setting, contraction in $\mathbf{MMD}^2$ reduces to checking whether the difference profile ψ_τ is Bernstein. Equivalently,

$$\mathbf{MMD}^2_{k_{\mathrm{rbf}}}(f\#\nu, f\#\mu) \le \tau\, \mathbf{MMD}^2_{k_{\mathrm{rbf}}}(\nu, \mu) \qquad \text{for every } d \ge 1 \text{ and every } \mu, \nu \in \mathcal{P}(\mathbb{R}^d)$$

holds if and only if ψ_τ is Bernstein.

Via finitely many atoms. Since $\Lambda = \delta_{h^2}$ has a single atom, Proposition E.4 rules out the required measure domination condition. Indeed, the case $\tau \le \lambda^2$ is impossible because atoms must vanish, while for $\tau > \lambda^2$ the atoms of Λ must lie on disjoint tails of geometric chains, so there must be infinitely many of them, not a single one. Hence Γ_τ cannot be CND in every ambient dimension for any $\tau \in (0, 1)$.

Via saturating profiles. Proposition E.3 gives the same conclusion, since $\lim_{t\to\infty} \phi(t) = 1$, so ϕ is saturating.

Contraction factor. Under the fixed affine map and the RBF kernel, there is no contraction factor in $\mathbf{MMD}^2$. Equivalently,

$$\tau^\star(\lambda) = \inf\{\tau > 0 :\ \Gamma_\tau \text{ is CND on } \mathbb{R}^d \text{ for every } d \ge 1\} \ge 1.$$

Hence no contraction holds in $\mathbf{MMD}^2_{k_{\mathrm{rbf}}}$, and therefore no contraction holds in $\mathbf{MMD}_{k_{\mathrm{rbf}}}$ either. □

G Some new kernels inducing contraction

Preamble. This section presents kernels that induce contraction and, to the best of our knowledge, are new. First, Section G.1 goes through the complete Bernstein functions listed in [29]. We use Theorem E.5, which gives the smallest contraction factor across dimensions as a supremum over one variable, to turn the contraction test into a numerical optimization problem. Second, Section G.2 applies the geometric chaining of Section E.7 to build kernels that induce contraction with a prescribed contraction factor. This gives the proof of Theorem 6.1 of the main text, where the kernel is built from RBF kernels.

G.1 From complete Bernstein functions

Methodology. We started from the tables of complete Bernstein functions in [29] and retained only the entries for which an explicit Lévy density $m(s)$ is available, since this is the regime needed to apply Theorem E.5. We then discarded the entries for which the profile ϕ or the density m involves hypergeometric functions or other special functions not listed below. The retained entries are reported in Tables 3–6.

Notation for special functions. In Tables 3–6, we use the notation for special functions of [29, Section 15.1]. The *Euler gamma function* is

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\, dt.$$

The *lower and upper incomplete gamma functions* are

$$\gamma(a,x) = \int_0^x t^{a-1} e^{-t}\, dt, \qquad \Gamma(a,x) = \int_x^\infty t^{a-1} e^{-t}\, dt$$

(Entries 14, 63, 64, 85, 86). The *modified Bessel functions* are

$$I_\nu(x) = \sum_{n=0}^\infty \frac{(x/2)^{\nu+2n}}{n!\,\Gamma(\nu+n+1)}, \qquad K_\nu(x) = \frac{\pi}{2}\,\frac{I_{-\nu}(x) - I_\nu(x)}{\sin(\nu\pi)},$$

so that $I_0 = I_\nu|_{\nu=0}$, $I_1 = I_\nu|_{\nu=1}$, and $K_0 = K_\nu|_{\nu=0}$ (Entries 7, 57, 77, 78). The *exponential integral* is

$$\mathrm{Ei}(x) = -\int_{-x}^\infty \frac{e^{-t}}{t}\, dt,$$

interpreted as a Cauchy principal value when $x > 0$ (Entries 29, 30, 33, 100–104). The *error and complementary error functions* are

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt, \qquad \mathrm{Erfc}(x) = 1 - \mathrm{Erf}(x)$$

(Entries 38, 108, 109). The *generalized exponential integral* is

$$E_n(x) = \int_1^\infty \frac{e^{-xt}}{t^n}\, dt$$

(Entry 99). The *inverse hyperbolic cosine* is

$$\mathrm{arcosh}(x) = \log\bigl(x + \sqrt{x^2-1}\bigr)$$

(Entries 77, 78). Finally, $\log\Gamma$ denotes the logarithm of the Euler gamma function, namely $x \mapsto \log\Gamma(x)$ (Entry 88).

Entry	Profile ϕ, Lévy density m, and parameter range	Ana.	Num.
1	$\phi(t) = t^\alpha$ $m(s) = \dfrac{\alpha}{\Gamma(1-\alpha)}\, s^{-1-\alpha}$ Valid for $0 < \alpha < 1$	V	V
2	$\phi(t) = (t+1)^\alpha - 1$ $m(s) = \dfrac{\alpha}{\Gamma(1-\alpha)}\, e^{-s} s^{-1-\alpha}$ Valid for $0 < \alpha < 1$	V	V
3	$\phi(t) = 1 - (1+t)^{\alpha-1}$ $m(s) = \dfrac{1}{\Gamma(1-\alpha)}\, e^{-s} s^{-\alpha}$ Valid for $0 < \alpha < 1$	X	X
4	$\phi(t) = \dfrac{t}{t+a}$ $m(s) = a\, e^{-as}$ Valid for $a > 0$	X	X
5	$\phi(t) = \dfrac{t}{\sqrt{t+a}}$ $m(s) = \dfrac{e^{-as}(2as+1)}{2\sqrt{\pi}\, s^{3/2}}$ Valid for $a > 0$		V
6	$\phi(t) = \dfrac{t}{(t+a)^\alpha}$ $m(s) = \dfrac{\sin(\alpha\pi)\Gamma(1-\alpha)}{\pi}\, e^{-as} s^{\alpha-2}(as+1-\alpha)$ Valid for $a > 0,\ 0 < \alpha < 1$		V
7	$\phi(t) = \sqrt{\dfrac{t}{t+a}}$ $m(s) = \dfrac{a}{2} e^{-as/2} \left(I_0\left(\dfrac{as}{2}\right) - I_1\left(\dfrac{as}{2}\right)\right)$ Valid for $a > 0$		X
14	$\phi(t) = \dfrac{\pi}{\sin(\alpha\pi)} \dfrac{t(c^\alpha - t^\alpha)}{c-t}$ $m(s) = \Gamma(\alpha+2) c^{\alpha+1} e^{cs} \Gamma(-\alpha-1, cs)$ Valid for $c > 0,\ -1 < \alpha < 1$		V
18	$\phi(t) = \sqrt{t}\left(1 - e^{-2a\sqrt{t}}\right)$ $m(s) = \dfrac{(2a^2 - s)e^{-a^2/s} + s}{2\sqrt{\pi}\, s^{5/2}}$ Valid for $a > 0$		V
19	$\phi(t) = \sqrt{t}\left(1 + e^{-2a\sqrt{t}}\right)$ $m(s) = \dfrac{(s - 2a^2)e^{-a^2/s} + s}{2\sqrt{\pi}\, s^{5/2}}$ Valid for $a > 0$		X
20	$\phi(t) = \dfrac{t\left(1 - e^{-2\sqrt{t+a}}\right)}{\sqrt{t+a}}$ $m(s) = \dfrac{e^{-1/s-as}\left(2 + s(e^{1/s} - 1)(1+2as)\right)}{2\sqrt{\pi}\, s^{5/2}}$ Valid for $a > 0$		V

Table 3: Complete Bernstein functions of [29] with an explicit Lévy density. Entry numbers refer to [29]. Our table departs from [29] in nine entries. For Entries 30, 104 and 108, we correct ϕ, and for Entries 82 and 84, we correct the constant factor in m. For Entry 85, we restrict the range $\nu \in \mathbb{R}$ to $\nu < 1$, where m is positive. For Entries 64, 131 and 132, we write m or ϕ in another form. The last two columns tell whether the entry induces contraction (V) or not (X), according to an analytic result of this paper (Ana.) and to the numerical search (Num.). An empty Ana. cell means that the paper has no analytic result for that entry.

Entry	Profile ϕ, Lévy density m, and parameter range	Ana.	Num.
26	$\phi(t) = \log\left(1 + \frac{t}{a}\right)$ $m(s) = \frac{e^{-as}}{s}$ Valid for $a > 0$	X	X
27	$\phi(t) = t\log\left(1 + \frac{a}{t}\right)$ $m(s) = \frac{1 - e^{-as}(1+as)}{s^2}$ Valid for $a > 0$		X
28	$\phi(t) = \log\left(\frac{b(t+a)}{a(t+b)}\right)$ $m(s) = \frac{e^{-as} - e^{-bs}}{s}$ Valid for $0 < a < b$		X
29	$\phi(t) = \frac{1}{a} - \frac{1}{t}\log\left(1 + \frac{t}{a}\right)$ $m(s) = -\operatorname{Ei}(-as)$ Valid for $a > 0$		X
30	$\phi(t) = \frac{a+t}{at}\log\left(1 + \frac{t}{a}\right) - \frac{1}{a}$ $m(s) = \frac{e^{-as}}{as} + \operatorname{Ei}(-as)$ Valid for $a > 0$		X
31	$\phi(t) = (t+a)\log(t+a) - t\log t - a\log a$ $m(s) = \frac{1 - e^{-as}}{s^2}$ Valid for $a > 0$		X
32	$\phi(t) = (t+b)\log(t+b) - b\log b - (t+a)\log(t+a) + a\log a$ $m(s) = \frac{e^{-as} - e^{-bs}}{s^2}$ Valid for $0 < a < b$		X
33	$\phi(t) = \frac{t}{a-t}\log\left(\frac{a}{t}\right)$ $m(s) = ae^{as}\operatorname{Ei}(-as) + s^{-1}$ Valid for $a > 0$		X
38	$\phi(t) = \frac{1}{2}\left(a^2 + a\sqrt{a^2+2t}\right) - a^2 + t\log\left(\sqrt{1 + \frac{a^2}{2t}} + \frac{a}{\sqrt{2t}}\right)$ $m(s) = \frac{1}{2s^2}\operatorname{Erf}\left(a\sqrt{\frac{s}{2}}\right)$ Valid for $a > 0$		V
45	$\phi(t) = t(1+t)\log\left(1 + \frac{1}{t}\right) - t$ $m(s) = \frac{e^{-s}(2+s) + s - 2}{s^3}$ No free parameter.		X
46	$\phi(t) = t\left(1 - t\log\left(1 + \frac{1}{t}\right)\right)$ $m(s) = \frac{2 - e^{-s}(2 + 2s + s^2)}{s^3}$ No free parameter.		X

Table 4: Continuation of Table 3.

Entry	Profile ϕ, Lévy density m, and parameter range	Ana.	Num.
57	$\phi(t) = \log\left(1 + 2t + 2\sqrt{t(1+t)}\right)$ $m(s) = s^{-1} e^{-s/2} I_0\left(\frac{s}{2}\right)$ No free parameter.		X
63	$\phi(t) = \sqrt{t}\, \arctan\left(\sqrt{\frac{a}{t}}\right)$ $m(s) = \frac{\gamma\left(\frac{3}{2}, as\right)}{2s^{3/2}}$ Valid for $a > 0$		X
64	$\phi(t) = \sqrt{t}\, \arctan\left(\sqrt{\frac{t}{a}}\right)$ $m(s) = \frac{\Gamma\left(\frac{3}{2}, as\right)}{2s^{3/2}}$ Valid for $a > 0$		V
77	$\phi(t) = \operatorname{arcosh}(1+t) = \log\left(1 + t + \sqrt{t(t+2)}\right)$ $m(s) = \frac{e^{-s} I_0(s)}{s}$ No free parameter.		X
78	$\phi(t) = \frac{1}{2} \operatorname{arcosh}^2(1+t)$ $m(s) = \frac{e^{-s} K_0(s)}{s}$ No free parameter.		X
82	$\phi(t) = \frac{\Gamma\left(\frac{t+a}{2a}\right)}{\Gamma\left(\frac{t}{2a}\right)}$ $m(s) = \frac{a e^{2as}}{\sqrt{\pi}\left(e^{2as} - 1\right)^{3/2}}$ Valid for $a > 0$		V
83	$\phi(t) = \frac{\Gamma(\alpha t + 1)}{\Gamma(\alpha t + 1 - \alpha)}$ $m(s) = \frac{e^{-s/\alpha}}{\Gamma(1-\alpha)\left(1 - e^{-s/\alpha}\right)^{1+\alpha}}$ Valid for $0 < \alpha < 1$		V
84	$\phi(t) = \frac{t\,\Gamma(\alpha t + 1 - \alpha)}{\Gamma(\alpha t + 1)}$ $m(s) = \frac{(1-\alpha) e^{s/\alpha}}{\alpha \Gamma(1+\alpha)\left(e^{s/\alpha} - 1\right)^{2-\alpha}}$ Valid for $0 < \alpha < 1$		V
85	$\phi(t) = t^{1-\nu} e^t \Gamma(\nu, t)$ $m(s) = \frac{\Gamma(2-\nu)}{\Gamma(1-\nu)} (1+s)^{\nu-2}$ Valid for $\nu < 1$		X
86	$\phi(t) = t^{1-\nu} e^{at} \Gamma(\nu, at)$ $m(s) = \frac{\Gamma(2-\nu)}{\Gamma(1-\nu)} (a+s)^{\nu-2}$ Valid for $a > 0,\ 0 < \nu < 1$		X
88	$\phi(t) = \log\Gamma(t+a+b) - \log\Gamma(t+a) + \log\Gamma(a) - \log\Gamma(a+b)$ $m(s) = \frac{e^{-as}\left(1 - e^{-bs}\right)}{s\left(1 - e^{-s}\right)}$ Valid for $a, b > 0$		X

Table 5: Continuation of Table 3.

Entry	Profile ϕ, Lévy density m, and parameter range	Ana.	Num.
99	$\phi(t) = te^t E_n(t)$ $m(s) = \dfrac{n}{(1+s)^{n+1}}$ Valid for $n \in \mathbb{N}$	X	X
100	$\phi(t) = -te^{at}\,\mathrm{Ei}(-at)$ $m(s) = \dfrac{1}{(a+s)^2}$ Valid for $a > 0$	X	X
101	$\phi(t) = te^{at}\left(\mathrm{Ei}(-a(b+t)) - \mathrm{Ei}(-at)\right)$ $m(s) = \dfrac{1-(b(s+a)+1)e^{-(s+a)b}}{(s+a)^2}$ Valid for $a, b > 0$		X
102	$\phi(t) = -te^{at}\,\mathrm{Ei}(-a(b+t))$ $m(s) = \dfrac{(b(s+a)+1)e^{-(s+a)b}}{(s+a)^2}$ Valid for $a, b > 0$	X	X
103	$\phi(t) = (-1)^{n+1}t^{n+1}e^{at}\,\mathrm{Ei}(-at) + \sum_{k=1}^{n}(-1)^{n-k}(k-1)!a^{-k}t^{n-k+1}$ $m(s) = \dfrac{(n+1)!}{(a+s)^{n+2}}$ Valid for $a > 0,\ n \in \mathbb{N}$	X	X
104	$\phi(t) = \gamma_E + \log(at) - e^{at}\,\mathrm{Ei}(-at)$ $m(s) = \dfrac{a}{s(a+s)}$ Valid for $a > 0$		X
108	$\phi(t) = \sqrt{\pi t}\,e^{at}\,\mathrm{Erfc}\left(\sqrt{at}\right)$ $m(s) = \dfrac{1}{2(a+s)^{3/2}}$ Valid for $a > 0$	X	X
109	$\phi(t) = \sqrt{\dfrac{\pi}{a}}\,t - \pi t^{3/2}e^{at}\,\mathrm{Erfc}\left(\sqrt{at}\right)$ $m(s) = \Gamma\left(\dfrac{5}{2}\right)(a+s)^{-5/2}$ Valid for $a > 0$	X	X
131	$\phi(t) = -(2a)^\nu\Gamma(-\nu)\,\dfrac{t\,\Gamma\left(1+\nu+\frac{t}{2a}\right)}{(2a\nu+t)\Gamma\left(1+\frac{t}{2a}\right)}$ $m(s) = \left(\dfrac{a}{\sinh(as)}\right)^{1+\nu} e^{(1-\nu)as}$ Valid for $a > 0,\ 0 < \nu < 1$		V
132	$\phi(t) = (2a)^\nu\Gamma(-\nu)\left(\dfrac{\Gamma(\mu)}{\Gamma(1-\mu)} - \dfrac{\Gamma\left(\mu+\frac{t}{2a}\right)}{\Gamma\left(1-\mu+\frac{t}{2a}\right)}\right)$ $m(s) = \left(\dfrac{a}{\sinh(as)}\right)^{1+\nu}$ Valid for $\mu = \dfrac{1+\nu}{2},\ a > 0,\ 0 < \nu < 1$		V

Table 6: Continuation of Table 3.

Reduction to a supremum. For $\lambda \in (0,1)$, Theorem E.5 gives the smallest contraction factor across dimensions of a complete Bernstein function as

$$\tau^\star(\lambda) = \max\left\{\lambda^2 \mathbf{1}_{\{b>0\}},\ \sup_{s>0} R_\lambda(s)\right\}, \qquad R_\lambda(s) := \frac{m(s/\lambda^2)}{\lambda^2\, m(s)}.$$

Every entry of Tables 3–6 has drift $b = 0$, so $\tau^\star(\lambda)$ is the supremum of R_λ over $s > 0$. Deciding whether an entry induces contraction thus amounts to computing, for each λ, the supremum of a function of one variable.

Two difficulties. Computing this supremum numerically raises two difficulties. First, s ranges over $(0, \infty)$, but a numerical search can only explore a bounded range. The supremum may even

Algorithm 2 Search for the supremum of the ratio R_λ

Require: Discount $\lambda \in (0,1)$, Lévy density $m : (0,\infty) \to (0,\infty)$
Require: Parameters of Table 7 $U, n_{\text{grid}}, K, c_{\text{loc}}$
Output: Estimate $\widehat{\tau}^\star(\lambda)$ of $\sup_{s>0} R_\lambda(s)$, best point $u^\star$, $s^\star = e^{u^\star}$, and whether $u^\star$ is an end of the window
Set

$$\ell_\lambda(u) \leftarrow \log m(e^u/\lambda^2) - \log m(e^u) - \log(\lambda^2)$$

Evaluate the objective ℓ_λ on a grid over the window.
Form a uniform grid $u_1, \dots, u_{n_{\text{grid}}}$ on $[-U, U]$
Evaluate $\ell_\lambda(u_j)$ on the grid and select the K grid points with the largest values
Set

$$\Delta u \leftarrow \frac{2U}{n_{\text{grid}} - 1}$$

for each of the K selected grid points u_j **do**
    Define the local interval

$$[a_j, b_j] \leftarrow \big[\max(-U,\, u_j - c_{\text{loc}}\Delta u),\ \min(U,\, u_j + c_{\text{loc}}\Delta u)\big]$$

    Maximize ℓ_λ on $[a_j, b_j]$ with `minimize_scalar`, giving the point $\tilde{u}_j$ and its value $\tilde{L}_j$
end for
Let $(u^\star, L^\star)$ be the pair $(\tilde{u}_j, \tilde{L}_j)$ with the largest value $\tilde{L}_j$
return $\exp(L^\star), u^\star, e^{u^\star}$, and whether $|u^\star| = U$

be approached only as $s \to 0^+$ or as $s \to \infty$. The search mitigates this with a wide window, and reports when its best point lies at an end of the window, since the supremum may then lie further out. Second, the shape of R_λ is not known in advance, and a single local search may stop at a local maximum below the supremum. The search mitigates this by combining a fine grid over the whole window with local searches around its best points.

Logarithmic scale. To treat small and large values of s alike, we write $s = e^u$ and search over $u \in \mathbb{R}$. A window $[-U, U]$ in u then covers s from e^{-U} to e^U, and a uniform grid in u puts about the same number of points in each interval $[10^k, 10^{k+1}]$ of s. We also maximize the logarithm of the ratio,

$$\ell_\lambda(u) := \log R_\lambda(e^u) = \log m\left(\frac{e^u}{\lambda^2}\right) - \log m(e^u) - \log(\lambda^2),$$

which avoids overflow and underflow when m is very large or very small. Since the logarithm is increasing,

$$\tau^\star(\lambda) = \sup_{s>0} R_\lambda(s) = \exp\Big(\sup_{u\in\mathbb{R}} \ell_\lambda(u)\Big).$$

The search. The search maximizes ℓ_λ on a window $[-U, U]$ in two steps.

1. **Grid.** It evaluates ℓ_λ at n_{grid} equally spaced points of $[-U, U]$, spaced $\Delta u = 2U/(n_{\text{grid}} - 1)$ apart, and selects the K points with the largest values.
2. **Local searches.** It then solves K local maximization problems, one around each selected point u_0, on the interval $[u_0 - c_{\text{loc}}\Delta u,\ u_0 + c_{\text{loc}}\Delta u]$ cut to $[-U, U]$. We use `minimize_scalar`, the routine from SciPy [35] for optimizing a function of one variable on an interval. The best of the K results is denoted $u^\star$.

The search returns $\exp(\ell_\lambda(u^\star))$ as its estimate of $\tau^\star(\lambda)$, together with whether $u^\star$ is an end of the window. Algorithm 2 gives the details, and Table 7 lists the values of U, n_{grid}, K and c_{loc}. The search does not guarantee that the supremum is found.

Setting	Value	Description
Window U	80	The window is $[-U, U]$.
Grid size n_{grid}	8500	Number of grid points on the window.
Selected grid points K	12	Number of best grid points around which ℓ_λ is maximized locally.
Local interval c_{loc}	120	The local interval around a selected grid point u_0 is $[u_0 - c_{\mathrm{loc}}\Delta u,\ u_0 + c_{\mathrm{loc}}\Delta u]$.
Local tolerance	10^{-10}	Precision on u at which `minimize_scalar` stops.

Table 7: Values of the constants of Algorithm 2.

Experimental setup. We run the search on CPUs, at 100 equally spaced values of λ from 0.01 to 0.99, with the constants of Table 7. Entries with parameters use $\alpha = 0.6$ in Entries 1–3, $\alpha = \nu = 0.5$ in the others, $b = 3$ in Entries 28 and 32, $c = 2$ in Entry 14, and $a = b = n = 1$ otherwise. Figure 9 varies one of these parameters.

Numerical precision. The values of the density m are computed with `mpmath` [24] at 80 significant digits, so that they stay accurate where they are very large or very small. Their logarithms are then rounded to double precision, in which the grid and the local searches work.

Checks against known results. Entries 1 and 2 give the power-distance and generalized multiquadric kernels, whose smallest contraction factors across dimensions are known in closed form (Propositions F.1 and F.2). For both entries, the search recovers these values up to rounding errors. Entry 2 shows that the search can find the supremum even when it is approached only as $s \to 0^+$. Proposition G.1 proves that Entries 3, 4 and 26 do not induce contraction, in agreement with the search. By Corollary E.5.1, an entry does not induce contraction if its Lévy density $m(s)$ has a finite limit as $s \to 0^+$. Entries 99, 100, 102, 103, 108 and 109 meet this condition. More entries can be confirmed with this corollary, but for brevity we only list those for which the limit is immediate.

G.1.1 Entries that do not induce contraction

The search finds two groups of entries that do not induce contraction. For the entries of Figure 6, the estimate of $\tau^\star(\lambda)$ exceeds 1 for some λ. Since the estimate never exceeds $\tau^\star(\lambda)$, these entries do not induce contraction. For the entries of Figure 7, the estimate equals 1 up to numerical precision for every λ. These entries do not induce contraction either, but they may be non-expansive.

Entry 78 also belongs to the second group, although its estimate stays below 1. At every λ, its best point is the left end of the window, $u^\star = -80$, so its estimate in Figure 7 may miss a larger value further left. Further left, the ratio keeps increasing towards 1. We conclude that, for Entry 78, $\tau^\star(\lambda) = 1$ for every $\lambda \in (0, 1)$, although its estimate in Figure 7, limited by the window, stays below 1.

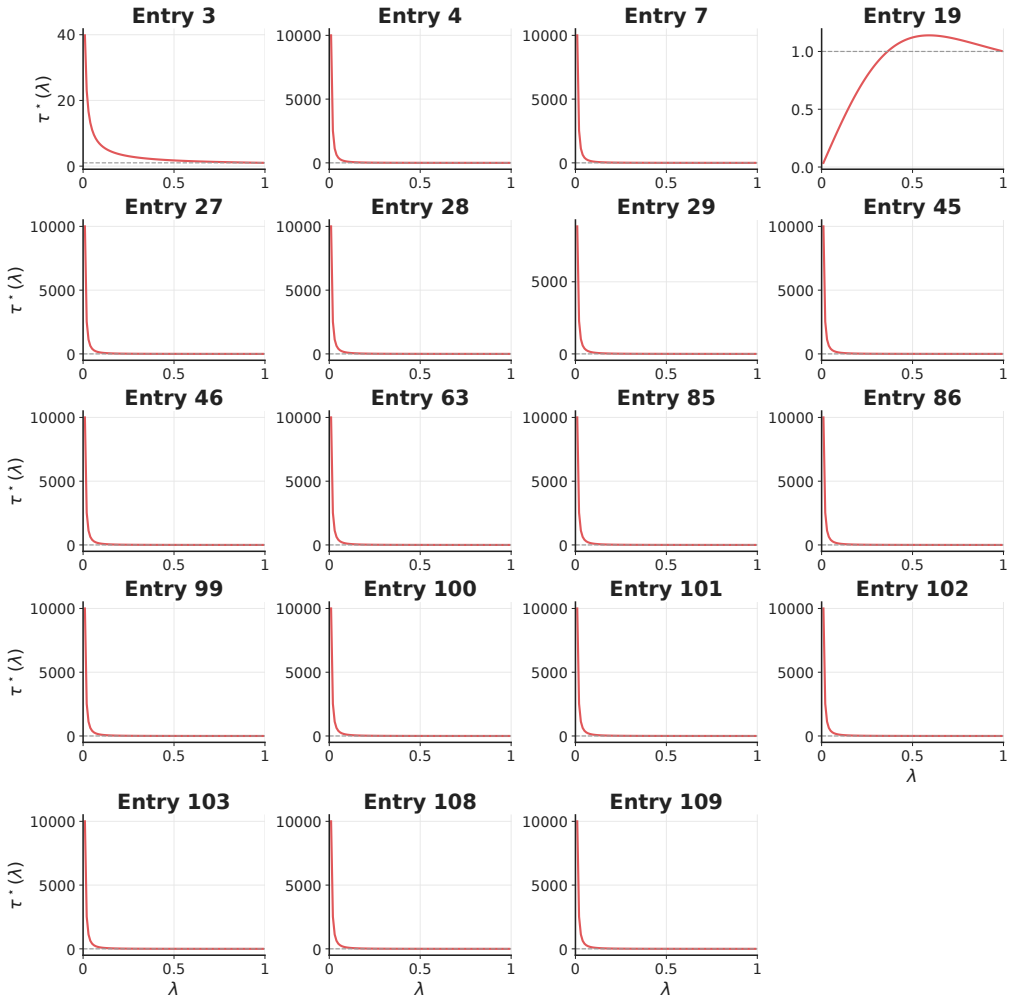

Figure 6: Entries whose estimate of $\tau^\star(\lambda)$ exceeds 1 for some λ. Entry numbers refer to [29].

G.1.2 Entries that induce contraction

We accept an entry when its estimate of $\tau^\star(\lambda)$ stays below 1 for every λ, and its best point is never at an end of the window. Table 8 lists the accepted entries, and Figures 8 and 11 show their estimates of $\tau^\star(\lambda)$ and their profiles ϕ. For an entry with a parameter, these results hold at the parameter values given in the experimental setup. Contraction may then hold only on part of the range allowed in [29]. Figure 9 shows, for seven such entries, how the estimate approaches or crosses 1 as the parameter varies.

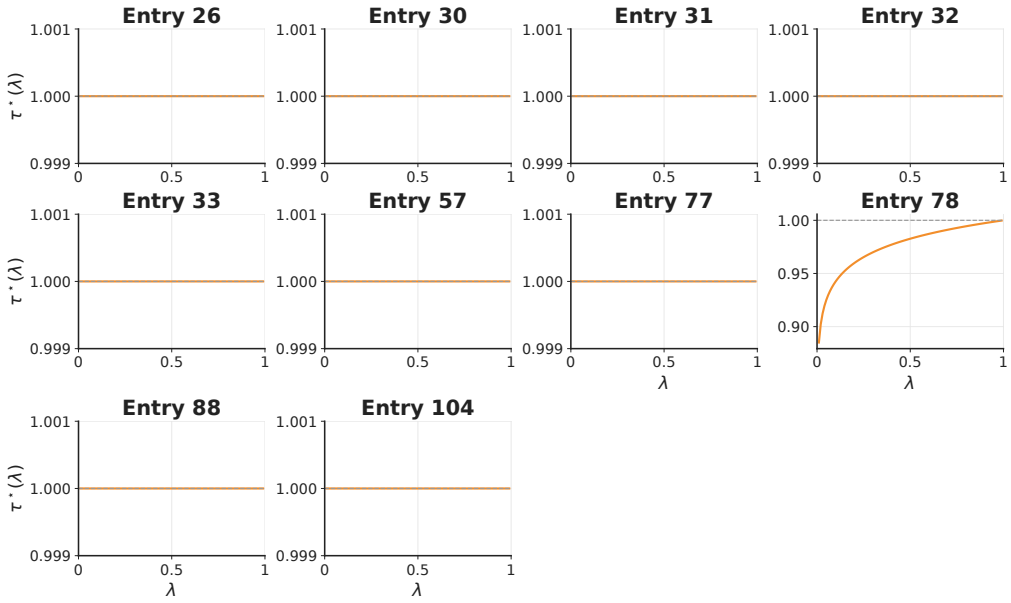

Figure 7: Entries whose estimate of $\tau^\star(\lambda)$ equals 1 up to numerical precision at every λ, and Entry 78. Entry numbers refer to [29].

For seven entries of Table 8, the parameter values giving contraction need no computation. Propositions F.1 and F.2 cover every value of the parameter of Entries 1 and 2. In Entries 5, 18, 38, 64 and 82, the parameter only changes the scale of t, which leaves $\tau^\star(\lambda)$ unchanged. For the other seven, Entries 6, 14, 20, 83, 84, 131 and 132, whether the entry induces contraction may depend on the value of its parameter. We therefore ran the search for many values of the parameter across its range. Figure 10 shows the largest estimate over λ for each value. When this curve crosses 1, a root search on the parameter locates where it does, which gives the range in Table 8.

Entry	Parameter values giving contraction
1	every value
2	every value
5	every value
6	$\alpha \leq 3/4$
14	$\alpha > 0$
18	every value
20	every value
38	every value
64	every value
82	every value
83	every value
84	α up to about 0.70
131	ν from about 0.30
132	every value

Table 8: Entries that induce contraction. Entry numbers refer to [29]. The second column gives the parameter values for which the entry induces contraction, among those allowed in [29]. Entry 1 gives the power-distance kernel studied in [25], and Entry 2 with $\alpha = 1/2$ gives the multiquadric kernel studied in [19]. Entry 5 is Entry 6 with $\alpha = 1/2$.

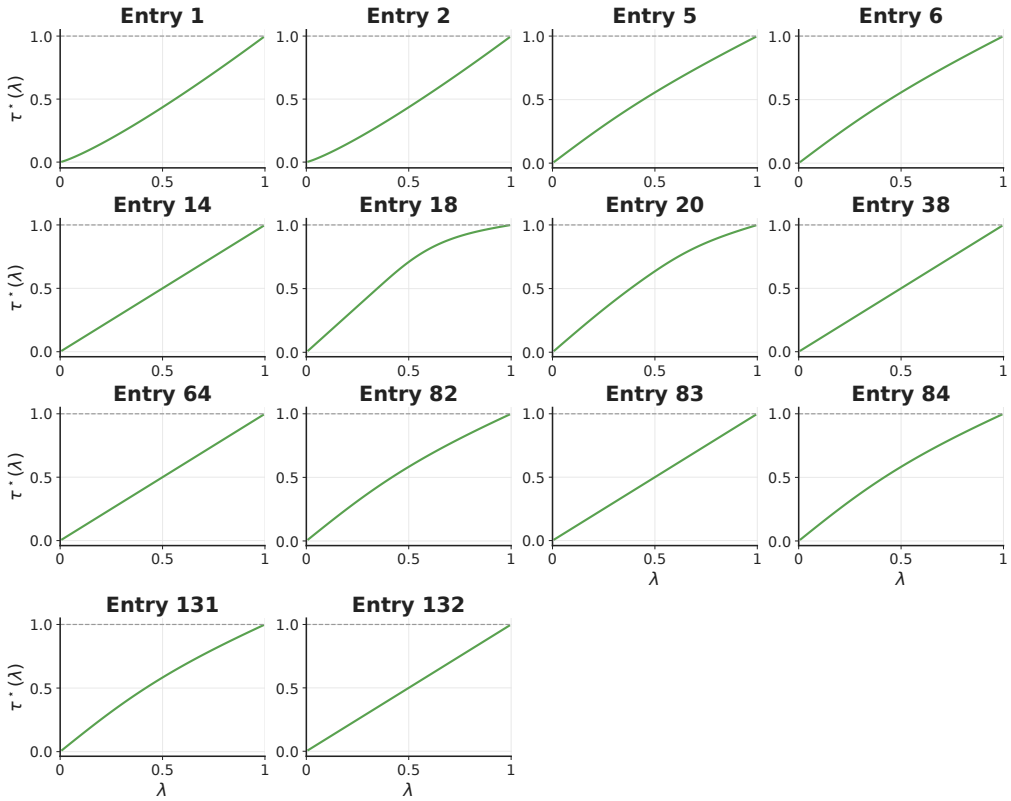

Figure 8: Estimates of $\tau^\star(\lambda)$ for the entries of Table 8. Entry numbers refer to [29].

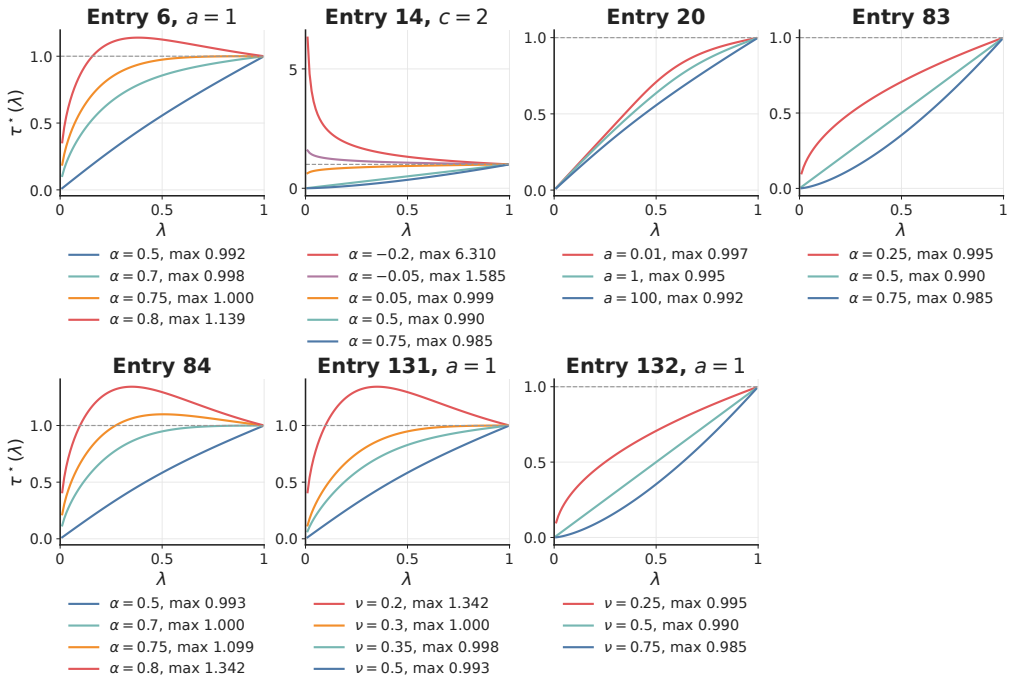

Figure 9: Estimates of $\tau^\star(\lambda)$ for seven entries with a parameter, at several values of the parameter. An entry induces contraction at a value of its parameter when its curve stays below the dashed line $\tau^\star(\lambda) = 1$. Entry numbers refer to [29].

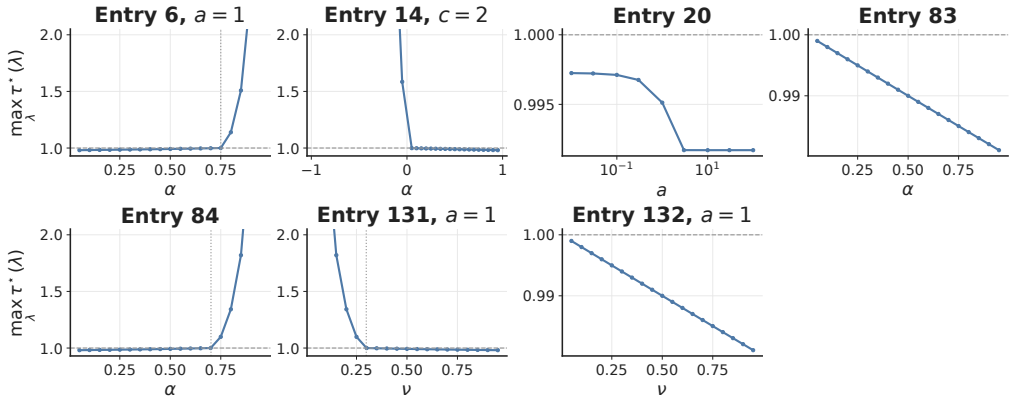

Figure 10: Each panel shows, for one entry of Table 8, the largest estimate of $\tau^\star(\lambda)$ over λ as a function of the entry's parameter. The entry induces contraction where the curve is below the dashed line. The dotted lines mark where the curve crosses 1, found by a root search. Entry numbers refer to [29].

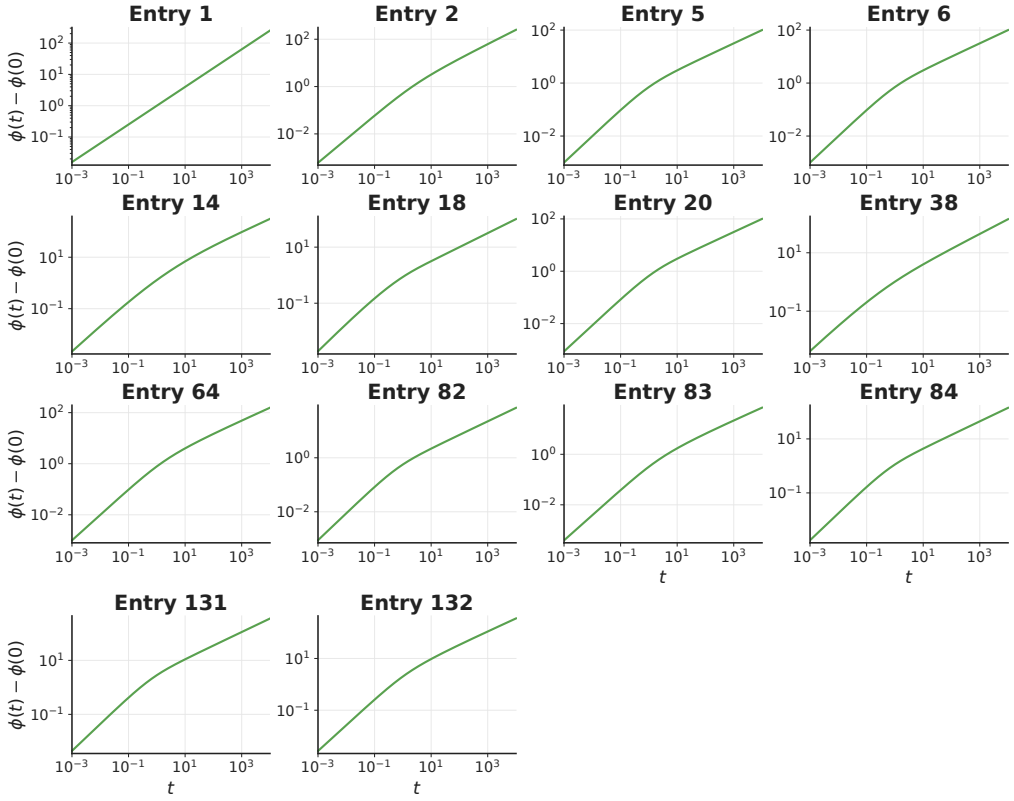

Figure 11: Profiles ϕ of the entries of Table 8, on logarithmic axes. Entry numbers refer to [29].

G.1.3 Entries 3, 4 and 26 in closed form

Proposition G.1 (No contraction under inverse multiquadric, rational and logarithmic kernels)**.** Assume $f(x) = q + \lambda x$ with $\lambda \in (0, 1)$. For each kernel k below, with its profile ϕ, the stated $\tau^\star(\lambda)$ is the smallest τ such that $\mathbf{MMD}_k^2(f\#\nu, f\#\mu) \le \tau\, \mathbf{MMD}_k^2(\nu, \mu)$ for every $d \ge 1$ and every $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$.

(i) Inverse multiquadric kernels, $\alpha \in (0, 1)$:

$$k_\alpha(x, y) = \big(1 + \|x - y\|^2\big)^{\alpha - 1}, \qquad \phi_\alpha(t) = 1 - (1 + t)^{\alpha - 1}, \qquad \tau^\star(\lambda) = \lambda^{2\alpha - 2}.$$

(ii) Rational kernels, $a > 0$:

$$k_a(x, y) = \frac{a}{\|x - y\|^2 + a}, \qquad \phi_a(t) = \frac{t}{t + a}, \qquad \tau^\star(\lambda) = \lambda^{-2}.$$

(iii) Logarithmic kernels, $a > 0$:

$$k_a(x, y) = -\log\left(1 + \frac{\|x - y\|^2}{a}\right), \qquad \phi_a(t) = \log\left(1 + \frac{t}{a}\right), \qquad \tau^\star(\lambda) = 1.$$

For (i) and (ii), $\tau^\star(\lambda) > 1$, so for some d, μ and ν the map f increases $\mathbf{MMD}_k^2$. For (iii), $\tau^\star(\lambda) = 1$, so f is non-expansive in $\mathbf{MMD}_{k_a}^2$, that is,

$$\mathbf{MMD}_{k_a}^2(f\#\nu, f\#\mu) \le \mathbf{MMD}_{k_a}^2(\nu, \mu) \qquad \text{for every } d \ge 1 \text{ and every } \mu, \nu \in \mathcal{P}(\mathbb{R}^d),$$

but f is not a contraction in $\mathbf{MMD}_{k_a}^2$.

Proof. **Lévy–Khintchine representations.** For (i), (ii) and (iii), Entries 3, 4 and 26 of Tables 3–4 give $b = 0$ and the Lévy densities

$$\text{(i) } m(s) = \frac{e^{-s} s^{-\alpha}}{\Gamma(1 - \alpha)}, \qquad \text{(ii) } m(s) = a\, e^{-as}, \qquad \text{(iii) } m(s) = \frac{e^{-as}}{s}, \qquad s > 0.$$

Application of Theorem E.5. Set $\rho = \lambda^2 \in (0, 1)$. Since $b = 0$, Theorem E.5 gives, in each case,

$$\tau^\star(\lambda) = \sup_{s > 0} R_\lambda(s), \qquad R_\lambda(s) := \frac{m(s/\rho)}{\rho\, m(s)}.$$

Using the formula for m in each case, we obtain

$$\text{(i)} \quad R_\lambda(s) = \frac{1}{\rho} \frac{\Gamma(1 - \alpha)^{-1} e^{-s/\rho} (s/\rho)^{-\alpha}}{\Gamma(1 - \alpha)^{-1} e^{-s} s^{-\alpha}} = \rho^{\alpha - 1} \exp\left(-s\left(\frac{1}{\rho} - 1\right)\right),$$

$$\text{(ii)} \quad R_\lambda(s) = \frac{1}{\rho} \frac{a e^{-as/\rho}}{a e^{-as}} = \rho^{-1} \exp\left(-as\left(\frac{1}{\rho} - 1\right)\right),$$

$$\text{(iii)} \quad R_\lambda(s) = \frac{1}{\rho} \frac{e^{-as/\rho}}{s/\rho} \frac{s}{e^{-as}} = \exp\left(-as\left(\frac{1}{\rho} - 1\right)\right).$$

Suprema of the ratios. For (i), (ii) and (iii), we show that the ratios R_λ stay below $\rho^{\alpha-1}$, ρ^{-1} and 1, respectively, and approach these values as $s \to 0^+$, which makes these values their suprema.

Since $0 < \rho < 1$, the factor $1/\rho - 1$ in each exponent above is positive. Each exponent is therefore a negative constant times s, so it is negative for every $s > 0$ and tends to 0 as $s \to 0^+$. Each exponential thus lies strictly between 0 and 1 for every $s > 0$, and tends to 1 as $s \to 0^+$.

In the expressions above, the ratios $R_\lambda(s)$ are these exponentials multiplied by $\rho^{\alpha-1}$ for (i), by ρ^{-1} for (ii) and by 1 for (iii). Hence, for every $s > 0$,

$$\text{(i) } 0 < R_\lambda(s) < \rho^{\alpha - 1}, \qquad \text{(ii) } 0 < R_\lambda(s) < \rho^{-1}, \qquad \text{(iii) } 0 < R_\lambda(s) < 1.$$

The ratios also tend to their upper bounds as $s \to 0^+$. Therefore, these upper bounds are their suprema, and Theorem E.5 yields

$$\text{(i) } \tau^\star(\lambda) = \rho^{\alpha - 1} = \lambda^{2\alpha - 2}, \qquad \text{(ii) } \tau^\star(\lambda) = \rho^{-1} = \lambda^{-2}, \qquad \text{(iii) } \tau^\star(\lambda) = 1.$$

Conclusions. For (i), since $\alpha \in (0, 1)$, we have $2\alpha - 2 < 0$, hence $\tau^\star(\lambda) = \lambda^{2\alpha - 2} > 1$ for every $0 < \lambda < 1$. For (ii), $\tau^\star(\lambda) = \lambda^{-2} > 1$ for every $0 < \lambda < 1$. □

G.2 From geometric chaining

Proposition G.2 (Geometric chaining builds kernels with prescribed contraction factor)**.** Fix $\lambda \in (0,1)$ and set $\rho := \lambda^2 \in (0,1)$. Let $\tau \in (0,1)$, let ϕ_0 be a Bernstein function on $[0,\infty)$ with $\phi_0(0) = 0$, and define

$$h_0(r) := \phi_0(r^2), \qquad r \geq 0.$$

Assume that the chained profile

$$\phi(t) := \sum_{n=0}^{\infty} \tau^{-n} \phi_0(\rho^n t), \qquad t \geq 0,$$

is well defined in the sense of Proposition E.7. Define its radial profile

$$h_{\mathrm{chain}}(r) := \phi(r^2), \qquad r \geq 0,$$

and, for an arbitrary constant $C \in \mathbb{R}$,

$$k_{\mathrm{chain}}(x,y) := C - h_{\mathrm{chain}}(\|x-y\|), \qquad x, y \in \mathbb{R}^d.$$

Then, for every $d \geq 1$, every $q \in \mathbb{R}^d$ and every affine map $f(x) = q + \lambda x$, one has

$$\boxed{\mathbf{MMD}^2_{k_{\mathrm{chain}}}(f\#\nu, f\#\mu) \leq \tau\, \mathbf{MMD}^2_{k_{\mathrm{chain}}}(\nu,\mu) \qquad \forall \mu,\nu \in \mathcal{P}(\mathbb{R}^d).}$$

Moreover, the chained radial profile is given by

$$\boxed{h_{\mathrm{chain}}(r) = \sum_{n=0}^{\infty} \tau^{-n} h_0(\lambda^n r), \qquad r \geq 0,}$$

and satisfies the exact identity

$$\boxed{\tau\, h_{\mathrm{chain}}(r) - h_{\mathrm{chain}}(\lambda r) = \tau\, h_0(r), \qquad r \geq 0.}$$

Thus, whenever the chained profile is well defined, every such affine map f is a τ-contraction in $\mathbf{MMD}^2$ under the radial kernel it produces.

Proof. By Proposition E.7, the chained profile ϕ defines a Bernstein function. Therefore Theorem E.6 applies and yields

$$\mathbf{MMD}^2_{k_{\mathrm{chain}}}(f\#\nu, f\#\mu) \leq \tau\, \mathbf{MMD}^2_{k_{\mathrm{chain}}}(\nu,\mu) \qquad \forall d \geq 1,\ \forall q \in \mathbb{R}^d,\ \forall \mu,\nu \in \mathcal{P}(\mathbb{R}^d),$$

for $f(x) = q + \lambda x$.

It remains to identify the chained radial profile. Since $\rho = \lambda^2$,

$$h_{\mathrm{chain}}(r) = \phi(r^2) = \sum_{n=0}^{\infty} \tau^{-n} \phi_0(\rho^n r^2) = \sum_{n=0}^{\infty} \tau^{-n} \phi_0((\lambda^n r)^2) = \sum_{n=0}^{\infty} \tau^{-n} h_0(\lambda^n r).$$

Finally, Proposition E.6 gives

$$\tau\, \phi(t) - \phi(\rho t) = \tau\, \phi_0(t) \qquad \forall t \geq 0.$$

Substituting $t = r^2$ and using $\rho r^2 = (\lambda r)^2$ yields

$$\tau\, h_{\mathrm{chain}}(r) - h_{\mathrm{chain}}(\lambda r) = \tau\, h_0(r), \qquad r \geq 0.$$

□

Proposition G.3 (The chained RBF kernel realizes any prescribed smallest contraction factor across dimensions)**.** Fix $\lambda \in (0,1)$ and set $\rho := \lambda^2 \in (0,1)$. Let $\tau \in (\rho, 1)$, $a > 0$, and $c > 0$, and define

$$\phi_{\mathrm{chain}}(t) := \sum_{n=0}^{\infty} c\, \tau^{-n} \big(1 - e^{-a\rho^n t}\big), \qquad h_{\mathrm{chain}}(r) := \phi_{\mathrm{chain}}(r^2).$$

For an arbitrary constant $C \in \mathbb{R}$, let

$$k_{\mathrm{chain}}(x, y) := C - h_{\mathrm{chain}}(\|x - y\|), \qquad x, y \in \mathbb{R}^d.$$

Then, for every $d \geq 1$, every $q \in \mathbb{R}^d$, and every affine map

$$f(x) = q + \lambda x,$$

one has

$$\mathbf{MMD}^2_{k_{\mathrm{chain}}}(f\#\nu, f\#\mu) \leq \tau\, \mathbf{MMD}^2_{k_{\mathrm{chain}}}(\nu, \mu) \qquad \forall \mu, \nu \in \mathcal{P}(\mathbb{R}^d).$$

Moreover, τ is the smallest contraction factor for k_{chain}, namely

$$\tau^\star(\lambda) = \tau.$$

Thus, for every prescribed factor $\tau \in (\lambda^2, 1)$, this construction yields a kernel whose smallest contraction factor across dimensions is exactly τ. For the chained RBF kernel, smaller choices of τ lead to faster contraction, and

$$\inf_{\tau > \lambda^2} \tau^\star(\lambda) = \lambda^2,$$

although this value is not attained since the construction requires $\tau > \lambda^2$.

If we introduce the lengthscales

$$\ell_n := \frac{1}{\sqrt{a\rho^n}} = \frac{1}{\sqrt{a}}\, \lambda^{-n}, \qquad n \geq 0,$$

then $a\rho^n = 1/\ell_n^2$, and

$$h_{\mathrm{chain}}(r) = \sum_{n=0}^{\infty} c\, \tau^{-n} \Big(1 - \exp\big(-r^2/\ell_n^2\big)\Big), \qquad r \geq 0.$$

Thus the geometric rescaling $r \mapsto \lambda^n r$ can be viewed as a series of RBF kernels with geometrically increasing lengthscales $(\ell_n)_{n \geq 0}$.

Proof.

Smallest contraction factor. The base Bernstein function

$$\phi_0(t) := c\big(1 - e^{-at}\big)$$

has zero drift and a Lévy measure with a single atom, $c\,\delta_a$. Thus the present construction is exactly the single geometric chain of atoms from Proposition E.5, with $s_0 = a$. Since $\tau > \rho = \lambda^2$, that proposition shows that ϕ_{chain} is a Bernstein function and that, under the induced radial kernel, the affine map $x \mapsto q + \lambda x$ is a τ-contraction in $\mathbf{MMD}^2$. In particular, for the kernel designed with parameter τ, this yields

$$\tau^\star(\lambda) \leq \tau.$$

This is also the RBF case of the general chaining results, namely Propositions E.6 and E.7 together with Proposition G.2.

It remains to prove

$$\tau^\star(\lambda) \geq \tau,$$

since together with the already established $\tau^\star(\lambda) \leq \tau$, this yields

$$\tau^\star(\lambda) = \tau.$$

Let $\bar{\tau} > 0$ be any contraction factor in $\mathbf{MMD}^2_{k_{\mathrm{chain}}}$. Write

$$s_n := \rho^n a, \qquad \Lambda(ds) = \sum_{n=0}^{\infty} c\, \tau^{-n}\, \delta_{s_n}(ds).$$

By condition 3 of Theorem E.4, one must have

$$\bar{\tau}\, \Lambda(\{x\}) \geq \Lambda(\{x/\rho\}) \qquad \text{for every } x > 0.$$

Applying this with $x = s_n$ for $n \geq 1$ gives

$$\bar{\tau}\,\Lambda(\{s_n\}) \geq \Lambda(\{s_{n-1}\}).$$

Since

$$\Lambda(\{s_n\}) = c\,\tau^{-n}, \qquad \Lambda(\{s_{n-1}\}) = c\,\tau^{-(n-1)},$$

it follows that

$$\bar{\tau}\,c\,\tau^{-n} \geq c\,\tau^{-(n-1)},$$

hence

$$\bar{\tau} \geq \tau.$$

Therefore $\tau^\star(\lambda) \geq \tau$, and so $\tau^\star(\lambda) = \tau$. Taking the infimum over $\tau > \lambda^2$ gives

$$\inf_{\tau > \lambda^2} \tau^\star(\lambda) = \lambda^2.$$

Explicit form as a series of RBF kernels. By definition,

$$\phi_{\text{chain}}(t) = \sum_{n=0}^{\infty} c\,\tau^{-n}\big(1 - e^{-a\rho^n t}\big),$$

and therefore

$$h_{\text{chain}}(r) = \phi_{\text{chain}}(r^2) = \sum_{n=0}^{\infty} c\,\tau^{-n}\Big(1 - e^{-a\rho^n r^2}\Big).$$

Introducing

$$\ell_n = \frac{1}{\sqrt{a\rho^n}} = \frac{1}{\sqrt{a}}\lambda^{-n}$$

gives $a\rho^n = 1/\ell_n^2$, so

$$h_{\text{chain}}(r) = \sum_{n=0}^{\infty} c\,\tau^{-n}\Big(1 - \exp\big(-r^2/\ell_n^2\big)\Big).$$

Finally,

$$k_{\text{chain}}(x, y) = C - \sum_{n=0}^{\infty} c\,\tau^{-n}\Big(1 - \exp\big(-\|x - y\|^2/\ell_n^2\big)\Big),$$

which is the form used in Theorem 6.1. □

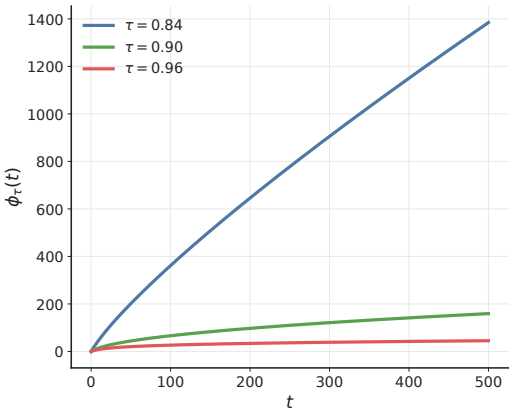

Figure 12: Bernstein functions $\phi_\tau(t) = \sum_{n=0}^{K} c\,\tau^{-n}(1 - e^{-a\gamma^{2n}t})$ for $\tau \in \{0.84, 0.90, 0.96\}$, truncated at $K = 700$ with $\gamma = 0.90$, $a = 0.20$.

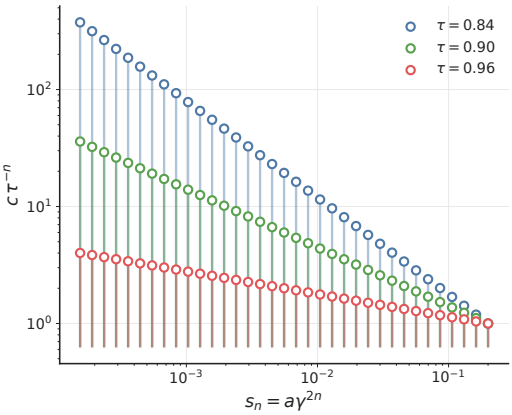

Figure 13: Lévy measure atoms $(s_n, c\,\tau^{-n})$ with $s_n = a\gamma^{2n}$, first 35 atoms shown on a log-log scale.

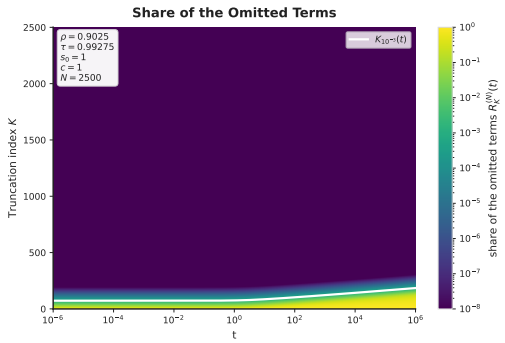

Figure 14: Share of the omitted terms $R_K^{(N)}(t)$ for the chained RBF kernel. The white curve shows the truncation threshold $K_{10^{-3}}(t)$, beyond which the omitted terms account for at most 10^{-3} of the approximated radial profile.

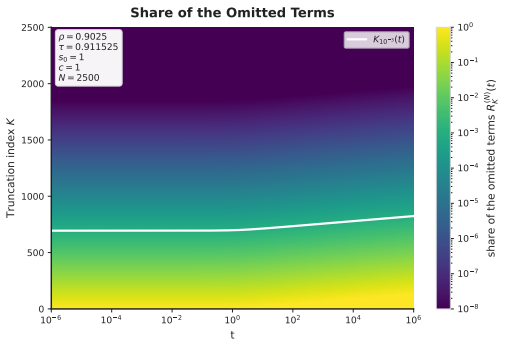

Figure 15: The same share of the omitted terms when ρ/τ is closer to one. The white curve shifts upward, showing that more terms are needed before the omitted terms account for less than 10^{-3} of the approximated radial profile.

Table 9: Contraction of truncated chains on a ball, for the two chains of Figures 14 and 15, with $\rho = 0.9025$, that is $\gamma = 0.95$, and $s_0 = c = 1$. The test of Section 5 uses 10 sets of 512 points drawn at random in the ball of radius 1, in dimension 1, the same for every number of terms. The estimate is the largest contraction threshold over these sets. The full chain is computed with 2500 terms. We compute $R_K^{(N)}(t)$ with $N = 5000$, for every number of terms and squared distances up to 4. The test is run with every number of terms up to 40, then with 50, 60, 80, 100, 150, 200, 300, 400, 500, 700, 1000 and 1500 terms.

	$\tau = 0.99275$	$\tau = 0.911525$
Estimated smallest contraction factor of the full chain	0.964	0.910
Terms needed to pass the test	13	12
Terms needed to be within 10^{-3} of the full chain	60	300
Terms needed for $R_K^{(N)}(t) \leq 10^{-3}$ on the ball	80	706

Numerical estimation. To use the geometric chaining construction as an MMD kernel, the infinite series has to be approximated numerically and therefore truncated. Consider the Bernstein function of the chained RBF kernel

$$\phi(t) = \sum_{n \geq 0} c\,\tau^{-n}\big(1 - \exp(-s_0\rho^n t)\big), \qquad 0 < \rho < \tau < 1,$$

and write its n-th term as

$$\phi_n(t) := c\,\tau^{-n}\big(1 - \exp(-s_0\rho^n t)\big).$$

For a truncation index K, we measure the relative contribution of the omitted terms by

$$R_K(t) := \frac{\sum_{n \geq K} \phi_n(t)}{\sum_{n \geq 0} \phi_n(t)}.$$

In practice, we approximate this quantity by replacing the infinite denominator and numerator with a sufficiently large finite cutoff N,

$$R_K^{(N)}(t) := \frac{\sum_{K \leq n \leq N} c\,\tau^{-n}\big(1 - \exp(-s_0\rho^n t)\big)}{\sum_{0 \leq n \leq N} c\,\tau^{-n}\big(1 - \exp(-s_0\rho^n t)\big)}, \qquad N \gg 1.$$

Given a tolerance level $\alpha > 0$, we define the truncation threshold

$$K_\alpha(t) := \min\left\{K \in \mathbb{N} \mid R_K^{(N)}(t) \leq \alpha\right\}.$$

In the numerical results below, we use $\alpha = 10^{-3}$ and report the corresponding threshold $K_{10^{-3}}(t)$.

As shown in Figure 14, the terms of the chain decay rapidly. For each fixed t, there is a finite index $K_{10^{-3}}(t)$ beyond which the omitted terms account for less than 0.1% of the approximated radial profile. Figure 15 shows the same diagnostic when ρ/τ is closer to one, where the share of the omitted terms decreases more slowly as K grows and a larger truncation index is needed to reach the same 0.1% threshold.

Contraction of truncated chains. A truncated chain is a finite sum of RBF kernels, so by Proposition 4.3 it does not induce contraction on all of $\mathbb{R}^d$. As discussed in Section 5, it can still induce contraction on a bounded set, which we check with the test detailed in Appendix D for the two chains of Figures 14 and 15 (Table 9). About a dozen terms are enough for the truncated chain to pass the test. With more terms, its estimated smallest contraction factor on the ball decreases towards that of the full chain. This limit is below τ, since contraction on a ball is a weaker requirement than on all of $\mathbb{R}^d$. Truncating with the rule above, which makes the profile accurate to 10^{-3} on the ball, already brings the estimated contraction factor of the truncated chain within 10^{-3} of that of the full chain. Both take many more terms when ρ/τ is closer to one.

H Contraction rate and discriminability

Preamble. A kernel with a smaller contraction factor is not always the better choice, because changing the kernel also changes which differences between two distributions the MMD can detect. This section studies this tension, which Section 7 introduces. It starts with two results on the kernels $k(x, y) = C - b\|x - y\|^2$, whose Bernstein function $\phi(t) = bt$ has a zero Lévy measure. First, Proposition H.1 shows that such a kernel only compares the means of two distributions. Second, Proposition H.2 shows that only these kernels can reach the smallest possible contraction factor γ^2. The rest of the section describes the experiments. Section H.1 presents the permutation test we use to measure the power of a kernel, that is, how often it detects that two distributions differ. Section H.2 then explains how we estimate the tradeoff between this power and the contraction factor. Finally, Section H.3 repeats these experiments in higher dimensions.

Proposition H.1 (The kernel $C - b\|x - y\|^2$ compares only means)**.** Let $b \geq 0$ and $C \in \mathbb{R}$, and consider the kernel

$$k(x, y) := C - b\|x - y\|^2, \qquad x, y \in \mathbb{R}^d.$$

Empirical level. Given samples $x_{1:n} = (x_1, \ldots, x_n)$ and $y_{1:m} = (y_1, \ldots, y_m)$, define the empirical means

$$\bar{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y}_m := \frac{1}{m}\sum_{j=1}^{m} y_j.$$

Then the biased empirical estimator satisfies

$$\boxed{\widehat{\mathbf{MMD}}^2_{k,\mathrm{b}}(x_{1:n}, y_{1:m}) = 2b\|\bar{x}_n - \bar{y}_m\|^2}.$$

In particular, the biased empirical $\mathbf{MMD}^2$ depends only on the empirical means of the two samples.

Population level. Similarly, at the population level, let $P, Q \in \mathcal{P}(\mathbb{R}^d)$ have finite second moments. Then

$$\boxed{\mathbf{MMD}^2_k(P, Q) = 2b\|\mathbb{E}_P[X] - \mathbb{E}_Q[X]\|^2}.$$

In particular, $\mathbf{MMD}^2_k(P, Q) = 0$ whenever P and Q have the same first moment.

Proof. **Empirical identity.** Let $N := n + m$, define the aggregate sample

$$z_i := x_i \quad (1 \leq i \leq n), \qquad z_{n+j} := y_j \quad (1 \leq j \leq m),$$

and let

$$\ell := \begin{pmatrix} \frac{1}{n}\mathbf{1}_n \\ -\frac{1}{m}\mathbf{1}_m \end{pmatrix} \in \mathbb{R}^N, \qquad \text{so that } \mathbf{1}_N^\top \ell = 0.$$

Let $K \in \mathbb{R}^{N\times N}$ be the Gram matrix of the aggregate sample, $K_{ij} := k(z_i, z_j)$. By Section A.3,

$$\widehat{\mathbf{MMD}}^2_{k,\mathrm{b}}(x_{1:n}, y_{1:m}) = \ell^\top K \ell.$$

Using $k(z_i, z_j) = C - b\|z_i - z_j\|^2$, we obtain

$$\ell^\top K \ell = C(\mathbf{1}_N^\top \ell)^2 - b\sum_{i,j=1}^{N} \ell_i \ell_j \|z_i - z_j\|^2 = -b\sum_{i,j=1}^{N} \ell_i \ell_j \|z_i - z_j\|^2,$$

since $\mathbf{1}_N^\top \ell = 0$. Applying Lemma I.1 then yields

$$\widehat{\mathbf{MMD}}^2_{k,\mathrm{b}}(x_{1:n}, y_{1:m}) = 2b\Big\|\sum_{i=1}^{N} \ell_i z_i\Big\|^2.$$

Finally,

$$\sum_{i=1}^{N} \ell_i z_i = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{m}\sum_{j=1}^{m} y_j = \bar{x}_n - \bar{y}_m,$$

which gives the result.

Population identity. Let $\sigma := P - Q$, so that $\sigma(\mathbb{R}^d) = 0$. By the definition of $\mathbf{MMD}_k^2$ as a quadratic form,

$$\mathbf{MMD}_k^2(P, Q) = \iint_{\mathbb{R}^d \times \mathbb{R}^d} k(x, y)\, d\sigma(x)\, d\sigma(y).$$

Expanding the kernel gives

$$k(x, y) = C - b\|x\|^2 - b\|y\|^2 + 2b\, x^\top y.$$

Since $\sigma(\mathbb{R}^d) = 0$, the constant term and the terms that depend on only one of x and y vanish in the quadratic form,

$$\iint C\, d\sigma(x)\, d\sigma(y) = C\, \sigma(\mathbb{R}^d)^2 = 0,$$

and

$$\iint \|x\|^2\, d\sigma(x)\, d\sigma(y) = \Big(\int \|x\|^2\, d\sigma(x)\Big)\sigma(\mathbb{R}^d) = 0,$$

with the same identity for the $\|y\|^2$ term. Therefore,

$$\mathbf{MMD}_k^2(P, Q) = 2b \iint x^\top y\, d\sigma(x)\, d\sigma(y) = 2b\Big\|\int x\, d\sigma(x)\Big\|^2.$$

Since $\int x\, d\sigma(x) = \mathbb{E}_P[X] - \mathbb{E}_Q[X]$, the claim follows. □

Proposition H.2 (Only the kernels $C - b\|x-y\|^2$ reach the factor γ^2)**.** Let $k(x, y) = C - \phi(\|x-y\|^2)$ for $x, y \in \mathbb{R}^d$, where ϕ is a Bernstein function with Lévy measure Λ. If $\Lambda \not\equiv 0$, then $\tau_k^\star(\mathcal{T}^\pi) > \gamma^2$. Equivalently, the value γ^2 can only be reached by a kernel of the form $k(x, y) = C - b\|x - y\|^2$.

Proof.

A lower bound from two point masses. We first derive a lower bound on every contraction factor. The second part then shows that this bound exceeds γ^2 when $\Lambda \not\equiv 0$.

A constant in ϕ only shifts C, so we take $\phi(0) = 0$. For two point masses, this gives $\mathbf{MMD}_k^2(\delta_a, \delta_b) = 2\phi(\|a - b\|^2)$. Take the one-state MDP of Lemma C.2 with reward 0, whose Bellman operator multiplies every return by γ. It has a single state-action pair, so the supremum in $\overline{\mathbf{MMD}}_k^2$ is over that pair alone.

Fix $x \in \mathbb{R}^d$ with $x \neq 0$, let $t = \|x\|^2$, and let η_1 and η_2 put the point masses δ_x and δ_0 on that pair. Before a Bellman step,

$$\overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2) = \mathbf{MMD}_k^2(\delta_x, \delta_0) = 2\phi(t).$$

After the step, $\mathcal{T}^\pi\eta_1$ and $\mathcal{T}^\pi\eta_2$ put $\delta_{\gamma x}$ and δ_0 on the pair, so

$$\overline{\mathbf{MMD}}_k^2(\mathcal{T}^\pi\eta_1, \mathcal{T}^\pi\eta_2) = \mathbf{MMD}_k^2(\delta_{\gamma x}, \delta_0) = 2\phi(\gamma^2 t).$$

If $\mathcal{T}^\pi$ is a τ-contraction for this MDP, the definition of contraction in Section 2 applied to η_1 and η_2 gives

$$2\phi(\gamma^2 t) = \overline{\mathbf{MMD}}_k^2(\mathcal{T}^\pi\eta_1, \mathcal{T}^\pi\eta_2) \le \tau\, \overline{\mathbf{MMD}}_k^2(\eta_1, \eta_2) = 2\tau\, \phi(t),$$

so $\phi(\gamma^2 t) \le \tau\, \phi(t)$. The factor $\tau_k^\star(\mathcal{T}^\pi)$ is the smallest one valid for every MDP, so it is valid for this one as well. Hence, whenever $\phi(t) > 0$,

$$\tau_k^\star(\mathcal{T}^\pi) \ge \frac{\phi(\gamma^2 t)}{\phi(t)}.$$

The bound exceeds γ^2 when $\Lambda \not\equiv 0$. In the Lévy–Khintchine representation of ϕ, the terms in b cancel,

$$\begin{aligned}\phi(\gamma^2 t) - \gamma^2\phi(t) &= \big(b\gamma^2 t - \gamma^2 bt\big) + \int_{(0,\infty)} \big(1 - e^{-\gamma^2 st}\big)\, \Lambda(ds) - \gamma^2 \int_{(0,\infty)} \big(1 - e^{-st}\big)\, \Lambda(ds) \\ &= \int_{(0,\infty)} \Big[\big(1 - e^{-\gamma^2 st}\big) - \gamma^2\big(1 - e^{-st}\big)\Big]\, \Lambda(ds).\end{aligned}$$

The bracket has the form $g(\gamma^2 u) - \gamma^2 g(u)$ with $g(u) := 1 - e^{-u}$ and $u = st$, and we show that it is positive for every $u > 0$. First, g is strictly concave on $[0, \infty)$, since $g''(u) = -e^{-u} < 0$, and $g(0) = 0$. Second, since $\gamma^2 \in (0, 1)$, the point $\gamma^2 u$ lies between 0 and u, as the convex combination $\gamma^2 u = \gamma^2 \cdot u + (1 - \gamma^2) \cdot 0$. Strict concavity at this point gives

$$g(\gamma^2 u) = g\big(\gamma^2 u + (1 - \gamma^2) \cdot 0\big) > \gamma^2 g(u) + (1 - \gamma^2)\, g(0) = \gamma^2 g(u).$$

So the bracket is positive for every $s > 0$. When $\Lambda \not\equiv 0$, the integral of this positive function against Λ is positive, and $\phi(t) \geq \int_{(0,\infty)} g(st)\, \Lambda(ds) > 0$. Hence

$$\tau_k^\star(\mathcal{T}^\pi) \geq \frac{\phi(\gamma^2 t)}{\phi(t)} > \gamma^2.$$

Since $\tau^\star(\gamma) \geq \tau_k^\star(\mathcal{T}^\pi)$, as noted after Theorem E.4, the smallest contraction factor across dimensions also exceeds γ^2.

When $\Lambda \equiv 0$. Then $\phi(t) = bt$, and k is the kernel

$$k(x, y) = C - b\, \|x - y\|^2.$$

For this kernel, $\mathbf{MMD}_k$ only compares means (Proposition H.1).

With $\Lambda \equiv 0$, the criterion of Theorem E.4 reduces to $b(\tau - \gamma^2) \geq 0$, which holds for every $\tau \in [\gamma^2, 1)$. Contraction then holds across dimensions, so

$$\tau_k^\star(\mathcal{T}^\pi) \leq \tau^\star(\gamma) \leq \gamma^2.$$

Conversely, when $b > 0$, the lower bound of the first part applies with $\phi(t) = bt > 0$, and its ratio is exactly

$$\frac{\phi(\gamma^2 t)}{\phi(t)} = \frac{b\gamma^2 t}{bt} = \gamma^2.$$

Hence $\tau_k^\star(\mathcal{T}^\pi) = \gamma^2$, and this kernel reaches the value γ^2. When $b = 0$ as well, the kernel is constant and $\mathbf{MMD}_k$ is identically zero. □

H.1 Permutation test and power estimation

$\mathbf{MMD}^2$ values are not comparable across kernels. For example, if $k' = c\,k$ with $c > 0$, then $\mathbf{MMD}^2_{k'} = c\,\mathbf{MMD}^2_k$, so a larger value can reflect only a change of units. Following Borgwardt et al. [7], Gretton et al. [15, 16], we therefore compare kernels through the power of a level-α permutation test built from the biased empirical statistic.

Level and power of a test. A hypothesis test of $H_0 \colon P = Q$ against $H_1 \colon P \neq Q$ is a binary rule $\Delta_k \in \{0, 1\}$, with $\Delta_k = 1$ meaning rejection. It has level α if $\sup_P \mathbb{P}_{P \times P}(\Delta_k = 1) \leq \alpha$, controlling the false-positive rate. Its power against (P, Q) with $P \neq Q$ is $\mathbb{P}_{P \times Q}(\Delta_k = 1)$, the probability of correctly detecting that $P \neq Q$.

Test statistic. Given samples $X_{1:n} \sim P^n$ and $Y_{1:m} \sim Q^m$, let $N := n + m$, form the aggregate sample $Z_{1:N} = (X_{1:n}, Y_{1:m})$, and let $K \in \mathbb{R}^{N \times N}$ be its Gram matrix, $K_{ij} := k(Z_i, Z_j)$. Define the two-sample weight vector $\ell \in \mathbb{R}^N$ as in Section A.3,

$$\ell_i := \begin{cases} \frac{1}{n}, & 1 \leq i \leq n, \\ -\frac{1}{m}, & n + 1 \leq i \leq N, \end{cases}$$

so that the test statistic is

$$T_k := \ell^\top K \ell = \widehat{\mathbf{MMD}}^2_{k,\mathrm{b}}(X_{1:n}, Y_{1:m}).$$

Null hypothesis and permutation test. Under H_0, the labels are exchangeable, so permuting $\{1, \ldots, N\}$ leaves the joint distribution of the aggregate sample unchanged. Given B_{perm} uniformly drawn permutations $\pi_1, \ldots, \pi_{B_{\mathrm{perm}}}$ of $\{1, \ldots, N\}$, each induces a permuted statistic $T_{k,\pi_b} := \ell^\top K_{\pi_b} \ell$, where $(K_{\pi_b})_{ij} := k(Z_{\pi_b(i)}, Z_{\pi_b(j)})$. The permutation p-value is

$$\hat{p} := \frac{1 + \#\big\{b \in \{1, \ldots, B_{\mathrm{perm}}\} : T_{k,\pi_b} \geq T_k\big\}}{B_{\mathrm{perm}} + 1},$$

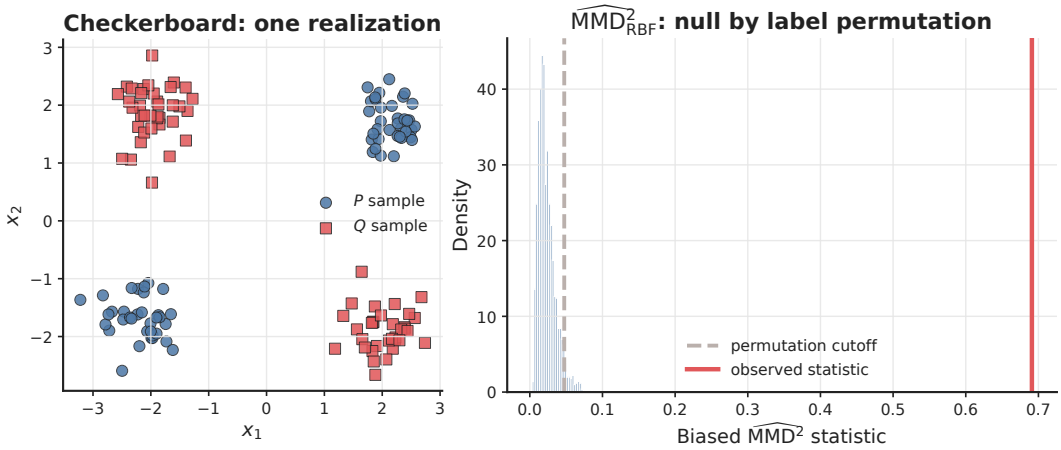

Figure 16: One example of the permutation test used for power estimation, using the RBF kernel with lengthscale $\ell = 1.05$ at level $\alpha = 0.05$. Left, one realization of the checkerboard two-sample problem ($n = m = 70$). Right, the permutation null for the biased $\widehat{\mathbf{MMD}}^2_{\mathrm{RBF}}$ statistic, obtained from 2,000 label permutations on the aggregate sample. The dashed line is the level-α cutoff and the red line is the observed statistic.

Algorithm 3 Estimating the power of the permutation test for one kernel configuration

1: Fix (P, Q), sample sizes (n, m), a level α, an outer repetition count B, and a permutation budget B_{perm}.
2: **for** $b = 1, \ldots, B$ **do**
3: Draw $X^{(b)}_{1:n} \sim P^n$ and $Y^{(b)}_{1:m} \sim Q^m$.
4: Build the Gram matrix $K^{(b)}$ of the aggregate sample and compute $T^{(b)}_k = \widehat{\mathbf{MMD}}^2_{k,\mathrm{b}}$.
5: Permute the labels B_{perm} times on the aggregate sample and recompute the corresponding statistics $T^{(b)}_{k,\pi}$.
6: Compute the permutation p-value

$$\hat{p}^{(b)} \;=\; \frac{1 + \#\{T^{(b)}_{k,\pi} \geq T^{(b)}_k\}}{B_{\mathrm{perm}} + 1}.$$

7: **end for**
8: Return $\widehat{D}_{k,n,m,\alpha}(P, Q) := \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\big[\hat{p}^{(b)} \leq \alpha\big]$.

and H_0 is rejected if and only if $\hat{p} \leq \alpha$.

Figure 16 illustrates one instance of this test. The aggregate sample is fixed, the labels are permuted to form the null distribution, and the observed statistic is compared with the resulting level-α cutoff.

Power estimation. The power of the level-α test for a given kernel k is $\mathbb{P}_{H_1}(\hat{p} \leq \alpha)$. It is estimated by repeating the following procedure B times independently. Each run draws a fresh dataset pair $(X^{(b)}_{1:n}, Y^{(b)}_{1:m})$, computes the observed statistic $T^{(b)}_k$, runs B_{perm} permutations to obtain the p-value $\hat{p}^{(b)}$, and records whether $\hat{p}^{(b)} \leq \alpha$. The empirical power $\widehat{D}_{k,n,m,\alpha}(P, Q)$ is the fraction of rejections across the B repetitions. Algorithm 3 summarizes this procedure.

H.2 Estimating the tradeoff between power and contraction

For a given distribution pair (P, Q) and kernel family $\{k_\theta\}$, lowering the smallest contraction factor on a bounded set and maximizing discriminative power pull in opposite directions. To make this tradeoff visible, we fix a contraction budget $\bar{\tau}$ and report, over all θ with $\widehat{\tau}^{\star}_{k_\theta,E}(\gamma) \leq \bar{\tau}$, the best

Algorithm 4 Measuring the tradeoff between power and contraction

Require: Distribution pair (P, Q), sample sizes (n, m), level α, discount grid $\{\gamma_l\}$, grid of contraction budgets $\bar{\tau}$, parameter grid $\{\theta_i\}$, ball radius R, support size N, trial count J, outer repetition count B, permutation budget B_{perm}.

Step 1: estimate power for each kernel parameter.
for each θ_i **do**
    Run Algorithm 3 with kernel k_{θ_i}. Set $\widehat{D}_i := \widehat{D}_{k_{\theta_i},n,m,\alpha}(P, Q)$.
end for

Step 2: estimate the smallest contraction factor on E for each kernel parameter and discount.
for each θ_i and each γ_l **do**
    Run Algorithm 1 with kernel k_{θ_i}, scaling parameter $\lambda = \gamma_l$, ball $B(0, R)$, support size N, and J trials. Set $\widehat{\tau}_{i,l} := \max_{1\le j\le J} \widehat{\tau}^{\star}_{S_j}(\gamma_l)$.
end for

Step 3: build the tradeoff heatmap.
for each γ_l and each budget $\bar{\tau}$ **do**
    Record $\max_{i\,:\,\widehat{\tau}_{i,l}\le\bar{\tau}} \widehat{D}_i$, or mark the cell as empty if no θ_i satisfies the constraint.
end for

Setting	Value	Description
Support size N	512	Number of points drawn uniformly in $B(0, R)$ for each trial support.
Trial count J	1	Number of independent trial supports used to estimate $\tau^{\star}_{k_\theta,E}(\gamma)$.
Eigenvalue tolerance $\varepsilon_{\mathrm{eig}}$	10^{-9}	Relative tolerance below which eigenvalues of A are discarded before forming the reduced problem.
Outer repetitions B	1000	Number of independent dataset draws used to estimate power.
Permutation budget B_{perm}	1000	Number of label permutations used to compute the p-value on each dataset draw.
Level α	0.05	Significance level of the permutation test.

Table 10: Numerical settings used in Algorithm 4.

achievable power

$$\widehat{D}^{\star}(\gamma, \bar{\tau}) \;:=\; \max\big\{\, \widehat{D}_{k_\theta,n,m,\alpha}(P, Q) : \widehat{\tau}^{\star}_{k_\theta,E}(\gamma) \le \bar{\tau} \,\big\},$$

leaving cells with no feasible θ undefined. Sweeping over $(\gamma, \bar{\tau})$ produces the tradeoff heatmap of Algorithm 4.

Estimation procedure. For each kernel family and parameter grid in Table 11, and for each γ, the smallest contraction factor $\tau^{\star}_{k_\theta,E}(\gamma)$ on the ball $E = B(0, R)$ is estimated by Algorithm 1. For each dataset in Table 12, R is chosen to contain most of the mass and is used to clip the sampled data. The numerical settings are given in Table 10. In particular, we use support size N and J independent random supports, retaining the maximum over the J estimates. For each θ, the power $\widehat{D}_{k_\theta,n,m,\alpha}(P, Q)$ is estimated by Algorithm 3. The full procedure is summarized in Algorithm 4.

Dataset	Distribution	Radius R
1D Mean Shift	$P = \mathcal{N}(-0.40, 0.18^2)$, $Q = \mathcal{N}(0.40, 0.18^2)$	1.40
1D Same Mean, Different Std	$P = \mathcal{N}(0, 0.25^2)$, $Q = \mathcal{N}(0, 0.55^2)$	2.00
1D Bimodal vs Unimodal	$P = \frac{1}{2}\mathcal{N}(-0.65, 0.15^2) + \frac{1}{2}\mathcal{N}(0.65, 0.15^2)$, $Q = \mathcal{N}(0, 0.65^2 + 0.15^2)$	2.00
Two-Gaussian Location	$P = \mathcal{N}\big((0.55, 0), 0.18^2 I_2\big)$, $Q = \mathcal{N}\big((-0.55, 0), 0.18^2 I_2\big)$	1.35
Covariance Swap	$P = \mathcal{N}\left(0, \begin{bmatrix} 0.28 & 0.18 \\ 0.18 & 0.28 \end{bmatrix}\right)$, $Q = \mathcal{N}\left(0, \begin{bmatrix} 0.28 & -0.18 \\ -0.18 & 0.28 \end{bmatrix}\right)$	2.40
Bimodal vs Unimodal	$P = \frac{1}{2}\mathcal{N}\big((-0.75, 0), 0.18^2 I_2\big) + \frac{1}{2}\mathcal{N}\big((0.75, 0), 0.18^2 I_2\big)$, $Q = \mathcal{N}\left(0, \begin{bmatrix} 0.75^2 + 0.18^2 & 0 \\ 0 & 0.18^2 \end{bmatrix}\right)$	2.40
Checkerboard	$P = \frac{1}{2}\mathcal{N}\big((-2.2, -1.7), 0.35^2 I_2\big) + \frac{1}{2}\mathcal{N}\big((2.2, 1.7), 0.35^2 I_2\big)$, $Q = \frac{1}{2}\mathcal{N}\big((-2.0, 1.9), 0.35^2 I_2\big) + \frac{1}{2}\mathcal{N}\big((2.0, -1.9), 0.35^2 I_2\big)$	4.40

Table 12: Datasets used to study the tradeoff between power and contraction. The displayed Gaussian laws are sampled and then clipped to the ball $B(0, R)$.

Kernel family and formula	Parameter θ	Grid over Θ
RBF $k_h(x, y) = \exp\big(-h^2\|x - y\|^2\big)$	Scale h	$\mathrm{geom}(10^{-3}, 10^3, 100)$
Generalized multiquadric $k_{\beta,h}(x, y) = -\big(1 + h^2\|x - y\|^2\big)^{\beta}$	Power β	$\mathrm{lin}(0.05, 1.00, 50)$
Power distance $k_\alpha(x, y) = -\|x - y\|^\alpha$	Power α	$\mathrm{lin}(0.20, 2.00, 50)$
Logarithmic kernel $k_a(x, y) = -\log\big(1 + \|x - y\|^2/a\big)$	Scale a	$\mathrm{geom}(10^{-3}, 10^3, 100)$

Table 11: Kernel families and the values of their parameter used to study the tradeoff between power and contraction. Here $\mathrm{geom}(a, b, n)$ denotes n geometrically spaced values from a to b, and $\mathrm{lin}(a, b, n)$ denotes n linearly spaced values from a to b.

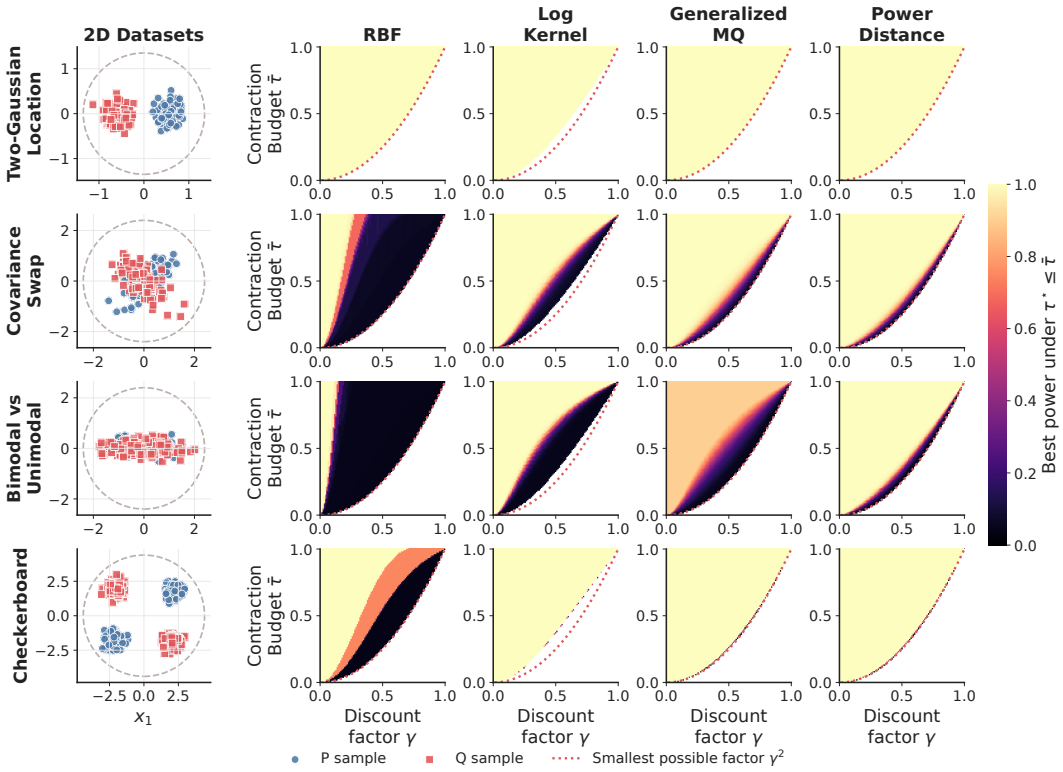

Figure 17: Two-dimensional tradeoff experiments. The left column shows one sampled dataset for each task, with dashed circles marking the clipping radius R. The remaining columns show, for each kernel family, the best empirical test power achievable under the constraint $\tau^\star_{k_\theta,E}(\gamma) \leq \bar{\tau}$. The dotted curve is the smallest possible contraction factor γ^2.

H.3 Higher dimensions

Return distributions are multivariate when rewards are vector valued [13, 37, 38], so the two samples compared by the test can live in higher dimensions. We therefore check whether the tradeoff persists as the dimension grows, on two pairs of distributions defined in any dimension, a mean shift and a checkerboard on the corners of a cube (Table 13). We use the power-distance kernels $k(x, y) = -\|x - y\|^\alpha$, whose smallest contraction factor across dimensions is γ^α (Proposition 4.1). The constraint of Appendix H.2 is then $\gamma^\alpha \leq \bar{\tau}$, with no estimation on a ball.

Figure 18 shows the results. For the mean shift, the power does not depend on α, so faster contraction costs no power in any dimension. For the checkerboard, the two distributions have the same mean. The kernels with larger α contract faster and are closer to comparing means, and they lose their power first. As the dimension grows, they lose it at smaller α, so the tradeoff becomes stronger. As the dimension grows, each distribution also spreads over 2^{d-1} corners, many more than the 100 points of a sample. By dimension 16, no kernel detects the difference, so the tradeoff can no longer be observed.

Dataset	**Distribution**
Cube Checkerboard	P and Q uniform on the corners $s/\sqrt{d}$, $s \in \{-1, 1\}^d$, of a cube, colored like a checkerboard, with the corners having an even number of -1 for P and those having an odd number for Q
Mean Shift	$P = \mathcal{N}(-0.4\,u, I_d)$ and $Q = \mathcal{N}(0.4\,u, I_d)$, with $u = (1, \ldots, 1)/\sqrt{d}$

Setting	**Value**	**Description**
Dimension d	$2, 3, 4, 8, 16, 32, 64$	Dimension of the samples.
Power α	$\mathrm{lin}(0.20, 2.00, 50)$	Parameter of $k_\alpha(x, y) = -\|x - y\|^\alpha$.
Sample sizes (n, m)	$(100, 100)$	Number of points in each sample.

Table 13: Distributions and settings of the experiments in higher dimensions. The outer repetitions, the permutation budget and the level are those of Table 10.

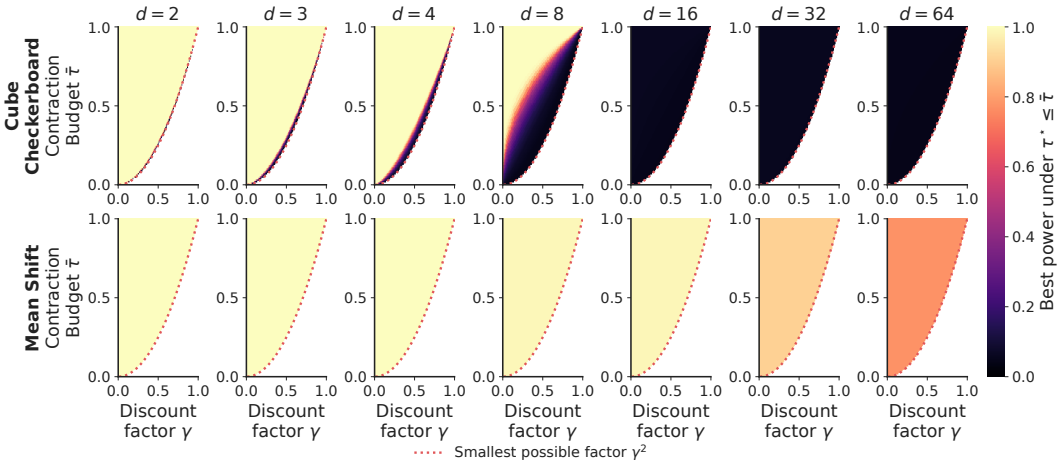

Figure 18: Tradeoff experiments in higher dimensions for the power-distance kernels $k(x, y) = -\|x - y\|^\alpha$, with the settings of Table 13. Each column is a dimension d. Each panel shows the best empirical test power achievable under the constraint $\gamma^\alpha \leq \bar{\tau}$. The dotted curve is the smallest possible contraction factor γ^2.

I Properties of radial kernels built from Bernstein functions

Preamble. This section studies when the radial kernels built from Bernstein functions induce a metric through $\mathbf{MMD}$, rather than only a pseudometric. For kernels of the form

$$k(x, y) = C - \phi(\|x - y\|^2),$$

the key issue is whether the Bernstein part $\phi(\|x - y\|^2)$ is *strictly* conditionally negative definite. Lemma I.2 shows that this happens exactly when the Lévy measure Λ is nonzero. Theorem I.1 then combines this with the general criterion for MMD to be a metric and shows that $\mathbf{MMD}_k$ is a metric if and only if $\Lambda \not\equiv 0$.

Lemma I.1 (Quadratic forms of squared Euclidean distances)**.** Let $x_1, \dots, x_n \in \mathbb{R}^d$ and $w_1, \dots, w_n \in \mathbb{R}$ satisfy $\sum_{i=1}^n w_i = 0$. Then

$$\sum_{i,j=1}^n w_i w_j \, \|x_i - x_j\|^2 = -2 \Big\| \sum_{i=1}^n w_i x_i \Big\|^2 \;\le\; 0.$$

Moreover, the left-hand side equals 0 if and only if $\sum_{i=1}^n w_i x_i = 0$.

Proof. Expand

$$\|x_i - x_j\|^2 = \|x_i\|^2 + \|x_j\|^2 - 2\langle x_i, x_j \rangle.$$

Hence

$$\sum_{i,j=1}^n w_i w_j \, \|x_i - x_j\|^2 = \sum_{i,j=1}^n w_i w_j \, \|x_i\|^2 + \sum_{i,j=1}^n w_i w_j \, \|x_j\|^2 - 2 \sum_{i,j=1}^n w_i w_j \, \langle x_i, x_j \rangle.$$

Using $\sum_{j=1}^n w_j = 0$,

$$\sum_{i,j=1}^n w_i w_j \, \|x_i\|^2 = \sum_{i=1}^n \Big(w_i \|x_i\|^2 \sum_{j=1}^n w_j \Big) = 0, \quad \sum_{i,j=1}^n w_i w_j \, \|x_j\|^2 = \sum_{j=1}^n \Big(w_j \|x_j\|^2 \sum_{i=1}^n w_i \Big) = 0.$$

Therefore

$$\sum_{i,j=1}^n w_i w_j \, \|x_i - x_j\|^2 = -2 \sum_{i,j=1}^n w_i w_j \, \langle x_i, x_j \rangle.$$

Finally,

$$\sum_{i,j=1}^n w_i w_j \, \langle x_i, x_j \rangle = \Big\langle \sum_{i=1}^n w_i x_i, \, \sum_{j=1}^n w_j x_j \Big\rangle = \Big\| \sum_{i=1}^n w_i x_i \Big\|^2,$$

which proves the identity and the inequality. The last claim follows since $\big\| \sum_{i=1}^n w_i x_i \big\|^2 = 0$ only when $\sum_{i=1}^n w_i x_i = 0$. □

Lemma I.2 (When a radial kernel built from a Bernstein function is CSND)**.** Let $\phi \colon [0, \infty) \to [0, \infty)$ be a Bernstein function with $\phi(0) = 0$, admitting the Lévy–Khintchine representation

$$\phi(t) \;=\; b\, t \;+\; \int_{(0,\infty)} \big(1 - e^{-t\, s}\big) \, \Lambda(ds),$$

where $b \ge 0$ and Λ is a nonnegative measure satisfying $\displaystyle\int_{(0,\infty)} \min\{1, s\} \, \Lambda(ds) < \infty$. Define the radial kernel

$$k(x, y) \;=\; \phi\big(\|x - y\|^2\big)$$

on $\mathbb{R}^d$. Then k is conditionally *strictly* negative definite (CSND) on $\mathbb{R}^d$ if and only if Λ is not identically zero (equivalently, ϕ is not of the form $\phi(t) = b\, t$). Moreover, if $\Lambda \equiv 0$, then k is CND but not CSND.

Proof. We split into two cases.

Case 1: Zero Lévy measure, $\Lambda \equiv 0$. Then $\phi(t) = b\,t$ and

$$k(x,y) = b\,\|x-y\|^2.$$

For any finite choice $x_1,\dots,x_n \in \mathbb{R}^d$ and weights $w_1,\dots,w_n$ with $\sum_i w_i = 0$,

$$\sum_{i,j=1}^{n} w_i w_j\, k(x_i,x_j) = b\sum_{i,j=1}^{n} w_i w_j\,\|x_i-x_j\|^2 = -2b\,\Big\|\sum_{i=1}^{n} w_i x_i\Big\|^2 \le 0$$

by Lemma I.1. Hence k is CND. Moreover, if $\sum_i w_i x_i = 0$ with $\sum_i w_i = 0$ and w nonzero, then the quadratic form vanishes, so k is not CSND.

Case 2: Nonzero Lévy measure, $\Lambda \not\equiv 0$. From the Lévy–Khintchine form,

$$k(x,y) = b\,\|x-y\|^2 + \int_{(0,\infty)} \big(1-e^{-s\|x-y\|^2}\big)\,\Lambda(ds).$$

Fix pairwise distinct $x_1,\dots,x_n \in \mathbb{R}^d$ and weights $w_1,\dots,w_n$ with $\sum_i w_i = 0$, not all zero, and set

$$S := \sum_{i,j=1}^{n} w_i w_j\, k(x_i,x_j).$$

Drift part. By Lemma I.1,

$$b\sum_{i,j} w_i w_j \|x_i-x_j\|^2 \;=\; -2b\Big\|\sum_{i=1}^{n} w_i x_i\Big\|^2 \;\le\; 0.$$

Interchanging the finite sum and the integral. For each (i,j) and all $s>0$, set $u := s\|x_i-x_j\|^2 \ge 0$. Since

$$1-e^{-u} \le 1 \qquad \text{and} \qquad 1-e^{-u} \le u,$$

we have

$$0 \le 1-e^{-s\|x_i-x_j\|^2} \le \min\{1,\; s\|x_i-x_j\|^2\}.$$

Moreover,

$$\min\{1,\, s\|x_i-x_j\|^2\} \le \max\{1,\|x_i-x_j\|^2\}\,\min\{1,s\},$$

so

$$\int_{(0,\infty)} \big|1-e^{-s\|x_i-x_j\|^2}\big|\,\Lambda(ds) \;\le\; \max\{1,\|x_i-x_j\|^2\}\int_{(0,\infty)} \min\{1,s\}\,\Lambda(ds) \;<\; \infty.$$

Hence each integral is finite, and since there are only finitely many pairs (i,j), we may exchange the finite sum and the integral,

$$\sum_{i,j} w_i w_j \int_{(0,\infty)} \big(1-e^{-s\|x_i-x_j\|^2}\big)\,\Lambda(ds) = \int_{(0,\infty)} \Big[\sum_{i,j} w_i w_j \big(1-e^{-s\|x_i-x_j\|^2}\big)\Big]\Lambda(ds).$$

Inner sum decomposition. For any fixed $s>0$,

$$\sum_{i,j} w_i w_j \big(1-e^{-s\|x_i-x_j\|^2}\big) = \underbrace{\sum_{i,j} w_i w_j}_{=0} - \sum_{i,j} w_i w_j\, e^{-s\|x_i-x_j\|^2} = -\,Q(s),$$

where

$$Q(s) := \sum_{i,j=1}^{n} w_i w_j\, e^{-s\|x_i-x_j\|^2}.$$

Strict positivity of $Q(s)$. Fix $s>0$. By [22, Cor. 2.1(c)], for pairwise distinct $x_1,\dots,x_n \in \mathbb{R}^d$,

$$\sum_{i,j=1}^{n} w_i w_j\, e^{-s\|x_i-x_j\|^2} \;>\; 0 \qquad \text{for every nonzero } (w_1,\dots,w_n) \in \mathbb{R}^n.$$

Therefore $Q(s) > 0$ for all $s > 0$ whenever the weights are not all zero, which is the case for the weights fixed above.

Conclusion. Combining the pieces,

$$S \;=\; -2b\Big\|\sum_i w_i x_i\Big\|^2 \;-\; \int_{(0,\infty)} Q(s)\,\Lambda(ds).$$

The first term is ≤ 0. Moreover, since $\Lambda \not\equiv 0$ and $Q(s) > 0$ for all $s > 0$, we have

$$\int_{(0,\infty)} Q(s)\,\Lambda(ds) > 0.$$

Hence $S < 0$, and therefore k is conditionally strictly negative definite.

□

Theorem I.1 (When radial kernels built from Bernstein functions make MMD a metric)**.** Let $d \ge 1$ and let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be symmetric. Assume that there exist a constant $C \in \mathbb{R}$ and a Bernstein function $\phi : [0,\infty) \to [0,\infty)$ with $\phi(0) = 0$ such that

$$k(x,y) \;=\; C - \phi\big(\|x-y\|^2\big).$$

By Theorem E.1, k is CPD, so $\mathbf{MMD}_k$ is defined as a quadratic form, as in Definition A.6. Write the Lévy–Khintchine representation of ϕ as

$$\phi(t) \;=\; b\,t \;+\; \int_{(0,\infty)} \big(1 - e^{-t\,s}\big)\,\Lambda(ds), \qquad b \ge 0,$$

where Λ is a nonnegative measure with $\int_{(0,\infty)} \min\{1,s\}\,\Lambda(ds) < \infty$.

Then $\mathbf{MMD}_k$ satisfies the metric axioms except possibly identity of indiscernibles (as in Proposition A.3). Moreover,

$$\boxed{\mathbf{MMD}_k \text{ is a metric if and only if } \Lambda \not\equiv 0.}$$

Proof.

Step 1: Reducing identity of indiscernibles to strictness. Since k is CPD, Proposition A.3 implies that $\mathbf{MMD}_k$ satisfies nonnegativity, symmetry, and the triangle inequality. It is a metric if and only if identity of indiscernibles holds, which by the same proposition is equivalent to k being CSPD.

Step 2: From CSPD of k to CSND of the Bernstein part. Adding a constant does not change whether a kernel is CPD or CSPD, hence

$$k \text{ is CSPD} \quad\Longleftrightarrow\quad -\phi(\|x-y\|^2) \text{ is CSPD} \quad\Longleftrightarrow\quad \phi(\|x-y\|^2) \text{ is CSND.}$$

Step 3: When the Bernstein part is CSND. Lemma I.2 proves this criterion on finite sets of points. For the general case required by Definition A.5, Theorem 4.1 of [17] shows that $\phi(\|x-y\|^2)$ is CSND when $\Lambda \not\equiv 0$, and Lemma I.2 already shows that it is not when $\Lambda \equiv 0$. Hence

$$\phi(\|x-y\|^2) \text{ is CSND on } \mathbb{R}^d \quad\Longleftrightarrow\quad \Lambda \not\equiv 0.$$

Combining Steps 1–3 yields

$$\mathbf{MMD}_k \text{ is a metric} \quad\Longleftrightarrow\quad \Lambda \not\equiv 0,$$

which proves the boxed statement. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The theoretical claims regarding reduction to CPD then Bernstein criteria are the core of the work. The empirical claims regarding contraction on bounded sets, new kernels found by numerical search and the tension between contraction rate and power are experimented and documented.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion discusses further directions, such as experiments on RL and benchmarking on real environments. Section 5 states that the results on bounded sets are numerical evidence on finite sets of points, not a proof.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All the proofs are in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The appendix provides detailed description of the algorithms and numerical values used for the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code will be released upon acceptance. The appendix contains all the information to reconstruct it.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The numerical experiments are well documented with tables of values we used and pseudo-code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Most experiments report deterministic numerical diagnostics and therefore do not require error bars. The detectability experiments estimate power over many independent draws, so the reported trends are not based on isolated random runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments are lightweight numerical computations, and the appendix gives enough compute information to reproduce them.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work does not involve human subjects, real-person datasets, private data, deployed systems, or high-risk model releases. The experiments are synthetic and the paper provides the information needed for reproducibility.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: We do not identify a direct societal impact beyond general downstream uses of reinforcement learning methods.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing datasets, models, or third-party research assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development does not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.