Injective Hilbert Space Embeddings of Probability Measures

Bharath K. Sriperumbudur

University of California, San Diego
&
MPI for Biological Cybernetics, Tübingen

Joint work with:
Arthur Gretton, Bernhard Schölkopf (MPI, Tübingen)
Kenji Fukumizu (Institute for Statistical Mathematics, Tokyo)
Gert Lanckriet (University of California, San Diego)

Probability Metrics

Setup:

- M : measurable space.
- $\mathcal{P}$: set of all Borel probability measures defined on M.

To do:

- Define a metric, γ on $\mathcal{P}$.
- γ is called the probability metric.

Popular examples:

- Kullback-Leibler divergence
- Jensen-Shannon divergence
- Total-variation distance (metric)
- Hellinger distance
- χ^2-distance

The above examples are special instances of Csiszár's f-divergence.

Applications

Two-sample problem:

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, respectively.
- Determine: are $\mathbb{P}$ and $\mathbb{Q}$ different?

- $\gamma(\mathbb{P}, \mathbb{Q})$: distance metric between $\mathbb{P}$ and $\mathbb{Q}$.

$$\begin{array}{c} H_0 : \mathbb{P} = \mathbb{Q} \\ \\ H_1 : \mathbb{P} \neq \mathbb{Q} \end{array} \equiv \begin{array}{c} H_0 : \gamma(\mathbb{P}, \mathbb{Q}) = 0 \\ \\ H_1 : \gamma(\mathbb{P}, \mathbb{Q}) > 0 \end{array}$$

- Test statistic: $\gamma(.,.)$

Other applications: Hypothesis testing (independence tests, goodness-of-fit tests), Central limit theorems, Density estimation, Markov chain Monte Carlo etc.

Applications

Two-sample problem:

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, respectively.
- Determine: are $\mathbb{P}$ and $\mathbb{Q}$ different?

- $\gamma(\mathbb{P}, \mathbb{Q})$: distance metric between $\mathbb{P}$ and $\mathbb{Q}$.

$$\begin{array}{c} H_0 : \mathbb{P} = \mathbb{Q} \\ H_1 : \mathbb{P} \neq \mathbb{Q} \end{array} \equiv \begin{array}{c} H_0 : \gamma(\mathbb{P}, \mathbb{Q}) = 0 \\ H_1 : \gamma(\mathbb{P}, \mathbb{Q}) > 0 \end{array}$$

- Test statistic: $\gamma(.,.)$

Other applications: Hypothesis testing (independence tests, goodness-of-fit tests), Central limit theorems, Density estimation, Markov chain Monte Carlo etc.

Applications

Two-sample problem:

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, respectively.
- Determine: are $\mathbb{P}$ and $\mathbb{Q}$ different?

- $\gamma(\mathbb{P}, \mathbb{Q})$: distance metric between $\mathbb{P}$ and $\mathbb{Q}$.

$$\begin{array}{l} H_0 : \mathbb{P} = \mathbb{Q} \\ H_1 : \mathbb{P} \neq \mathbb{Q} \end{array} \equiv \begin{array}{l} H_0 : \gamma(\mathbb{P}, \mathbb{Q}) = 0 \\ H_1 : \gamma(\mathbb{P}, \mathbb{Q}) > 0 \end{array}$$

- Test statistic: $\gamma(.,.)$

Other applications: Hypothesis testing (independence tests, goodness-of-fit tests), Central limit theorems, Density estimation, Markov chain Monte Carlo etc.

Maximum Mean Discrepancy

Let (M, ρ) be a metric space. The maximum mean discrepancy (MMD) between $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$ is defined as

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int_M f \, d\mathbb{P} - \int_M f \, d\mathbb{Q} \right|, \tag{1}$$

where $\mathcal{F} = \{f : M \to \mathbb{R} | f \in \cap_{\mathbb{P} \in \mathcal{P}} L^1(M, \mathbb{P})\}$.

- $\gamma_{\mathcal{F}}$ is also called the integral probability metric [Müller, 1997].
- Motivated from the notion of weak convergence of probability measures on metric spaces.
- $\gamma_{\mathcal{F}}$ is a pseudo-metric on $\mathcal{P}$, i.e., $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0 \not\Rightarrow \mathbb{P} = \mathbb{Q}$. $\mathcal{F}$ determines the metric property of $\gamma_{\mathcal{F}}$.

Maximum Mean Discrepancy

Let (M, ρ) be a metric space. The maximum mean discrepancy (MMD) between $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$ is defined as

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int_M f \, d\mathbb{P} - \int_M f \, d\mathbb{Q} \right|, \tag{1}$$

where $\mathcal{F} = \{f : M \to \mathbb{R} | f \in \cap_{\mathbb{P} \in \mathcal{P}} L^1(M, \mathbb{P})\}$.

- $\gamma_{\mathcal{F}}$ is also called the integral probability metric [Müller, 1997].
- Motivated from the notion of weak convergence of probability measures on metric spaces.
- $\gamma_{\mathcal{F}}$ is a pseudo-metric on $\mathcal{P}$, i.e., $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0 \nRightarrow \mathbb{P} = \mathbb{Q}$. $\mathcal{F}$ determines the metric property of $\gamma_{\mathcal{F}}$.

Examples

$\gamma_{\mathcal{F}}$ is a metric on $\mathcal{P}$ for

- $\mathcal{F} = C_b(M)$: definition of weak convergence.
- $\mathcal{F} = C_{bu}(M)$: by the Portmanteau theorem.
- $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$: total variation distance.
- $\mathcal{F} = \{f : \|f\|_L \leq 1\}$: Monge-Wasserstein/Rubinstein-Kantorovich metric.
- $\mathcal{F} = \{f : \|f\|_\infty + \|f\|_L \leq 1\}$: Dudley metric.
- $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\}$: Kolmogorov distance.
- $\mathcal{F} = \{e^{i\langle\omega,.\rangle} : \omega \in \mathbb{R}^d\}$: maximal difference between the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$.

What if $\mathcal{F}$ is an RKHS?

Set up: [Gretton et al., 2007]

- $\mathcal{H}$: reproducing kernel Hilbert space (RKHS).
- k : reproducing kernel; $k : M \times M \to \mathbb{R}$.
- $\mathcal{F}$: a unit ball in $\mathcal{H}$, i.e., $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$.

Theorem

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\} \subset (\mathcal{H}, k)$ *defined on a measurable space* M.
- k *is measurable and bounded.*

Then

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int_M f \, d\mathbb{P} - \int_M f \, d\mathbb{Q} \right| = \left\| \int_M k \, d\mathbb{P} - \int_M k \, d\mathbb{Q} \right\|_{\mathcal{H}}, \quad (2)$$

where $\|.\|_{\mathcal{H}}$ *represents the RKHS norm.*

What if $\mathcal{F}$ is an RKHS?

Set up: [Gretton et al., 2007]

- $\mathcal{H}$: reproducing kernel Hilbert space (RKHS).
- k : reproducing kernel; $k : M \times M \to \mathbb{R}$.
- $\mathcal{F}$: a unit ball in $\mathcal{H}$, i.e., $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\}$.

Theorem

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\} \subset (\mathcal{H}, k)$ *defined on a measurable space* M*.*
- *k is measurable and bounded.*

Then

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int_M f \, d\mathbb{P} - \int_M f \, d\mathbb{Q} \right| = \left\| \int_M k \, d\mathbb{P} - \int_M k \, d\mathbb{Q} \right\|_{\mathcal{H}}, \qquad (2)$$

where $\|.\|_{\mathcal{H}}$ *represents the RKHS norm.*

Why RKHS?

- Given $\mathbb{P}$ and $\mathbb{Q}$, computing $\gamma(\mathbb{P}, \mathbb{Q})$ is not straightforward when $\mathcal{F} = C_b(M)$, $C_{bu}(M)$, $\{\|f\|_L \leq 1\}$, $\{\|f\|_L + \|f\|_\infty \leq 1\}$.
- When $\mathcal{F} = \{f : \|f\|_\mathcal{H} \leq 1\}$, then $\gamma(\mathbb{P}, \mathbb{Q})$ is entirely determined by the kernel, k.
- k is measurable and bounded: $\gamma(\hat{\mathbb{P}}, \hat{\mathbb{Q}})$ is a $\sqrt{mn/(m+n)}$-consistent estimator of $\gamma(\mathbb{P}, \mathbb{Q})$ [Gretton et al., 2007].
- $M = \mathbb{R}^d$ and k is translation-invariant: the rate is independent of d.
- Easy to handle structured domains like graphs and strings.

Why RKHS?

- Given $\mathbb{P}$ and $\mathbb{Q}$, computing $\gamma(\mathbb{P}, \mathbb{Q})$ is not straightforward when $\mathcal{F} = C_b(M)$, $C_{bu}(M)$, $\{\|f\|_L \leq 1\}$, $\{\|f\|_L + \|f\|_\infty \leq 1\}$.
- When $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$, then $\gamma(\mathbb{P}, \mathbb{Q})$ is entirely determined by the kernel, k.
- k is measurable and bounded: $\gamma(\hat{\mathbb{P}}, \hat{\mathbb{Q}})$ is a $\sqrt{mn/(m+n)}$-consistent estimator of $\gamma(\mathbb{P}, \mathbb{Q})$ [Gretton et al., 2007].
- $M = \mathbb{R}^d$ and k is translation-invariant: the rate is independent of d.
- Easy to handle structured domains like graphs and strings.

Why RKHS?

- Given $\mathbb{P}$ and $\mathbb{Q}$, computing $\gamma(\mathbb{P}, \mathbb{Q})$ is not straightforward when $\mathcal{F} = C_b(M)$, $C_{bu}(M)$, $\{\|f\|_L \leq 1\}$, $\{\|f\|_L + \|f\|_\infty \leq 1\}$.
- When $\mathcal{F} = \{f : \|f\|_\mathcal{H} \leq 1\}$, then $\gamma(\mathbb{P}, \mathbb{Q})$ is entirely determined by the kernel, k.
- k is measurable and bounded: $\gamma(\hat{\mathbb{P}}, \hat{\mathbb{Q}})$ is a $\sqrt{mn/(m+n)}$-consistent estimator of $\gamma(\mathbb{P}, \mathbb{Q})$ [Gretton et al., 2007].
- $M = \mathbb{R}^d$ and k is translation-invariant: the rate is independent of d.
- Easy to handle structured domains like graphs and strings.

RKHS Embedding

- $\mathbb{P} \in \mathcal{P}$ is embedded as $\int_M k\, d\mathbb{P} \in \mathcal{H}$,

$$\Pi : \mathcal{P} \to \mathcal{H}, \ \ \mathbb{P} \mapsto \int_M k\, d\mathbb{P}. \tag{3}$$

- Example: $\mathbb{P} = \delta_x$ (Dirac measure at $x \in \mathbb{M}$) $\mapsto k(.,x)$ (kernel function at x).

Question: When is Π injective? In other words, when is $\gamma_{\mathcal{F}}$ a metric?

For what k, $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$?

- By choosing the right RKHS, $\mathbb{P}$ and $\mathbb{Q}$ can be distinguished by their mean elements in $\mathcal{H}$.

RKHS Embedding

- $\mathbb{P} \in \mathcal{P}$ is embedded as $\int_M k \, d\mathbb{P} \in \mathcal{H}$,

$$\Pi : \mathcal{P} \to \mathcal{H}, \;\; \mathbb{P} \mapsto \int_M k \, d\mathbb{P}. \tag{3}$$

- Example: $\mathbb{P} = \delta_x$ (Dirac measure at $x \in \mathbb{M}$) $\mapsto k(., x)$ (kernel function at x).

Question: When is Π injective? In other words, when is $\gamma_{\mathcal{F}}$ a metric?

For what k, $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$?

- By choosing the right RKHS, $\mathbb{P}$ and $\mathbb{Q}$ can be distinguished by their mean elements in $\mathcal{H}$.

RKHS Embedding

- $\mathbb{P} \in \mathcal{P}$ is embedded as $\int_M k\, d\mathbb{P} \in \mathcal{H}$,

$$\Pi : \mathcal{P} \to \mathcal{H}, \;\; \mathbb{P} \mapsto \int_M k\, d\mathbb{P}. \qquad (3)$$

- Example: $\mathbb{P} = \delta_x$ (Dirac measure at $x \in \mathbb{M}$) $\mapsto k(.,x)$ (kernel function at x).

Question: When is Π injective? In other words, when is $\gamma_{\mathcal{F}}$ a metric?

For what k, $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$?

- By choosing the right RKHS, $\mathbb{P}$ and $\mathbb{Q}$ can be distinguished by their mean elements in $\mathcal{H}$.

Characteristic Kernel

Definition

k is characteristic to a set $\mathcal{D} \subset \mathcal{P}$ of probability measures defined on M if

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q} \text{ for } \mathbb{P}, \mathbb{Q} \in \mathcal{D} \tag{4}$$

.

Example

Let $M = \mathbb{R}^d$ and $k(\omega, x) = e^{i\omega^T x}$.

$$\Pi[\mathbb{P}] = \int_M k \, d\mathbb{P} = \int_{\mathbb{R}^d} e^{i\langle \cdot, x \rangle} \, d\mathbb{P}. \tag{5}$$

The notion of characteristic kernel is a generalization of the characteristic function.

Characteristic Kernel

Definition

k is characteristic to a set $\mathcal{D} \subset \mathcal{P}$ of probability measures defined on M if

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q} \text{ for } \mathbb{P}, \mathbb{Q} \in \mathcal{D} \tag{4}$$

.

Example

Let $M = \mathbb{R}^d$ and $k(\omega, x) = e^{i\omega^T x}$.

$$\Pi[\mathbb{P}] = \int_M k\, d\mathbb{P} = \int_{\mathbb{R}^d} e^{i\langle \cdot, x\rangle}\, d\mathbb{P}. \tag{5}$$

The notion of characteristic kernel is a generalization of the characteristic function.

Sufficient Conditions

- Let M be compact. If $\mathcal{H}$ is dense in $C_b(M)$ w.r.t. the L^∞ norm (i.e. k is universal [Steinwart, 2002]), then k is characteristic to $\mathcal{P}$. [Gretton et al., 2007].
  - Gaussian and Laplacian kernels on any compact subset of $\mathbb{R}^d$.
- If $\mathcal{H} + \mathbb{R}$ is dense in $L^q(M)$, $q \geq 1$, then k is characteristic to $\mathcal{P}$ [Fukumizu et al., 2008].
  - More general condition than universality.
  - Gaussian and Laplacian kernels on the entire $\mathbb{R}^d$.

Issues:

- Difficult to check the conditions.
- Universality is an overly restrictive assumption.

Sufficient Conditions

- Let M be compact. If $\mathcal{H}$ is dense in $C_b(M)$ w.r.t. the L^∞ norm (i.e. k is universal [Steinwart, 2002]), then k is characteristic to $\mathcal{P}$. [Gretton et al., 2007].
  - Gaussian and Laplacian kernels on any compact subset of $\mathbb{R}^d$.
- If $\mathcal{H} + \mathbb{R}$ is dense in $L^q(M)$, $q \geq 1$, then k is characteristic to $\mathcal{P}$ [Fukumizu et al., 2008].
  - More general condition than universality.
  - Gaussian and Laplacian kernels on the entire $\mathbb{R}^d$.

Issues:

- Difficult to check the conditions.
- Universality is an overly restrictive assumption.

Sufficient Conditions

- Let M be compact. If $\mathcal{H}$ is dense in $C_b(M)$ w.r.t. the L^∞ norm (i.e. k is universal [Steinwart, 2002]), then k is characteristic to $\mathcal{P}$. [Gretton et al., 2007].
  - Gaussian and Laplacian kernels on any compact subset of $\mathbb{R}^d$.
- If $\mathcal{H} + \mathbb{R}$ is dense in $L^q(M)$, $q \geq 1$, then k is characteristic to $\mathcal{P}$ [Fukumizu et al., 2008].
  - More general condition than universality.
  - Gaussian and Laplacian kernels on the entire $\mathbb{R}^d$.

Issues:

- Difficult to check the conditions.
- Universality is an overly restrictive assumption.

Background & Notation

Assumption

$M = \mathbb{R}^d$. $k(x,y) = \psi(x-y)$ *where* ψ *is a bounded continuous real-valued positive definite function on* $\mathbb{R}^d$.

Theorem (Bochner)

ψ *is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure,* Λ *on* $\mathbb{R}^d$*, i.e.,*

$$\psi(x) = \int_{\mathbb{R}^d} e^{-ix^T\omega}\, d\Lambda(\omega),\ \forall x \in \mathbb{R}^d. \tag{6}$$

Characteristic function: $\phi_{\mathbb{P}}(\omega) = \int_{\mathbb{R}^d} e^{i\omega^T x}\, d\mathbb{P}(x),\ \forall \omega \in \mathbb{R}^d$.

- If $\psi \in L^1(\mathbb{R}^d)$, then $d\Lambda = \frac{1}{(2\pi)^{d/2}} \Psi\, d\omega$.

Background & Notation

Assumption

$M = \mathbb{R}^d$. $k(x, y) = \psi(x - y)$ *where* ψ *is a bounded continuous real-valued positive definite function on* $\mathbb{R}^d$.

Theorem (Bochner)

ψ *is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure,* Λ *on* $\mathbb{R}^d$*, i.e.,*

$$\psi(x) = \int_{\mathbb{R}^d} e^{-ix^T\omega}\, d\Lambda(\omega),\ \ \forall\, x \in \mathbb{R}^d. \tag{6}$$

Characteristic function: $\phi_{\mathbb{P}}(\omega) = \int_{\mathbb{R}^d} e^{i\omega^T x}\, d\mathbb{P}(x),\ \forall\, \omega \in \mathbb{R}^d$.

- If $\psi \in L^1(\mathbb{R}^d)$, then $d\Lambda = \frac{1}{(2\pi)^{d/2}}\Psi\, d\omega$.

Background & Notation

Assumption

$M = \mathbb{R}^d$. $k(x,y) = \psi(x-y)$ where ψ is a bounded continuous real-valued positive definite function on $\mathbb{R}^d$.

Theorem (Bochner)

ψ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure, Λ on $\mathbb{R}^d$, i.e.,

$$\psi(x) = \int_{\mathbb{R}^d} e^{-ix^T\omega}\, d\Lambda(\omega),\ \ \forall\, x \in \mathbb{R}^d. \tag{6}$$

Characteristic function: $\phi_{\mathbb{P}}(\omega) = \int_{\mathbb{R}^d} e^{i\omega^T x}\, d\mathbb{P}(x),\ \forall\, \omega \in \mathbb{R}^d$.

- If $\psi \in L^1(\mathbb{R}^d)$, then $d\Lambda = \frac{1}{(2\pi)^{d/2}}\Psi\, d\omega$.

Main Result

Theorem

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\} \subset (\mathcal{H}, k)$.
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*

Then, k is characteristic to $\mathcal{P}$ $\Leftrightarrow$ supp$(\Lambda) = \mathbb{R}^d$.

- If k is such that supp$(\Lambda) = \mathbb{R}^d$, then $\nexists\, \mathbb{P} \neq \mathbb{Q}$ such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0$.
- Can we have k with supp$(\Lambda) \neq \mathbb{R}^d$ such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$? The theorem says NO.
- Complete characterization of translation-invariant kernels in $\mathbb{R}^d$.
- Examples: Gaussian, Laplacian, B_{2n+1}-splines, Matérn class etc.

Main Result

Theorem

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\} \subset (\mathcal{H}, k)$.
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*

Then, k is characteristic to $\mathcal{P}$ $\Leftrightarrow$ supp$(\Lambda) = \mathbb{R}^d$.

- If k is such that supp$(\Lambda) = \mathbb{R}^d$, then $\nexists\, \mathbb{P} \neq \mathbb{Q}$ such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0$.
- Can we have k with supp$(\Lambda) \neq \mathbb{R}^d$ such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$? The theorem says NO.
- Complete characterization of translation-invariant kernels in $\mathbb{R}^d$.
- Examples: Gaussian, Laplacian, B_{2n+1}-splines, Matérn class etc.

Main Result

Theorem

Let

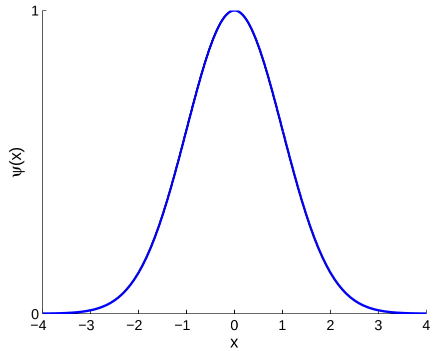
- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\} \subset (\mathcal{H}, k)$.
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*

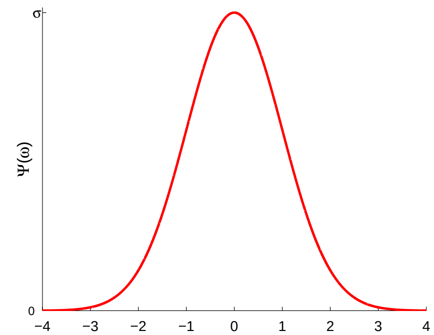
Then, k is characteristic to $\mathcal{P} \Leftrightarrow \text{supp}(\Lambda) = \mathbb{R}^d$.

- If k is such that $\text{supp}(\Lambda) = \mathbb{R}^d$, then $\nexists\, \mathbb{P} \neq \mathbb{Q}$ such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0$.
- Can we have k with $\text{supp}(\Lambda) \neq \mathbb{R}^d$ such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$? The theorem says NO.
- Complete characterization of translation-invariant kernels in $\mathbb{R}^d$.
- Examples: Gaussian, Laplacian, B_{2n+1}-splines, Matérn class etc.

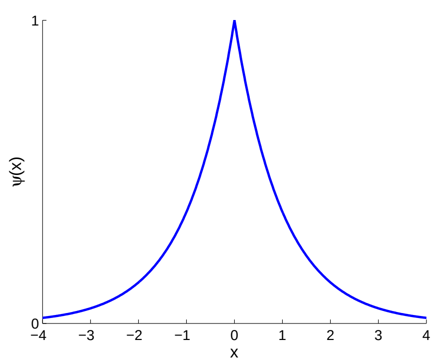
Characteristic kernel: Examples

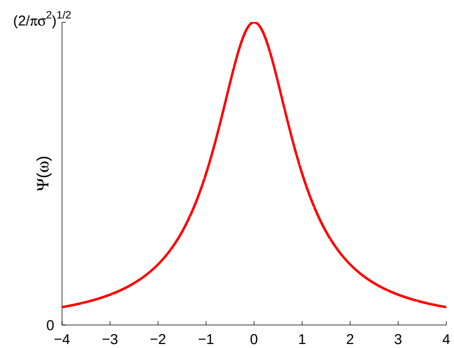
- Gaussian kernel: $\psi(x) = e^{-x^2/2\sigma^2}$; $\Psi(\omega) = \sigma e^{-\sigma^2\omega^2/2}$.





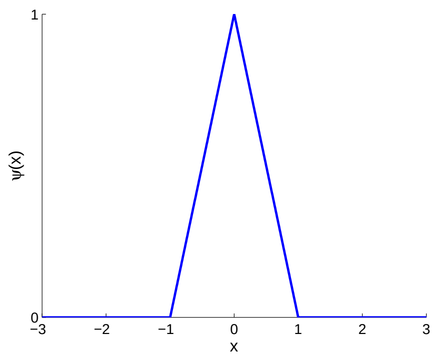
- Laplacian kernel: $\psi(x) = e^{-\sigma|x|}$; $\Psi(\omega) = \sqrt{\frac{2}{\pi}} \frac{\sigma}{\sigma^2+\omega^2}$.

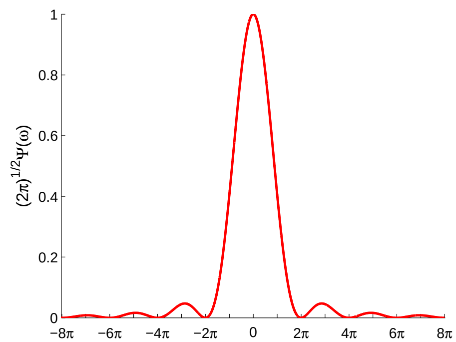




Characteristic kernel: Examples

- B_1-spline kernel: $\psi(x) = (1 - |x|)\mathbb{1}_{[-1,1]}(x)$; $\Psi(\omega) = \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{\sin^2(\frac{\omega}{2})}{\omega^2}$.

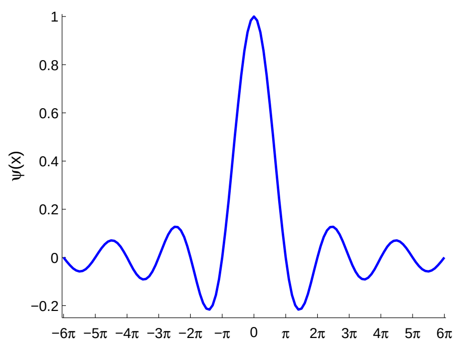


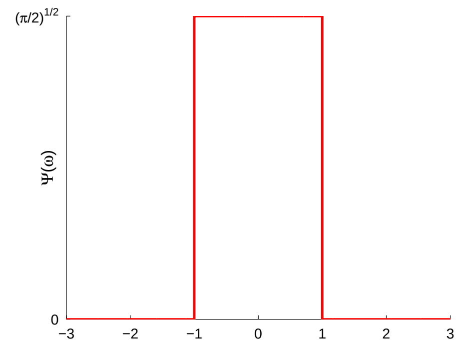


- $\Psi(\omega) = 0$ at $\omega = 2l\pi$, $l \in \mathbb{Z}$; $\text{supp}(\Psi) = \mathbb{R}$.

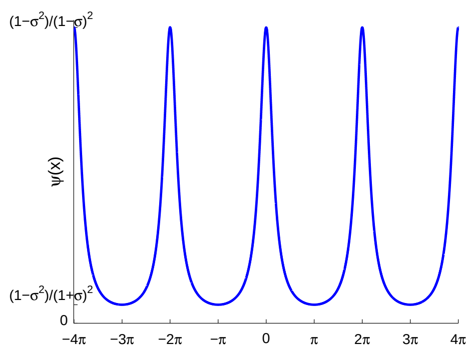
Non-characteristic kernel: Examples

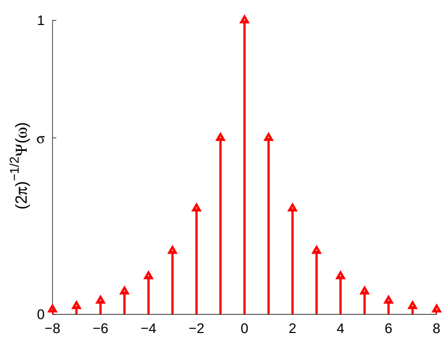
- Sinc kernel: $\psi(x) = \frac{\sin(\sigma x)}{x}$; $\Psi(\omega) = \sqrt{\frac{\pi}{2}}\mathbb{1}_{[-\sigma,\sigma]}(\omega)$.





- Poisson kernel: $\psi(x) = \frac{1-\sigma^2}{\sigma^2 - 2\sigma\cos(x) + 1}$; $\Psi(\omega) = \sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{|j|}\delta(\omega - j)$.

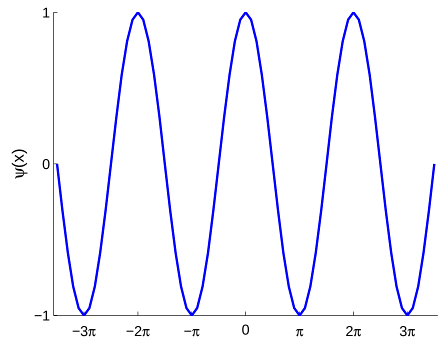


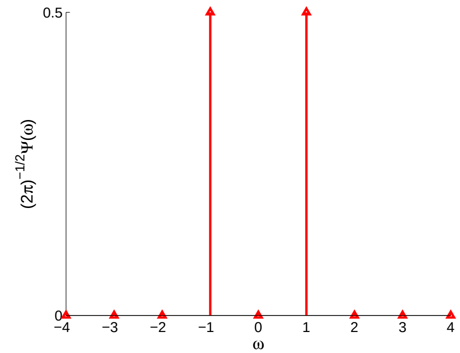


- Periodic kernels on $\mathbb{R}^d$ are not characteristic to $\mathcal{P}$.

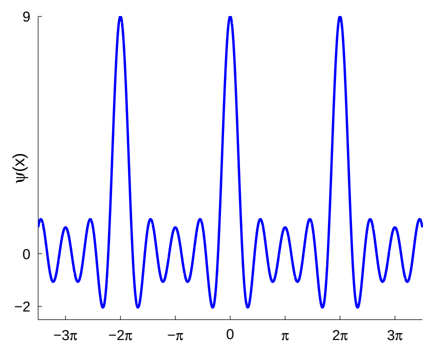
Non-characteristic kernel: Examples

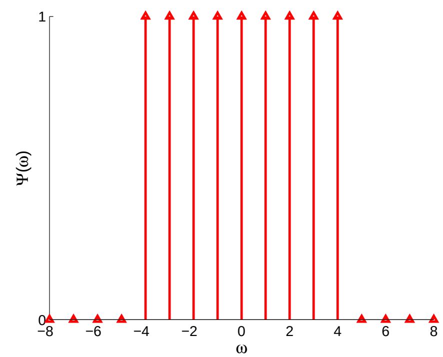
- Cosine kernel: $\psi(x) = \cos(\sigma x)$; $\Psi(\omega) = \sqrt{\frac{\pi}{2}}[\delta(\omega - \sigma) + \delta(\omega + \sigma)]$.





- Dirichlet kernel: $\psi(x) = \frac{\sin(nx+0.5x)}{\sin(0.5x)}$; $\Psi(\omega) = \sqrt{2\pi}\sum_{j=-n}^{n}\delta(\omega - j)$.





Fourier Representation of MMD

Lemma

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\} \subset (\mathcal{H}, k)$.
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*
- $\phi_{\mathbb{P}}, \phi_{\mathbb{Q}}$: *characteristic functions of* $\mathbb{P}$ *and* $\mathbb{Q}$.

Fourier Representation of MMD

Lemma

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\} \subset (\mathcal{H}, k)$.
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*
- $\phi_{\mathbb{P}}, \phi_{\mathbb{Q}}$: *characteristic functions of* $\mathbb{P}$ *and* $\mathbb{Q}$.

Then

$$\int_{\mathbb{R}^d} k(., x)\, d\mathbb{P}(x) = \mathscr{F}^{-1}\left[\overline{\phi_{\mathbb{P}}}\Lambda\right], \tag{7}$$

Fourier Representation of MMD

Lemma

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\} \subset (\mathcal{H}, k)$.
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*
- $\phi_{\mathbb{P}}$, $\phi_{\mathbb{Q}}$ *: characteristic functions of* $\mathbb{P}$ *and* $\mathbb{Q}$.

Then

$$\int_{\mathbb{R}^d} k(., x)\, d\mathbb{P}(x) = \mathscr{F}^{-1}\left[\overline{\phi}_{\mathbb{P}}\Lambda\right], \tag{7}$$

and

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \left\|\mathscr{F}^{-1}[(\overline{\phi}_{\mathbb{P}} - \overline{\phi}_{\mathbb{Q}})\Lambda]\right\|_{\mathcal{H}}, \tag{8}$$

where $-$ *represents complex conjugation,* $\mathscr{F}^{-1}$ *represents the inverse Fourier transform.*

Proof

Sufficiency: Assume $\psi \in L^1(\mathbb{R}^d)$.

- Λ is absolutely continuous w.r.t. the Lebesgue measure and has density, Ψ.
- $\mathscr{F}[\psi] = \Psi$
- $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow (\phi_{\mathbb{P}} - \phi_{\mathbb{Q}})\Psi = 0$.
- If $\text{supp}(\Lambda) = \mathbb{R}^d$, then $\Psi(\omega) > 0$ a.e. $\Rightarrow \phi_{\mathbb{P}} = \phi_{\mathbb{Q}}$ a.e. $\Rightarrow \mathbb{P} = \mathbb{Q}$.

$\psi \notin L^1(\mathbb{R}^d)$ can be addressed using distribution theory.

Necessity:

- We need to show that k is characteristic $\Rightarrow$ $\text{supp}(\Lambda) = \mathbb{R}^d$.
- Equivalent to showing that $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d \Rightarrow k$ is not characteristic.
- We show that for any k with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$, $\exists \mathbb{P} \neq \mathbb{Q}$ such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0$.

Proof

Sufficiency: Assume $\psi \in L^1(\mathbb{R}^d)$.

- Λ is absolutely continuous w.r.t. the Lebesgue measure and has density, Ψ.
- $\mathscr{F}[\psi] = \Psi$
- $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = 0 \Rightarrow (\phi_{\mathbb{P}} - \phi_{\mathbb{Q}})\Psi = 0.$
- If $\text{supp}(\Lambda) = \mathbb{R}^d$, then $\Psi(\omega) > 0$ a.e. $\Rightarrow \phi_{\mathbb{P}} = \phi_{\mathbb{Q}}$ a.e. $\Rightarrow \mathbb{P} = \mathbb{Q}$.

$\psi \notin L^1(\mathbb{R}^d)$ can be addressed using distribution theory.

Necessity:

- We need to show that k is characteristic $\Rightarrow \text{supp}(\Lambda) = \mathbb{R}^d$.
- Equivalent to showing that $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d \Rightarrow k$ is not characteristic.
- We show that for any k with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$, $\exists\, \mathbb{P} \neq \mathbb{Q}$ such that $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = 0$.

Proof Idea: Necessity

Lemma

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\} \subset (\mathcal{H}, k)$.
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*
- $\mathcal{D} = \{\mathbb{P} : \phi_{\mathbb{P}} \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)\} \subset \mathcal{P}$.

Then for any $\mathbb{Q} \in \mathcal{D}$, $\exists \mathbb{P} \neq \mathbb{Q}$, $\mathbb{P} \in \mathcal{D}$ *given by*

$$p = q + \mathscr{F}^{-1}[\theta] \quad (9)$$

such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0$ *if and only if* $\exists\, \theta : \mathbb{R}^d \to \mathbb{C}$, $\theta \neq 0$ *that satisfies:*

(i) $\theta \in (L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)) \cap C_b(\mathbb{R}^d)$ *is conjugate symmetric,*

(ii) $\mathscr{F}^{-1}[\theta] \in L^1(\mathbb{R}^d) \cap (L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d))$,

(iii) $\theta\Lambda = 0$,

(iv) $\theta(0) = 0$,

(v) $\inf_{x \in \mathbb{R}^d}\{\mathscr{F}^{-1}[\theta](x) + q(x)\} \geq 0$.

Proof Idea: Necessity

Lemma

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\} \subset (\mathcal{H}, k)$.
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*
- $\mathcal{D} = \{\mathbb{P} : \phi_{\mathbb{P}} \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)\} \subset \mathcal{P}$.

Then for any $\mathbb{Q} \in \mathcal{D}$, $\exists \mathbb{P} \neq \mathbb{Q}$, $\mathbb{P} \in \mathcal{D}$ *given by*

$$p = q + \mathscr{F}^{-1}[\theta] \tag{9}$$

such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0$ *if and only if* $\exists\, \theta : \mathbb{R}^d \to \mathbb{C}$, $\theta \neq 0$ *that satisfies:*

(i) $\theta \in (L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)) \cap C_b(\mathbb{R}^d)$ *is conjugate symmetric,*

(ii) $\mathscr{F}^{-1}[\theta] \in L^1(\mathbb{R}^d) \cap (L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d))$,

(iii) $\theta\Lambda = 0$,

(iv) $\theta(0) = 0$,

(v) $\inf_{x \in \mathbb{R}^d}\{\mathscr{F}^{-1}[\theta](x) + q(x)\} \geq 0$.

Proof Idea: Necessity

Lemma

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\} \subset (\mathcal{H}, k)$.
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*
- $\mathcal{D} = \{\mathbb{P} : \phi_{\mathbb{P}} \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)\} \subset \mathcal{P}$.

Then for any $\mathbb{Q} \in \mathcal{D}$, $\exists \mathbb{P} \neq \mathbb{Q}$, $\mathbb{P} \in \mathcal{D}$ *given by*

$$p = q + \mathscr{F}^{-1}[\theta] \tag{9}$$

such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = 0$ *if and only if* $\exists\, \theta : \mathbb{R}^d \to \mathbb{C}$, $\theta \neq 0$ *that satisfies:*

(i) $\theta \in (L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)) \cap C_b(\mathbb{R}^d)$ *is conjugate symmetric,*

(ii) $\mathscr{F}^{-1}[\theta] \in L^1(\mathbb{R}^d) \cap (L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d))$,

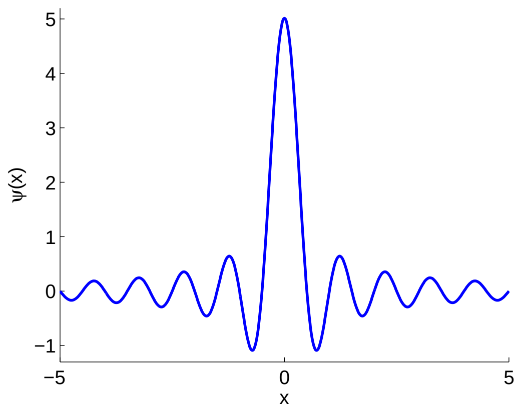
(iii) $\theta\Lambda = 0$,

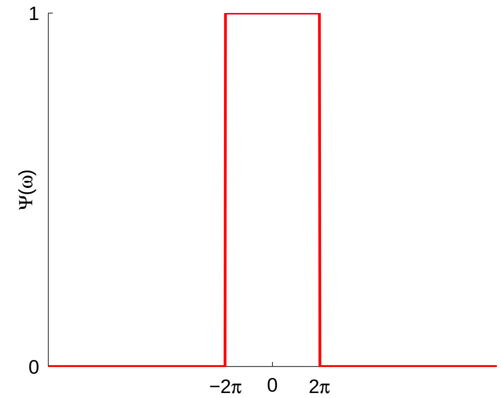
(iv) $\theta(0) = 0$,

(v) $\inf_{x \in \mathbb{R}^d} \{\mathscr{F}^{-1}[\theta](x) + q(x)\} \ge 0$.

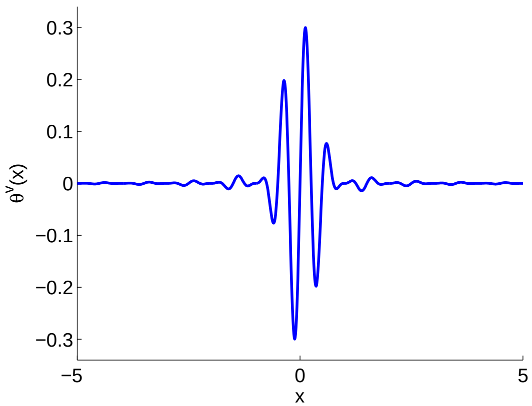
Proof Idea of Necessity: Example

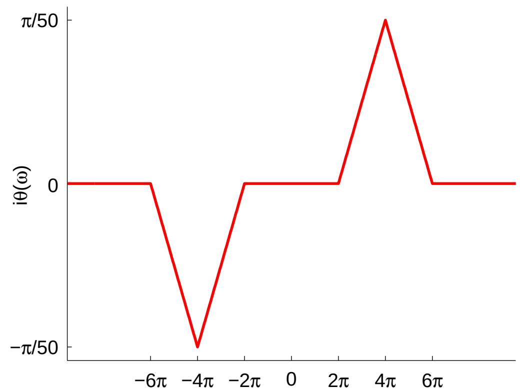
- $\psi(x) = \sqrt{\frac{2}{\pi}} \frac{sin(2\pi x)}{x}$; $\Psi(\omega) = \mathbf{1}_{[-2\pi,2\pi]}(\omega)$.





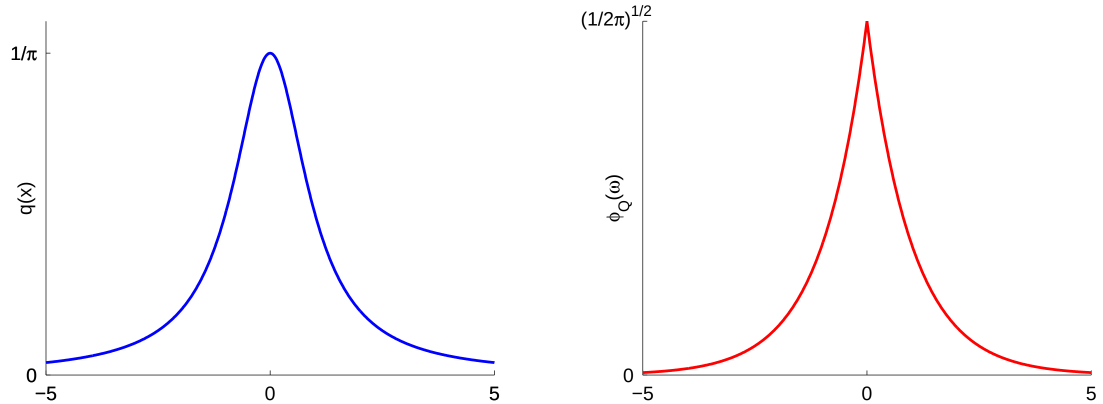
- $\theta(\omega) = \frac{1}{100i}\left[\mathbb{1}_{[-2\pi,2\pi]}(\omega)(2\pi - |\omega|)\right] * [\delta(\omega - 4\pi) - \delta(\omega + 4\pi)]$; $\mathscr{F}^{-1}[\theta](x) = \frac{\sqrt{2}}{50\sqrt{\pi}} \sin(4\pi x) \frac{\sin^2(\pi x)}{x^2}$.



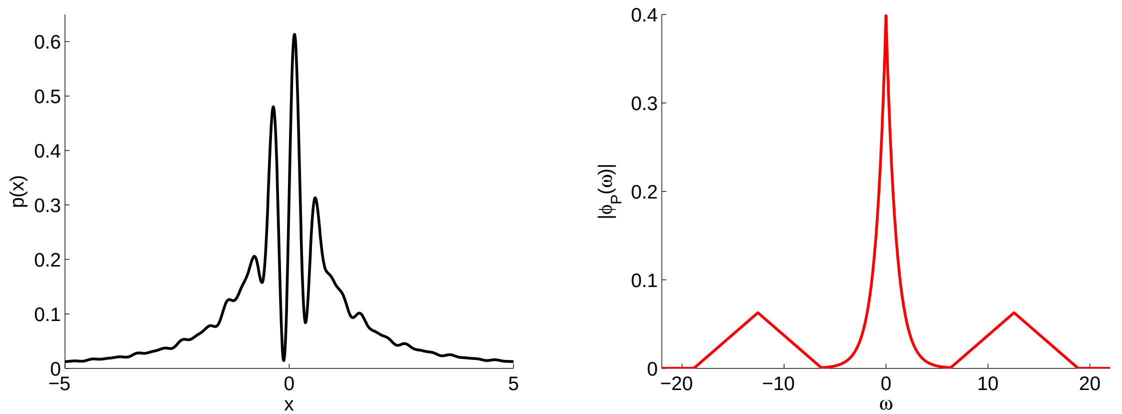


Example : cntd.

- $q(x) = \frac{1}{\pi(1+x^2)}$; $\phi_{\mathbb{Q}}(\omega) = \frac{1}{\sqrt{2\pi}} e^{-|\omega|}$.



- $p(x) = q(x) + \mathscr{F}^{-1}[\theta](x)$; $\phi_{\mathbb{P}}(\omega) = \phi_{\mathbb{Q}}(\omega) + \theta(\omega)$.



Useful Result

Corollary

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\} \subset (\mathcal{H}, k)$
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*
- *supp*(ψ) *is compact.*

Then k is characteristic to $\mathcal{P}$.

- All compactly supported continuous kernels are characteristic to $\mathcal{P}$.
- Computationally advantageous in practice.

So far, $\text{supp}(\Lambda) = \mathbb{R}^d \Leftrightarrow k$ is characteristic to $\mathcal{P}$.

- Can k with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ be characteristic to some $\mathcal{D} \subsetneq \mathcal{P}$?

Useful Result

Corollary

Let

- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\} \subset (\mathcal{H}, k)$
- $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; *bounded and continuous.*
- *supp*(ψ) *is compact.*

Then k is characteristic to $\mathcal{P}$.

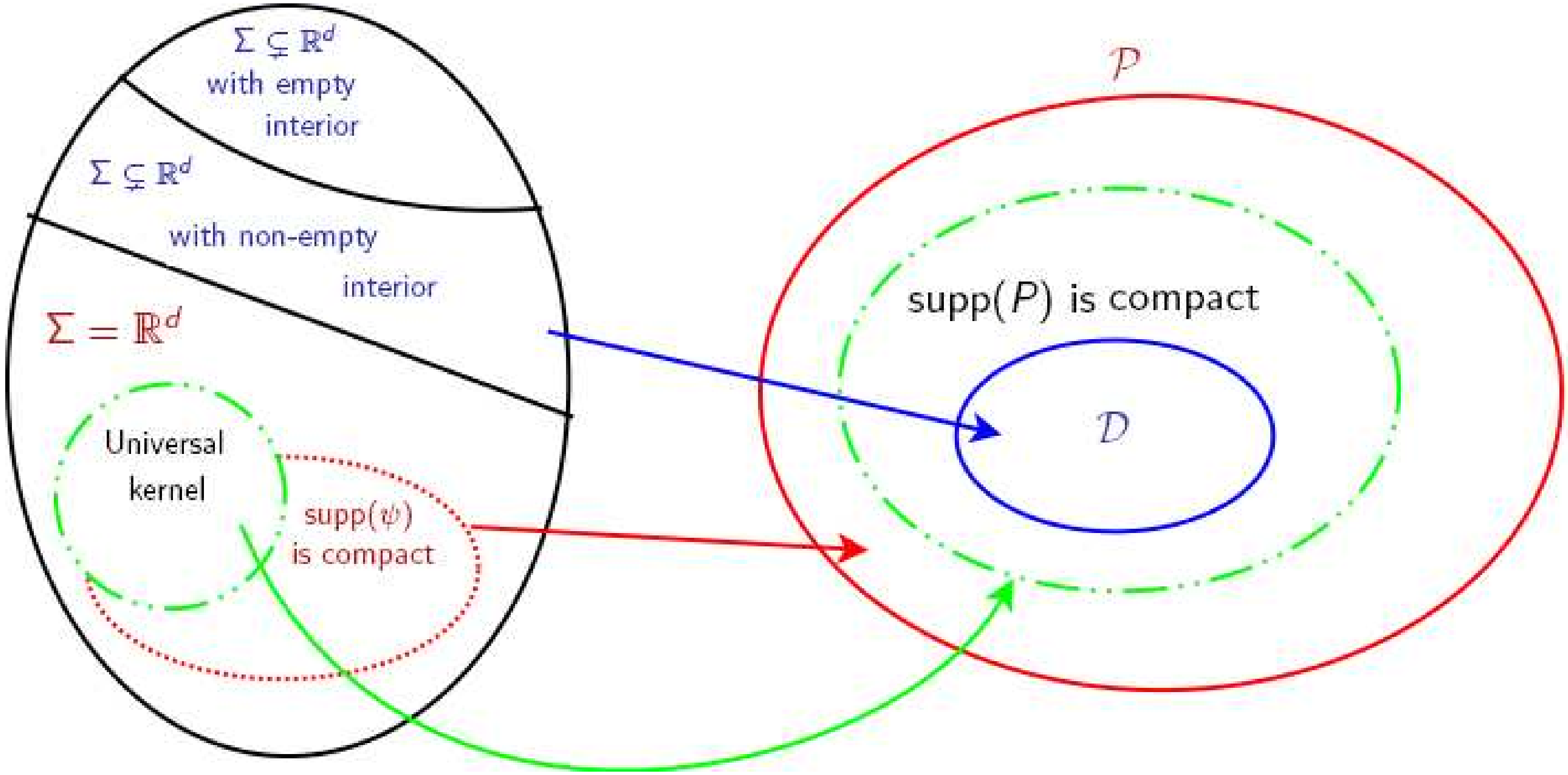
- All compactly supported continuous kernels are characteristic to $\mathcal{P}$.
- Computationally advantageous in practice.

So far, $\text{supp}(\Lambda) = \mathbb{R}^d \Leftrightarrow k$ is characteristic to $\mathcal{P}$.

- Can k with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ be characteristic to some $\mathcal{D} \subsetneq \mathcal{P}$?

Summing Up

$\Sigma := \text{supp}(\Lambda)$



Dissimilar Distributions with Small MMD : Example

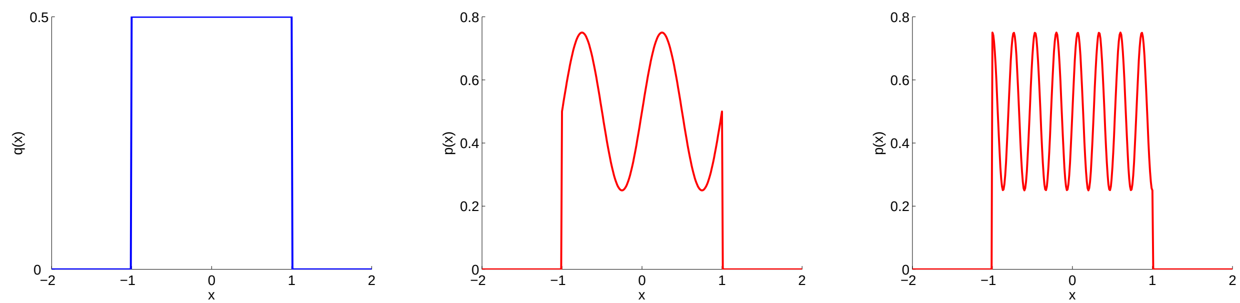
Question: How good is the "characteristic property" in the finite sample setting?

Dissimilar Distributions with Small MMD : Example

Question: How good is the "characteristic property" in the finite sample setting?

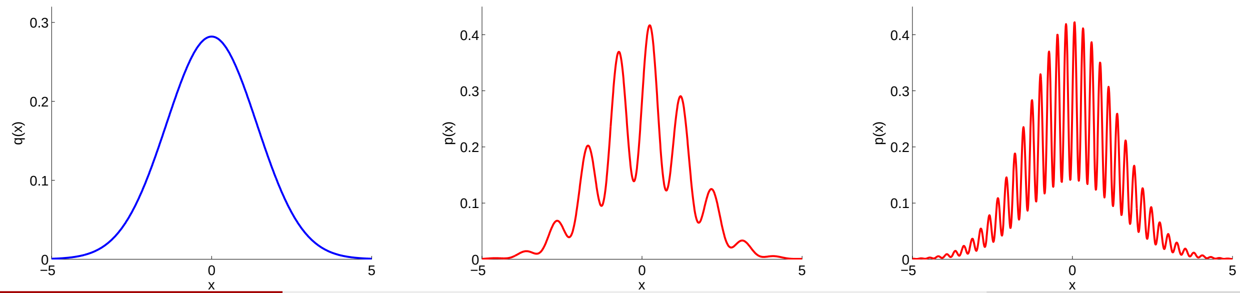
$$p(x) = q(x) + \alpha q(x) \sin(\nu \pi x). \tag{10}$$

- $q = \mathcal{U}[-1, 1]$



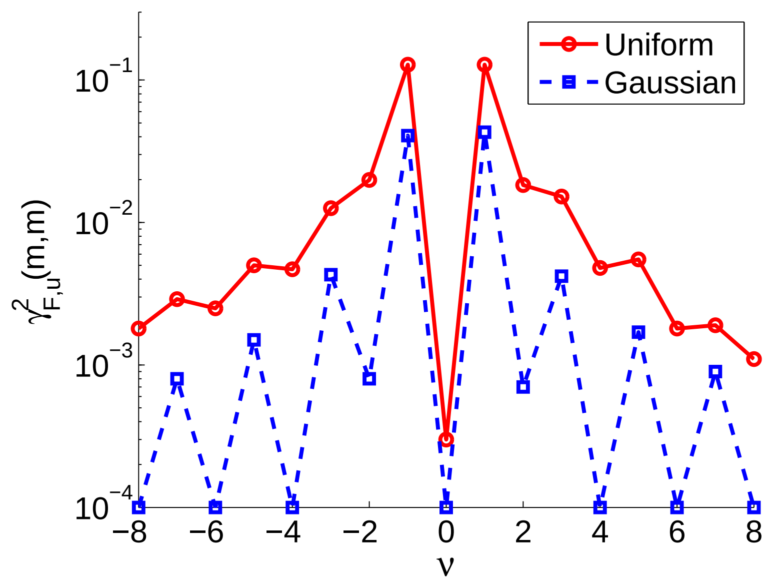
$\nu = 0$ $\qquad$ $\nu = 2$ $\qquad$ $\nu = 7.5$

- $q = \mathcal{N}(0, 2)$

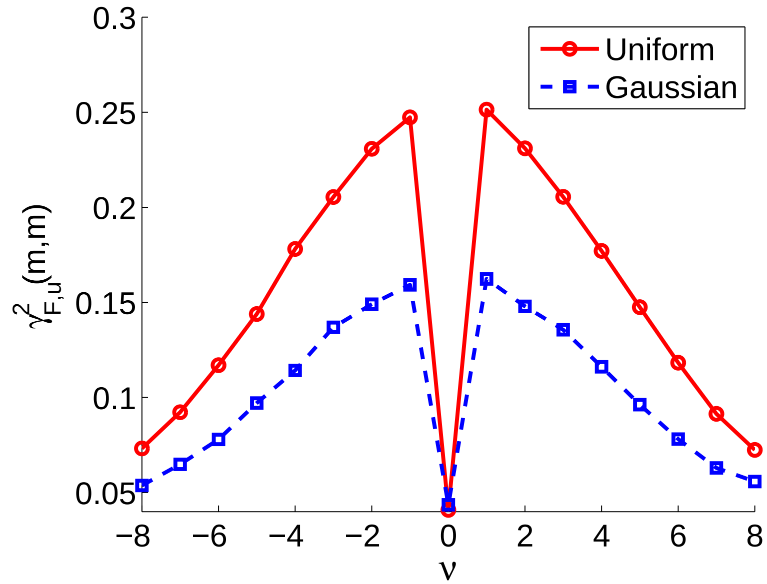


Example : cntd.

$\gamma_{\mathcal{F}}(\hat{\mathbb{P}}, \hat{\mathbb{Q}})$ vs. ν:



B_1-spline



Gaussian

Large ν: $\gamma_{\mathcal{F}}(\hat{\mathbb{P}}, \hat{\mathbb{Q}})$ becomes indistinguishable from zero though $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) > 0$.

Summary

- Maximum mean discrepancy, $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left|\int_M f\, d\mathbb{P} - \int_M f\, d\mathbb{Q}\right|$.
- When $\mathcal{F}$ is a unit ball in an RKHS $(\mathcal{H}, k)$, then $\gamma_{\mathcal{F}}$ is entirely determined by k.
- When $M = \mathbb{R}^d$, $\gamma_{\mathcal{F}}$ is a metric on $\mathcal{P}$ if and only if the Fourier spectrum of a translation-invariant kernel has the entire domain as its support.
- In the finite sample setting, characteristic kernels may have difficulty in distinguishing certain distributions.

Extensions & Open Questions

Extensions:

- M is a compact subset of $\mathbb{R}^d$ but with periodic boundary conditions, e.g. Torus, $\mathbb{T}^d$.
- M: locally compact Abelian group, compact non-abelian group, semigroup.
- Relation of RKHS based $\gamma_{\mathcal{F}}$ to probability metrics induced by other $\mathcal{F}$.
- Role of the speed of decay of the spectrum of k on $\gamma_{\mathcal{F}}$.
- Dependence of $\gamma_{\mathcal{F}}$ on the kernel parameter.

Thank You

References

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008).
Kernel measures of conditional dependence.
In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA. MIT Press.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. (2007).
A kernel method for the two sample problem.
In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press.

Müller, A. (1997).
Integral probability metrics and their generating classes of functions.
Advances in Applied Probability, 29:429–443.

Steinwart, I. (2002).
On the influence of the kernel on the consistency of support vector machines.
Journal of Machine Learning Research, 2:67–93.