On the Relation Between Universality, Characteristic Kernels and RKHS Embedding of Measures

Bharath K. Sriperumbudur[⋆], Kenji Fukumizu[†] and
Gert R. G. Lanckriet[⋆]

[⋆]UC San Diego [†]The Institute of Statistical Mathematics

AISTATS 2010

Outline

- RKHS embedding of probability measures
- *Characteristic kernels*
- *Universal kernels*
  - *Various notions* of universality
  - *Novel characterization of universality*
  - Relation to RKHS embedding of *signed measures*

RKHS Embedding of Probability Measures

- *Input space* : X
- *Feature space* : $\mathcal{H}$ (with reproducing kernel, k)
- *Feature map* : Φ

$$\Phi : X \to \mathcal{H} \qquad\qquad x \mapsto \Phi(x) := k(\cdot, x)$$

RKHS Embedding of Probability Measures

- *Input space* : X
- *Feature space* : $\mathcal{H}$ (with reproducing kernel, k)
- *Feature map* : Φ

$$\Phi : X \to \mathcal{H} \qquad\qquad x \mapsto \Phi(x) := k(\cdot, x)$$

Extension to probability measures:

$$\mathbb{P} \mapsto \Phi(\mathbb{P}) := \int_X k(\cdot, x)\, d\mathbb{P}(x)$$

RKHS Embeddings of Probability Measures

- *Input space* : X
- *Feature space* : $\mathcal{H}$ (with reproducing kernel, k)
- *Feature map* : Φ

$$\Phi : X \to \mathcal{H} \qquad\qquad x \mapsto \Phi(x) := k(\cdot, x)$$

Extension to probability measures:

$$\mathbb{P} \mapsto \Phi(\mathbb{P}) := \underbrace{\int_X k(\cdot, x)\, d\mathbb{P}(x)}_{E_{Y\sim\mathbb{P}}[\Phi(Y)] = E_{Y\sim\mathbb{P}}[k(\cdot, Y)]}$$

RKHS Embeddings of Probability Measures

- *Input space* : X
- *Feature space* : $\mathcal{H}$ (with reproducing kernel, k)
- *Feature map* : Φ

$$\Phi : X \to \mathcal{H} \qquad\qquad x \mapsto \Phi(x) := k(\cdot, x)$$

Extension to probability measures:

$$\mathbb{P} \mapsto \Phi(\mathbb{P}) := \int_X k(\cdot, x)\, d\mathbb{P}(x)$$

Advantage: $\Phi(\mathbb{P})$ can distinguish $\mathbb{P}$ by *high-order moments.*

$$\begin{aligned} k(y,x) &= c_0 + c_1(xy) + c_2(xy)^2 + \cdots \quad (c_i \neq 0) \qquad \text{e.g. } k(y,x) = e^{xy} \\ \Phi(\mathbb{P})(y) &= c_0 + c_1 \left(\int_X x\, d\mathbb{P}(x) \right) y + c_2 \left(\int_X x^2\, d\mathbb{P}(x) \right) y^2 + \cdots \end{aligned}$$

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?

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?

- $\gamma(\mathbb{P}, \mathbb{Q}) = \|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\|_{\mathcal{H}}$: distance metric between $\mathbb{P}$ and $\mathbb{Q}$.

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

- *Test:* Say H_0 if $\widehat{\gamma}(\mathbb{P}, \mathbb{Q}) < \varepsilon$. Otherwise say H_1.

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?

- $\gamma(\mathbb{P}, \mathbb{Q}) = \|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\|_{\mathcal{H}}$: distance metric between $\mathbb{P}$ and $\mathbb{Q}$.

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

- *Test:* Say H_0 if $\widehat{\gamma}(\mathbb{P}, \mathbb{Q}) < \varepsilon$. Otherwise say H_1.

Other applications:

- *Hypothesis testing* : Independence test, Goodness of fit test, etc.

- Feature selection, message passing, density estimation, etc.

Characteristic Kernels

Define: k is *characteristic* if

$$\mathbb{P} \mapsto \int_X k(\cdot, x)\, d\mathbb{P}(x) \text{ is injective.}$$

In other words,

$$\int_X k(\cdot, x)\, d\mathbb{P}(x) = \int_X k(\cdot, x)\, d\mathbb{Q}(x) \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

- When $k(\cdot, x) = e^{\sqrt{-1}\langle \cdot, x \rangle}$, $\Phi(\mathbb{P})$ is the characteristic function of $\mathbb{P}$.

Characteristic Kernels

Define: k is *characteristic* if

$$\mathbb{P} \mapsto \int_X k(\cdot, x)\, d\mathbb{P}(x) \text{ is injective.}$$

In other words,

$$\int_X k(\cdot, x)\, d\mathbb{P}(x) = \int_X k(\cdot, x)\, d\mathbb{Q}(x) \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

- When $k(\cdot, x) = e^{\sqrt{-1}\langle \cdot, x\rangle}$, $\Phi(\mathbb{P})$ is the characteristic function of $\mathbb{P}$.
- Not all kernels are characteristic, e.g., $k(x, y) = x^T y$.

$$\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \not\Rightarrow \mathbb{P} = \mathbb{Q}$$

- *When is k characteristic?* [Gretton et al., 2007, Sriperumbudur et al., 2008, Fukumizu et al., 2008, Fukumizu et al., 2009, Sriperumbudur et al., 2009].

Universal Kernels

- *Regularization approach to supervised learning*

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega[f], \tag{1}$$

where $\lambda > 0$ and $\{(x_i, y_i)\}_{i=1}^{n}$ is the training data.

Universal Kernels

- *Regularization approach to supervised learning*

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega[f], \tag{1}$$

where $\lambda > 0$ and $\{(x_i, y_i)\}_{i=1}^{n}$ is the training data.

- *Representer theorem :* The solution to (1) is of the form

$$f = \sum_{i=1}^{n} c_i k(\cdot, x_i),$$

where $\{c_i\}_{i=1}^{n} \subset \mathbb{R}$ are the parameters typically obtained from the training data.

Universal Kernels

- *Regularization approach to supervised learning*

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega[f], \tag{1}$$

  where $\lambda > 0$ and $\{(x_i, y_i)\}_{i=1}^{n}$ is the training data.

- *Representer theorem :* The solution to (1) is of the form

$$f = \sum_{i=1}^{n} c_i k(\cdot, x_i),$$

  where $\{c_i\}_{i=1}^{n} \subset \mathbb{R}$ are the parameters typically obtained from the training data.

- *Question:* Can f approximate any *target function* arbitrarily "well" as $n \to \infty$?

Universal Kernels

- *Regularization approach to supervised learning*

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega[f], \tag{1}$$

  where $\lambda > 0$ and $\{(x_i, y_i)\}_{i=1}^n$ is the training data.

- *Representer theorem :* The solution to (1) is of the form

$$f = \sum_{i=1}^{n} c_i k(\cdot, x_i),$$

  where $\{c_i\}_{i=1}^n \subset \mathbb{R}$ are the parameters typically obtained from the training data.

- *Question:* Can f approximate any *target function* arbitrarily "well" as $n \to \infty$?

- We need $\mathcal{H}$ *to be "dense" in the space of target functions — k is universal.*

Various Notions of Universality

- *Prior work*
  - *c-universality* [Steinwart, 2001]
  - *cc-universality* [Micchelli et al., 2006]
- Proposed notion: c_0-*universality*
- *Characterization of c-, cc- and* c_0-*universality* : Relation to RKHS embedding of measures
  - Translation invariant kernels on $\mathbb{R}^d$
  - Radial kernels on $\mathbb{R}^d$

c-universality [Steinwart, 2001]

- X : compact metric space
- k : continuous on $X \times X$
- *Target function space* : $C(X)$, continuous functions on X

Define k to be *c-universal* if $\mathcal{H}$ is dense in $C(X)$ w.r.t. the uniform norm ($\|f\|_u := \sup_{x \in X} |f(x)|$).

c-universality [Steinwart, 2001]

- X : compact metric space
- k : continuous on $X \times X$
- *Target function space* : $C(X)$, continuous functions on X

Define k to be *c-universal* if $\mathcal{H}$ is dense in $C(X)$ w.r.t. the uniform norm ($\|f\|_u := \sup_{x\in X} |f(x)|$).

- *Sufficient conditions* are obtained based on the Stone-Weierstraß theorem. *Not easy to check!*
- *Examples:* Gaussian and Laplacian kernels on any compact subset of $\mathbb{R}^d$.

c-universality [Steinwart, 2001]

- X : compact metric space
- k : continuous on $X \times X$
- *Target function space* : $C(X)$, continuous functions on X

Define k to be *c-universal* if $\mathcal{H}$ is dense in $C(X)$ w.r.t. the uniform norm ($\|f\|_u := \sup_{x \in X} |f(x)|$).

- *Sufficient conditions* are obtained based on the Stone-Weierstraß theorem. *Not easy to check!*
- *Examples:* Gaussian and Laplacian kernels on any compact subset of $\mathbb{R}^d$.

Issue: X is compact which excludes many interesting spaces, such as $\mathbb{R}^d$.

cc-universality [Micchelli et al., 2006]

- X : Hausdorff space
- k : continuous on $X \times X$
- *Target function space* : $C(X)$

Define k to be *cc-universal* if $\mathcal{H}$ is dense in $C(X)$ endowed with the *topology of compact convergence.*

cc-universality [Micchelli et al., 2006]

- X : Hausdorff space
- k : continuous on $X \times X$
- *Target function space* : $C(X)$

Define k to be *cc-universal* if $\mathcal{H}$ is dense in $C(X)$ endowed with the *topology of compact convergence.*

In other words, for any compact set $Z \subset X$, $\mathcal{H}_{|Z} := \{f_{|Z} : f \in \mathcal{H}\}$ is dense in $C(Z)$ w.r.t. $\|\cdot\|_u$.

cc-universality [Micchelli et al., 2006]

- X : Hausdorff space
- k : continuous on $X \times X$
- *Target function space* : $C(X)$

Define k to be *cc-universal* if $\mathcal{H}$ is dense in $C(X)$ endowed with the *topology of compact convergence.*

- *Necessary and sufficient conditions* are obtained, which are related to the injectivity of RKHS embedding of measures.
- *Examples:* Gaussian, Laplacian and Sinc kernels on $\mathbb{R}^d$.

cc-universality [Micchelli et al., 2006]

- X : Hausdorff space
- k : continuous on $X \times X$
- *Target function space* : $C(X)$

Define k to be *cc-universal* if $\mathcal{H}$ is dense in $C(X)$ endowed with the *topology of compact convergence.*

- *Necessary and sufficient conditions* are obtained, which are related to the injectivity of RKHS embedding of measures.
- *Examples:* Gaussian, Laplacian and Sinc kernels on $\mathbb{R}^d$.

Issue: Topology of compact convergence is *weaker* than the topology of uniform convergence.

Proposed Notion: c_0-universality

- X : locally compact Hausdorff (LCH) space
- *Target function space* : $C_0(X)$, the space of bounded continuous functions that "vanish at infinity" (for every $\epsilon > 0$, $\{x \in X : |f(x)| \geq \epsilon\}$ is compact).
- k is bounded and $k(\cdot, x) \in C_0(X)$ for all $x \in X$.

Proposed Notion: c_0-universality

- X : locally compact Hausdorff (LCH) space
- *Target function space* : $C_0(X)$, the space of bounded continuous functions that "vanish at infinity" (for every $\epsilon > 0$, $\{x \in X : |f(x)| \geq \epsilon\}$ is compact).
- k is bounded and $k(\cdot, x) \in C_0(X)$ for all $x \in X$.

Define k to be *c_0-universal* if $\mathcal{H}$ is dense in $C_0(X)$ w.r.t. $\|\cdot\|_u$.

- Handles non-compact X and ensures uniform convergence over entire X.

Embedding Characterization of Universality

Theorem

- *k is c_0-universal if and only if*

$$\mu \mapsto \int_X k(\cdot, x)\, d\mu(x),\ \mu \in M_b(X),$$

is injective. $M_b(X)$ is the space of finite signed Radon measures on X.

Embedding Characterization of Universality

Theorem

- k is c_0-universal if and only if

$$\mu \mapsto \int_X k(\cdot, x)\, d\mu(x),\ \mu \in M_b(X),$$

 is injective. $M_b(X)$ is the space of finite signed Radon measures on X.

- k is cc-universal if and only if

$$\mu \mapsto \int_X k(\cdot, x)\, d\mu(x),\ \mu \in M_{bc}(X),$$

 is injective. $M_{bc}(X) = \{\mu \in M_b(X) \mid supp(\mu) \text{ is compact}\}$.

Embedding Characterization of Universality

Theorem

- *k is c_0-universal if and only if*

$$\mu \mapsto \int_X k(\cdot, x)\, d\mu(x),\ \mu \in M_b(X),$$

  is injective. $M_b(X)$ is the space of finite signed Radon measures on X.

- *k is cc-universal if and only if*

$$\mu \mapsto \int_X k(\cdot, x)\, d\mu(x),\ \mu \in M_{bc}(X),$$

  is injective. $M_{bc}(X) = \{\mu \in M_b(X) \,|\, supp(\mu) \text{ is compact}\}$.

- *k is c-universal if and only if*

$$\mu \mapsto \int_X k(\cdot, x)\, d\mu(x),\ \mu \in M_b(X),$$

  is injective.

Postive Definite Characterization of Universality

Theorem

- *k is c_0-universal (resp. c-universal) if and only if*

$$\int_X \int_X k(x, y)\, d\mu(x)\, d\mu(y) > 0, \; \forall\, \mu \in M_b(X) \backslash \{0\}.$$

Postive Definite Characterization of Universality

Theorem

- *k is c_0-universal (resp. c-universal) if and only if*

$$\int_X \int_X k(x,y)\, d\mu(x)\, d\mu(y) > 0,\ \forall\, \mu \in M_b(X)\backslash\{0\}.$$

- *k is cc-universal if and only if*

$$\int_X \int_X k(x,y)\, d\mu(x)\, d\mu(y) > 0,\ \forall\, \mu \in M_{bc}(X)\backslash\{0\}.$$

Postive Definite Characterization of Universality

Theorem

- *k is c_0-universal (resp. c-universal) if and only if*

$$\int_X \int_X k(x,y)\, d\mu(x)\, d\mu(y) > 0,\ \forall\, \mu \in M_b(X)\backslash\{0\}.$$

- *k is cc-universal if and only if*

$$\int_X \int_X k(x,y)\, d\mu(x)\, d\mu(y) > 0,\ \forall\, \mu \in M_{bc}(X)\backslash\{0\}.$$

- *If k is c-, cc- or c_0-universal, then it is strictly positive definite.*

X is an LCH space: Summary

$$
\begin{array}{ccc}
\begin{array}{c}\clubsuit \text{ holds} \\ \forall\, \mu \in M_b(X)\end{array} & \begin{array}{c}\Longrightarrow \\ \not\Longleftarrow\end{array} & \begin{array}{c}\clubsuit \text{ holds} \\ \forall\, \mu \in M_{bc}(X)\end{array} \\
\Updownarrow & & \Updownarrow \\
c_0\text{-}\mathit{universal} & \begin{array}{c}\Longrightarrow \\ \not\Longleftarrow\end{array} & cc\text{-}\mathit{universal} \\
\Downarrow \;\; \not\Uparrow & & \Downarrow \;\; \not\Uparrow \\
\mathit{characteristic} & \begin{array}{c}\overset{?}{\Longrightarrow} \\ \not\Longleftarrow\end{array} & \mathit{strictly\ pd}
\end{array}
$$

$$\clubsuit \; : \; \iint_X k(x,y)\, d\mu(x)\, d\mu(y) > 0$$

Translation Invariant Kernels on $\mathbb{R}^d$

$X = \mathbb{R}^d$ and $k(x, y) = \psi(x - y)$, where

$$\psi(x) = \int_{\mathbb{R}^d} e^{\sqrt{-1}x^T\omega}\, d\Lambda(\omega),\ x \in \mathbb{R}^d,$$

and Λ is a non-negative finite Borel measure.

Translation Invariant Kernels on $\mathbb{R}^d$

$X = \mathbb{R}^d$ and $k(x, y) = \psi(x - y)$, where

$$\psi(x) = \int_{\mathbb{R}^d} e^{\sqrt{-1}x^T\omega}\, d\Lambda(\omega),\ x \in \mathbb{R}^d,$$

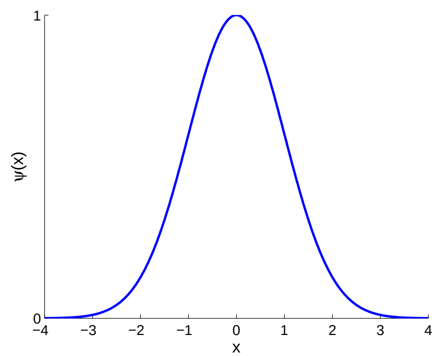
and Λ is a non-negative finite Borel measure.

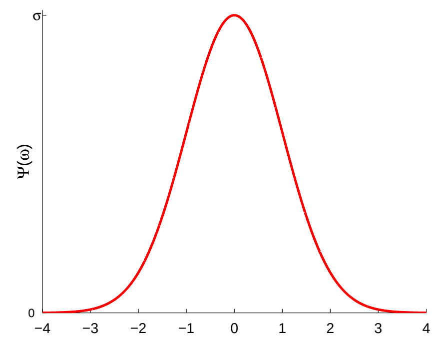
Theorem

- *k is c_0-universal if and only if $supp(\Lambda) = \mathbb{R}^d$.*
- *k is c_0-universal if and only if it is characteristic.*
- *If $supp(\Lambda)$ has a non-empty interior, then k is cc-universal. [Micchelli et al., 2006]*

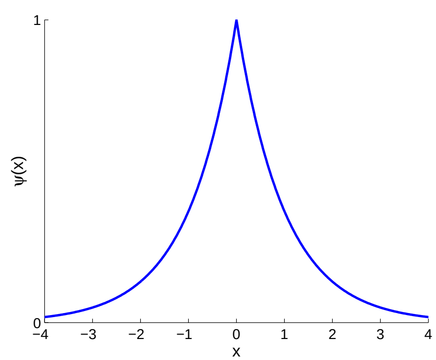
Examples

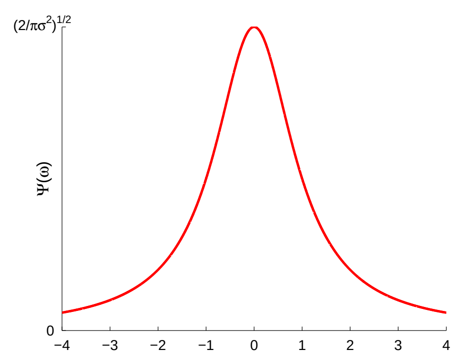
- Gaussian kernel: $\psi(x) = e^{-x^2/2\sigma^2}$; $\Psi(\omega) = \sigma e^{-\sigma^2\omega^2/2}$; $d\Lambda(\omega) = \Psi(\omega)\, d\omega$.





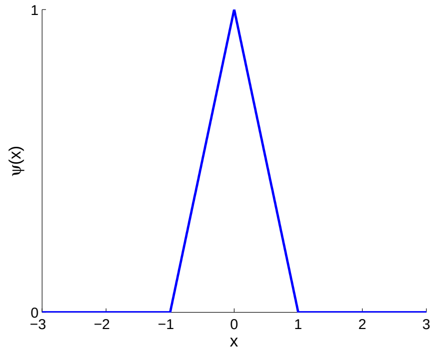
- Laplacian kernel: $\psi(x) = e^{-\sigma|x|}$; $\Psi(\omega) = \sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$.

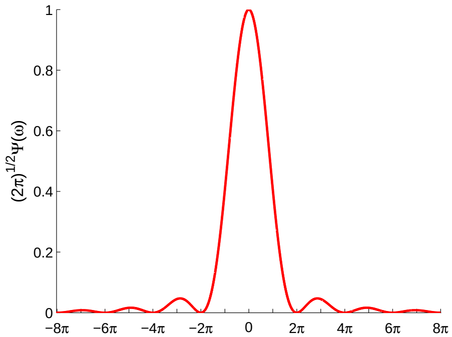




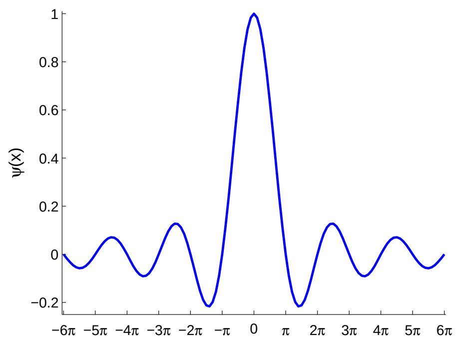
Examples

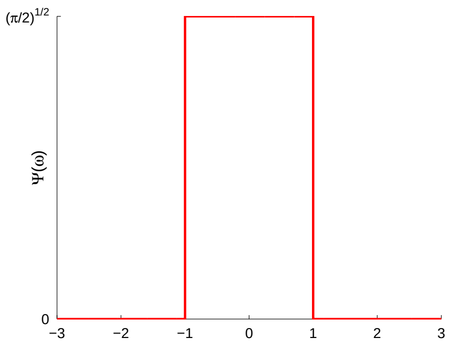
- B_1-spline kernel: $\psi(x) = (1 - |x|)\mathbb{1}_{[-1,1]}(x)$; $\Psi(\omega) = \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{\sin^2(\frac{\omega}{2})}{\omega^2}$.





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





Translation Invariant Kernels on $\mathbb{R}^d$: Summary

$$
\begin{tikzcd}[column sep=huge, row sep=huge]
\operatorname{supp}(\Lambda) = \mathbb{R}^d \arrow[r, Rightarrow, shift left] \arrow[d, Leftrightarrow] & (\operatorname{supp}(\Lambda))^\circ \neq \emptyset \arrow[l, Rightarrow, shift left, "/" marking] \arrow[d, Rightarrow, shift right] \\
c_0\text{-}\mathit{universal} \arrow[r, Rightarrow, shift left] \arrow[d, Leftrightarrow] & cc\text{-}\mathit{universal} \arrow[l, Rightarrow, shift left, "/" marking] \arrow[u, Rightarrow, shift right, "?"'] \arrow[d, Rightarrow, shift right] \\
\mathit{characteristic} \arrow[r, Rightarrow, shift left] & \mathit{strictly\ pd} \arrow[l, Rightarrow, shift left, "/" marking] \arrow[u, Rightarrow, shift right, "?\ (\spadesuit)"']
\end{tikzcd}
$$

$$\spadesuit \ : \ \psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$$

Radial Kernels on $\mathbb{R}^d$

Let

$$k(x, y) = \int_{[0,\infty)} e^{-t\|x-y\|_2^2} \, d\nu(t),$$

where ν is a finite non-negative Borel measure on $[0, \infty)$.

- *Examples:* Gaussian kernel, Inverse multi-quadratic kernel, $k(x, y) = (c^2 + \|x - y\|_2^2)^{-\beta}$, $\beta > \frac{d}{2}$, $c > 0$, etc.

Radial Kernels on $\mathbb{R}^d$

Let

$$k(x,y) = \int_{[0,\infty)} e^{-t\|x-y\|_2^2}\, d\nu(t),$$

where ν is a finite non-negative Borel measure on $[0,\infty)$.

- *Examples:* Gaussian kernel, Inverse multi-quadratic kernel, $k(x,y) = (c^2 + \|x-y\|_2^2)^{-\beta}$, $\beta > \frac{d}{2}$, $c > 0$, etc.

Theorem

The following conditions are equivalent.

- *supp*$(\nu) \neq \{0\}$.
- *k is c_0-universal.*
- *k is cc-universal.*
- *k is characteristic.*
- *k is strictly pd.*

Radial Kernels on $\mathbb{R}^d$: Summary

$$
\begin{tikzcd}
\mathrm{supp}(\nu) \neq \{0\} \arrow[d, Leftrightarrow] & \\
c_0\textit{-universal} \arrow[r, Leftrightarrow] \arrow[d, Leftrightarrow] & \textit{cc-universal} \arrow[d, Leftrightarrow] \\
\textit{characteristic} \arrow[r, Leftrightarrow] & \textit{strictly pd}
\end{tikzcd}
$$

Summary

- *Characteristic kernel*
  - Injective RKHS embedding of probability measures.
  - *Applications:* Hypothesis testing, feature selection, etc.
- *Universal kernel*
  - Consistency of learning algorithms.
  - Injective RKHS embedding of finite signed Radon measures.
- *Clarified the relation* between various notions of universality and characteristic kernels.

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.
- Fukumizu, K., Sriperumbudur, B. K., Gretton, A., and Schölkopf, B. (2009).
  Characteristic kernels on groups and semigroups.
  In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 473–480.
- 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.
- Micchelli, C. A., Xu, Y., and Zhang, H. (2006).
  Universal kernels.
  Journal of Machine Learning Research, 7:2651–2667.
- Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Lanckriet, G. R. G., and Schölkopf, B. (2009).
  Kernel choice and classifiability for RKHS embeddings of probability distributions.
  In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1750–1758. MIT Press.
- Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G. R. G., and Schölkopf, B. (2008).
  Injective Hilbert space embeddings of probability measures.
  In Servedio, R. and Zhang, T., editors, *Proc. of the 21st Annual Conference on Learning Theory*, pages 111–122.
- Steinwart, I. (2001).
  On the influence of the kernel on the consistency of support vector machines.
  Journal of Machine Learning Research, 2:67–93.