Kernel Choice and Classifiability for RKHS Embeddings of Probability Distributions

Bharath K. Sriperumbudur[⋆], Kenji Fukumizu[†], Arthur Gretton[‡,×], Gert R. G. Lanckriet[⋆] and Bernhard Schölkopf[×]

[⋆]UC San Diego [†]The Institute of Statistical Mathematics
[‡] CMU [×]MPI for Biological Cybernetics

NIPS 2009

RKHS Embeddings of Probability Measures

- *Input space* : X
- *Feature space* : $\mathcal{H}$
- *Feature map* : Φ

$$\Phi : X \to \mathcal{H} \qquad\qquad x \mapsto \Phi(x).$$

Extension to probability measures:

$$\mathbb{P} \mapsto \Phi(\mathbb{P})$$

Distance between $\mathbb{P}$ *and* $\mathbb{Q}$*:*

$$\gamma(\mathbb{P}, \mathbb{Q}) = \|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\|_{\mathcal{H}}.$$

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

Applications

- *Hypothesis testing*
  - Testing for independence and conditional independence
  - Goodness of fit test
- *Density estimation* : quality of the estimate, convergence results.
- *Central limit theorems*
- *Information theory*

Popular examples:

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

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

Integral Probability Metrics

- The *integral probability metric* [Müller, 1997] between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\mathbb{P}} f - \mathbb{E}_{\mathbb{Q}} f \right|.$$

- Many popular probability metrics can be obtained by appropriately choosing $\mathcal{F}$.
  - *Total variation distance* : $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$.
  - *Wasserstein distance* : $\mathcal{F} = \{f : \|f\|_L \leq 1\}$.
  - *Dudley metric* : $\mathcal{F} = \{f : \|f\|_L + \|f\|_\infty \leq 1\}$.

- *well-studied* in statistics and probability theory.

$\mathcal{F}$ is a Reproducing Kernel Hilbert Space

- $\mathcal{H}$: reproducing kernel Hilbert space (RKHS).
- k : measurable, bounded, real-valued reproducing kernel.
- $\mathcal{F}$: a unit ball in $\mathcal{H}$, i.e., $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$.

Maximum mean discrepancy (MMD): [Gretton et al., 2007]

$$\gamma_k(\mathbb{P}, \mathbb{Q}) := \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \|\mathrm{E}_{\mathbb{P}} k - \mathrm{E}_{\mathbb{Q}} k\|_{\mathcal{H}},$$

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

RKHS embedding of probability measures:

$$\mathbb{P} \mapsto \mathrm{E}_{\mathbb{P}} k =: \Phi(\mathbb{P}).$$

Advantages

- *Easy to compute* γ_k unlike other $\mathcal{F}$.
- *k is measurable and bounded:* $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ is a $\sqrt{\frac{mn}{m+n}}$-consistent estimator of $\gamma_k(\mathbb{P}, \mathbb{Q})$ [Gretton et al., 2007].
- *k is translation-invariant on* $\mathbb{R}^d$*:* the rate is independent of d.
- Easy to handle structured domains like graphs and strings.

Characteristic Kernels

When is γ_k a metric?

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathrm{E}_{\mathbb{P}} k = \mathrm{E}_{\mathbb{Q}} k \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Define: k is *characteristic* if

$$\mathrm{E}_{\mathbb{P}} k = \mathrm{E}_{\mathbb{Q}} k \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

- Not all kernels are characteristic, e.g. $k(x, y) = x^T y$.

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_2.$$

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

Outline

- Characterization of characteristic kernels *(visit poster!)*
- *Choice of characteristic kernels*
- Characteristic kernels and binary classification

Choice of Characteristic Kernels

Examples: Gaussian, Laplacian, B_{2l+1}-splines, Poisson kernel, etc.

Suppose k is a Gaussian kernel, $k_\sigma(x,y) = e^{-\frac{\|x-y\|_2^2}{2\sigma^2}}$.

- γ_k is a *function of σ.*
- So γ_k is a family of metrics. *Which one do we use in practice?*
- Note that $\gamma_k \to 0$ as $\sigma \to 0$ or $\sigma \to \infty$.
- Define

$$\gamma(\mathbb{P},\mathbb{Q}) = \sup_{\sigma \in \mathbb{R}_+} \gamma_{k_\sigma}(\mathbb{P},\mathbb{Q}).$$

Classes of Characteristic Kernels

Generalized MMD:

$$\gamma(\mathbb{P},\mathbb{Q}) := \sup_{k\in\mathcal{K}} \gamma_k(\mathbb{P},\mathbb{Q}).$$

Examples for $\mathcal{K}$:

- $\mathcal{K}_g := \{e^{-\sigma\|x-y\|_2^2}, x, y \in \mathbb{R}^d \;:\; \sigma \in \mathbb{R}_+\}$.
- $\mathcal{K}_{rbf} := \{\int_0^\infty e^{-\lambda\|x-y\|_2^2}\, d\mu_\sigma(\lambda), x, y \in \mathbb{R}^d,\ \mu_\sigma \in \mathscr{M}^+ \;:\; \sigma \in \Sigma \subset \mathbb{R}^d\}$, where $\mathscr{M}^+$ is the set of all finite nonnegative Borel measures, μ_σ on $\mathbb{R}_+$ that is not concentrated at zero.
- $\mathcal{K}_{lin} := \{k_\lambda = \sum_{i=1}^{l} \lambda_i k_i | k_\lambda \text{ is pd}, \sum_{i=1}^{l} \lambda_i = 1\}$.
- $\mathcal{K}_{con} := \{k_\lambda = \sum_{i=1}^{l} \lambda_i k_i | \lambda_i \geq 0, \sum_{i=1}^{l} \lambda_i = 1\}$.

Computation

- $$\gamma(\mathbb{P},\mathbb{Q}) = \sup_{k\in\mathcal{K}} \Big[\iint k(x,y)\,d\mathbb{P}(x)\,d\mathbb{P}(y) + \iint k(x,y)\,d\mathbb{Q}(x)\,d\mathbb{Q}(y) - 2\iint k(x,y)\,d\mathbb{P}(x)\,d\mathbb{Q}(y) \Big]^{1/2}.$$

- Suppose $\{X_i\}_{i=1}^m \overset{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{Q}$.
- Let $\mathbb{P}_m := \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$ and $\mathbb{Q}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, where δ_x represents the Dirac measure at x.
- The empirical estimate of $\gamma(\mathbb{P},\mathbb{Q})$:

$$\gamma(\mathbb{P}_m,\mathbb{Q}_n) = \sup_{k\in\mathcal{K}} \left[\sum_{i,j=1}^{m} \frac{k(X_i,X_j)}{m^2} + \sum_{i,j=1}^{n} \frac{k(Y_i,Y_j)}{n^2} - 2\sum_{i,j=1}^{m,n} \frac{k(X_i,Y_j)}{mn} \right]^{1/2}.$$

Question

- *When is γ a metric?*
- *Answer:* If any $k \in \mathcal{K}$ is characteristic, then γ is a metric.

Question

- For a fixed k that is measurable and bounded, [Gretton et al., 2007] have shown that

$$|\gamma_k(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_k(\mathbb{P}, \mathbb{Q})| = O\left(\sqrt{\frac{m+n}{mn}}\right).$$

- *When does* $\gamma(\mathbb{P}_m, \mathbb{Q}_n) \xrightarrow{a.s.} \gamma(\mathbb{P}, \mathbb{Q})$*? What is the rate of convergence?*

Statistical Consistency: Result

Theorem

For any $\mathcal{K}$ *and* $\nu := \sup_{k\in\mathcal{K},x\in M} k(x,x) < \infty$, *with probability at least* $1-\delta$, *the following holds:*

$$\begin{aligned}|\gamma(\mathbb{P}_m,\mathbb{Q}_n)-\gamma(\mathbb{P},\mathbb{Q})| \quad &\leq \quad \sqrt{\frac{8U_m(\mathcal{K})}{m}}+\sqrt{\frac{8U_n(\mathcal{K})}{n}} \\ &+\left(\sqrt{8\nu}+\sqrt{36\nu\log\frac{4}{\delta}}\right)\sqrt{\frac{m+n}{mn}},\end{aligned}$$

where

$$U_m(\mathcal{K}) := \mathbb{E}\left[\sup_{k\in\mathcal{K}}\left|\frac{1}{m}\sum_{i<j}^{m}\rho_i\rho_j k(X_i,X_j)\right|\middle|X_1,\ldots,X_m\right],$$

is the Rademacher chaos complexity and ρ_i *are Rademacher random variables.*

Statistical Consistency: Result

Proposition

Suppose $\mathcal{K}$ is a VC-subgraph class. Then

$$|\gamma(\mathbb{P}_m, \mathbb{Q}_n) - \gamma(\mathbb{P}, \mathbb{Q})| = O\left(\sqrt{\frac{m+n}{mn}}\right).$$

In addition, $\gamma(\mathbb{P}_m, \mathbb{Q}_n) \xrightarrow{a.s.} \gamma(\mathbb{P}, \mathbb{Q})$.

Examples: [Ying and Campbell, 2009, Srebro and Ben-David, 2006]

- $\mathcal{K}_g$, $\mathcal{K}_{rbf}$, $\mathcal{K}_{lin}$, $\mathcal{K}_{con}$, etc.

The Two-Sample Problem

- *Given* : $\{X_1, \ldots, X_m\} \overset{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} \mathbb{Q}$.
- *Determine:* are $\mathbb{P}$ and $\mathbb{Q}$ different?

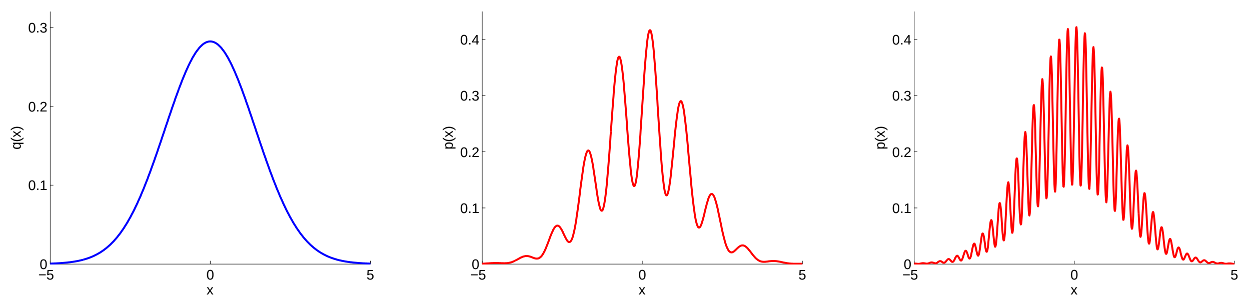
- $\gamma(\mathbb{P}, \mathbb{Q})$: distance metric between $\mathbb{P}$ and $\mathbb{Q}$.

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

- *Test:* Say H_0 if $\widehat{\gamma}(\mathbb{P}, \mathbb{Q}) < \varepsilon$. Otherwise say H_1.
- *Good Test:* Low Type-II error for user-defined Type-I error.

Experiments

- $q = \mathcal{N}(0, \sigma_q^2)$.
- $p(x) = q(x)(1 + \sin \nu x)$.

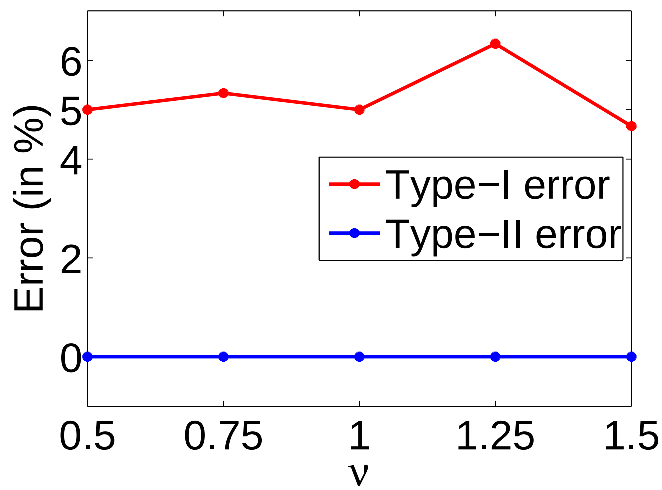


$\nu = 0$ $\nu = 2$ $\nu = 7.5$

- $k(x, y) = \exp(-(x - y)^2/\sigma)$.
- *Test statistics:* $\gamma(\mathbb{P}_m, \mathbb{Q}_m)$ and $\gamma_k(\mathbb{P}_m, \mathbb{Q}_m)$ for various σ.

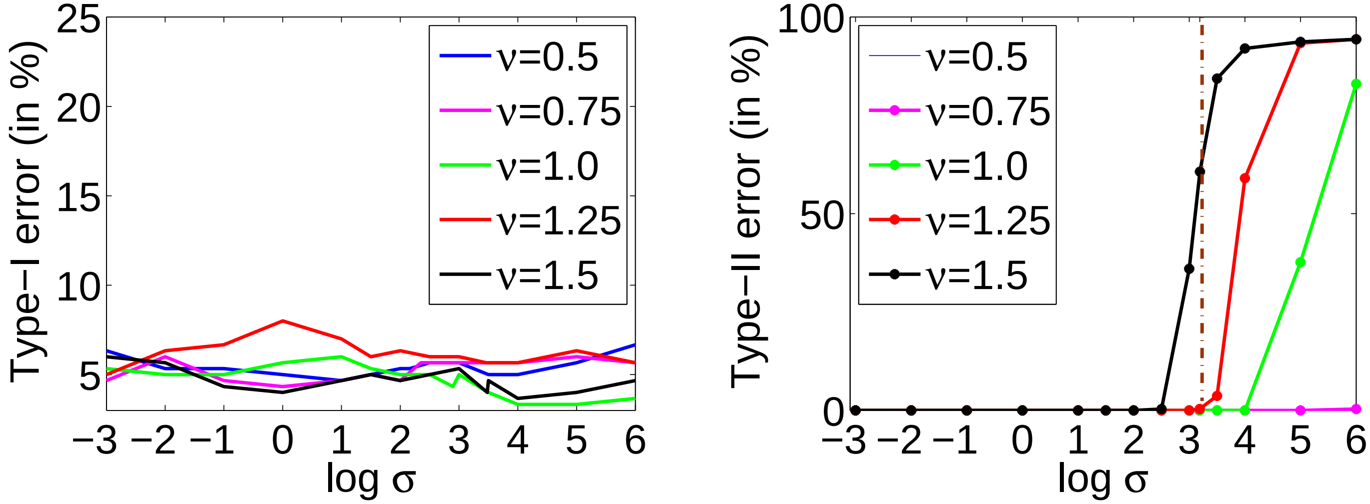
Experiments

$$\gamma(\mathbb{P}, \mathbb{Q})$$



Experiments

$$\gamma_k(\mathbb{P},\mathbb{Q})$$



Outline

- Characterization of characteristic kernels *(visit poster!)*
- Choice of characteristic kernels
- *Characteristic kernels and binary classification*

γ_k and Parzen Window Classifier

Let

- RKHS $(\mathcal{H}, k)$: *k measurable and bounded.*
- $\mathcal{F}_k = \{f : \|f\|_{\mathcal{H}} \leq 1\}$.
- $\mathbb{P}$, $\mathbb{Q}$: class-conditional distributions
- $R_{\mathcal{F}_k}$: Bayes risk of a classifier in $\mathcal{F}_k$.

Then,

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = -R_{\mathcal{F}_k}.$$

- The MMD between class conditionals $\mathbb{P}$ and $\mathbb{Q}$ is negative of the Bayes risk associated with a *Parzen window classifier.*
- Characteristic k is *important*.

γ_k *and Support Vector Machine*

- RKHS $(\mathcal{H}, k)$: *k measurable and bounded.*
- f_{svm} be the solution to the program,

$$\begin{aligned} \inf_{f \in \mathcal{H}} \quad & \|f\|_{\mathcal{H}} \\ s.t. \quad & Y_i f(X_i) \geq 1, \forall\, i. \end{aligned}$$

If k is characteristic, then

$$\frac{1}{\|f_{svm}\|_{\mathcal{H}}} \leq \frac{1}{2}\gamma_k(\mathbb{P}_m, \mathbb{Q}_n).$$

Achievability of Bayes Risk

- $\mathcal{G}_\star$: set of all real-valued measurable functions on M.
- $(\mathcal{H}, k)$: RKHS with measurable and bounded k.
- *Achievability of Bayes risk* :

$$\inf_{g \in \mathcal{H}} R(g) = \inf_{g \in \mathcal{G}_\star} R(g). \qquad (\star\star)$$

Under some technical conditions,

- *$(\star\star) \Rightarrow k$ is characteristic.*
- Suppose $1 \in \mathcal{H}$. *k is characteristic $\Rightarrow (\star\star)$.*

Summary

- *Characteristic kernel*
  - A class of kernels that characterize the probability measure associated with a random variable.
  - MMD is a metric.
- *How to choose characteristic kernels in practice?*
  - Generalized MMD.
  - Performs *better than* MMD in a two-sample test.
- *Characteristic kernels are important in binary classification.*
  - Parzen window classifier and hard-margin SVM.
  - Achievability of Bayes risk.

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.
- 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.
- Müller, A. (1997).
  Integral probability metrics and their generating classes of functions.
  Advances in Applied Probability, 29:429–443.
- Srebro, N. and Ben-David, S. (2006).
  Learning bounds for support vector machines with learned kernels.
  In Lugosi, G. and Simon, H. U., editors, *Proc. of the 19^{th} Annual Conference on Learning Theory*, pages 169–183.
- 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 21^{st} Annual Conference on Learning Theory*, pages 111–122.
- Ying, Y. and Campbell, C. (2009).
  Generalization bounds for learning the kernel.
  In *Proc. of the 22^{nd} Annual Conference on Learning Theory*.