“Large” Reproducing Kernel Hilbert Spaces

Advanced Topics in Machine Learning: COMPGI13

Bharath K. Sriperumbudur and Arthur Gretton

Gatsby Unit

20 March 2012

So far...

- Introduction to RKHS
- RKHS based learning algorithms
  - Kernel PCA
  - Kernel regression
  - SVMs for classification and regression
  - Hypothesis testing (two-sample and independence tests)
  - Feature selection, Clustering, ICA
- Representer theorem

This Lecture

Why RKHS?

How to choose an RKHS?

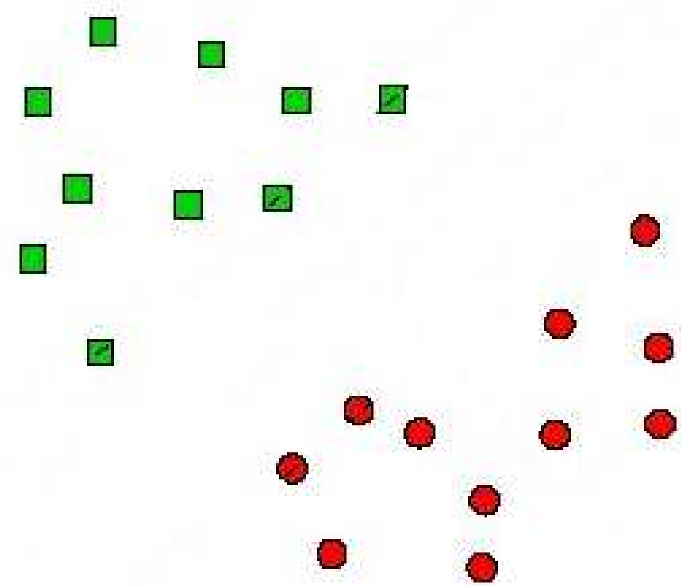
- Polynomial kernels
- Radial basis kernels
- Spline kernel
- Laplacian kernel

"Large" reproducing kernel Hilbert spaces

Binary Classification

- Given: $\mathcal{D} := \{(x_j, y_j)\}_{j=1}^{N}$, $x_j \in \mathcal{X}$, $y_j \in \{-1, +1\}$
- Goal: Learn a function $f : \mathcal{X} \to \mathbb{R}$ such that

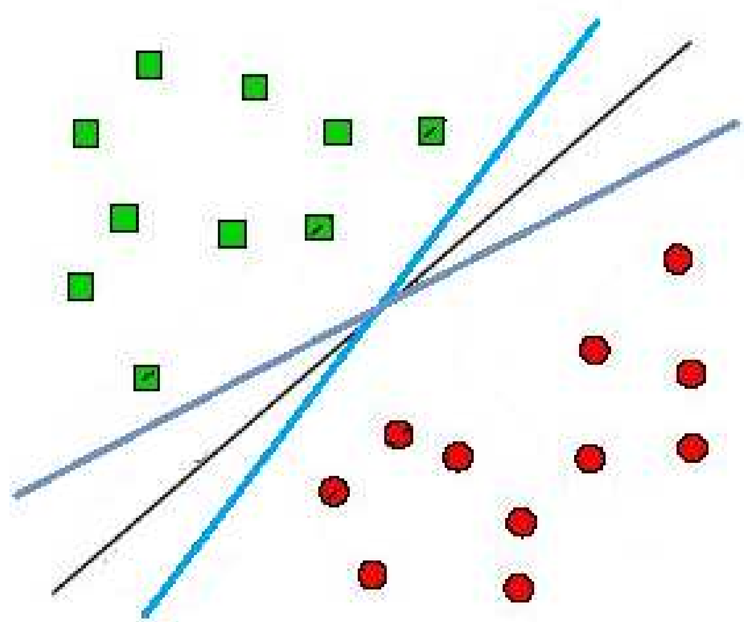
$$y_j = \operatorname{sign}(f(x_j)), \; \forall j = 1, \ldots, N$$



Linear Classifiers

- Linear classifier: $f(x) = \langle w, x \rangle + b$, $w, x \in \mathbb{R}^d$, $b \in \mathbb{R}$
- Find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$y_j \left(\langle w, x_j \rangle + b \right) \geq 0, \; \forall j = 1, \ldots, N.$$



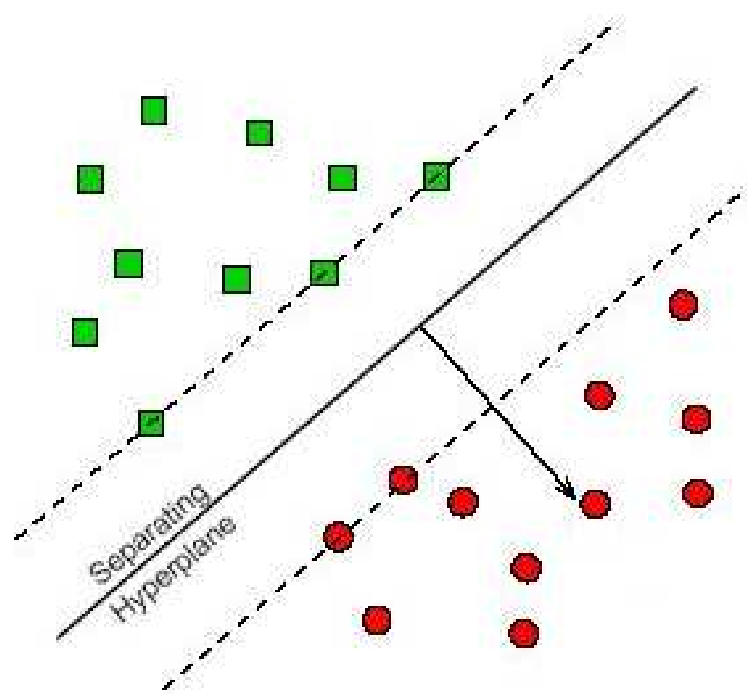
Maximum Margin Classifiers

- Popular Idea: Maximize the *margin* (distance from f to $\mathcal{D}$):

$$\max_{w,b} \min_{j \in \{1,\ldots,N\}} \frac{|\langle w, x_j \rangle + b|}{\|w\|}$$

- Result: Linear support vector machine (SVM)

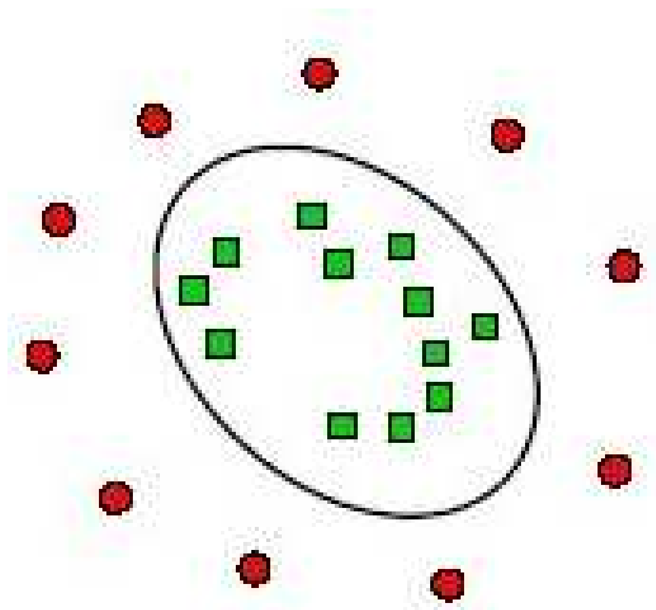
$$\min_{w,b} \{\|w\| \;:\; y_j\left(\langle w, x_j \rangle + b\right) \geq 1, \forall j = 1, \ldots, N\}$$



Non-linear Classifiers

- Issue: Linear SVM is not suitable to classify samples that cannot be linearly separated, i.e.,

$$\nexists\, w \in \mathbb{R}^d,\ b \in \mathbb{R} \text{ s.t. } y_j = \text{sign}(\langle w, x_j \rangle + b),\ \forall\, j = 1, \ldots, N$$



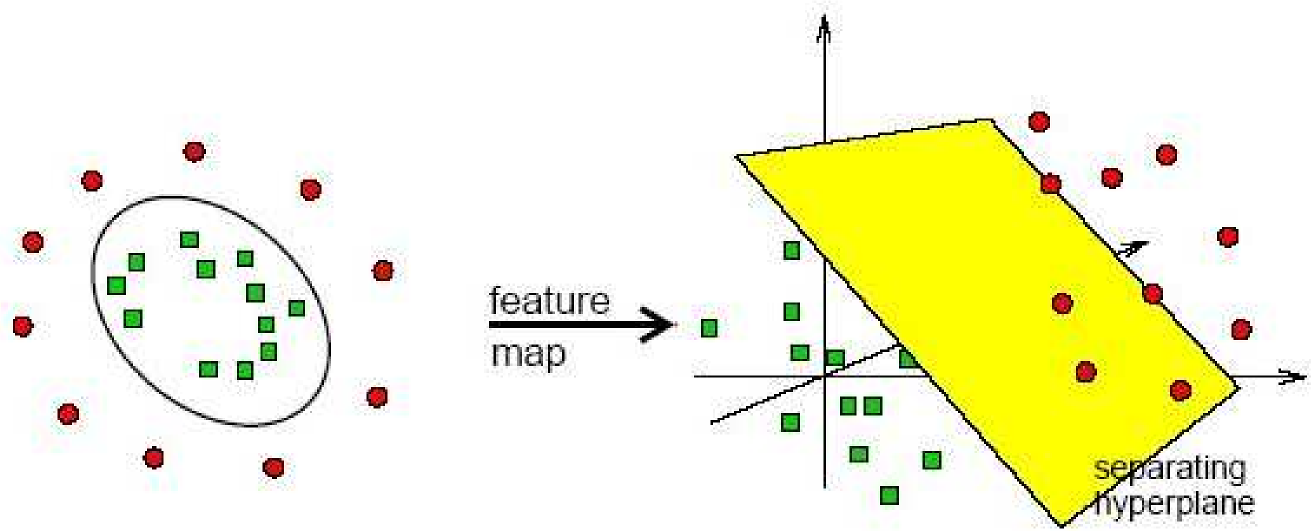
Kernel Classifiers

- Idea: $\mathcal{X} \mapsto \Phi(\mathcal{X}) \subset \mathcal{H}$ and build a linear SVM in the Hilbert space, $\mathcal{H}$. Φ is called the feature map.

$$\min_{\{\alpha_j\}_{j=1}^N} \quad \frac{1}{2}\sum_{l,j=1}^{N} \alpha_l \alpha_j y_l y_j \langle \Phi(x_l), \Phi(x_j)\rangle_{\mathcal{H}} - \sum_{j=1}^{N} \alpha_j$$

$$\text{s.t.} \quad \sum_{j=1}^{N} y_j \alpha_j = 0,\ \alpha_j \geq 0,\ \forall j$$

where $f(x) = \sum_{j=1}^{N} y_j \alpha_j \langle \Phi(x_j), \Phi(x)\rangle_{\mathcal{H}} + b$.



Problem of Learning

- Given a set $D := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of input/output pairs in $X \times Y$.
- Goal: "Learn" a function $f : X \to Y$ such that $f(x)$ is a good approximation of the possible response y for an arbitrary x.

Without any assumptions on the seen and unseen data, no learning is possible.

- Assumption: The past and future pairs (x, y) are independently generated by the same, but of course unknown probability distribution $\mathbf{P}$ on $X \times Y$.

Problem of Learning

- Given a set $D := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of input/output pairs in $X \times Y$.
- Goal: "Learn" a function $f : X \to Y$ such that $f(x)$ is a good approximation of the possible response y for an arbitrary x.

Without any assumptions on the seen and unseen data, no learning is possible.

- Assumption: The past and future pairs (x, y) are independently generated by the same, but of course unknown probability distribution $\mathbf{P}$ on $X \times Y$.

Problem of Learning

- Given a set $D := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of input/output pairs in $X \times Y$.
- Goal: "Learn" a function $f : X \to Y$ such that $f(x)$ is a good approximation of the possible response y for an arbitrary x.

Without any assumptions on the seen and unseen data, no learning is possible.

- Assumption: The past and future pairs (x, y) are independently generated by the same, but of course unknown probability distribution $\mathbf{P}$ on $X \times Y$.

Loss Function

- We need a means to assess the quality of an estimated response $f(x)$ when the true input and output pair is (x, y).
- Loss function: $L : Y \times Y \to [0, \infty)$
  - Squared-loss: $L(y, f(x)) = (y - f(x))^2$
  - Hinge-loss: $L(y, f(x)) = \max(0, 1 - yf(x))$
- Smaller the value of L, better is the approximation of $f(x)$ to y for a given pair (x, y).

Loss Function

- We need a means to assess the quality of an estimated response $f(x)$ when the true input and output pair is (x, y).
- Loss function: $L : Y \times Y \to [0, \infty)$
  - Squared-loss: $L(y, f(x)) = (y - f(x))^2$
  - Hinge-loss: $L(y, f(x)) = \max(0, 1 - yf(x))$
- Smaller the value of L, better is the approximation of $f(x)$ to y for a given pair (x, y).

Loss Function

- We need a means to assess the quality of an estimated response $f(x)$ when the true input and output pair is (x, y).
- Loss function: $L : Y \times Y \rightarrow [0, \infty)$
  - Squared-loss: $L(y, f(x)) = (y - f(x))^2$
  - Hinge-loss: $L(y, f(x)) = \max(0, 1 - yf(x))$
- Smaller the value of L, better is the approximation of $f(x)$ to y for a given pair (x, y).

Risk Functional

- Knowing $L(y, f(x))$ to be small for a particular (x, y) is not sufficient. Need to quantify how small the function

$$(x, y) \mapsto L(y, f(x))$$

  is.

- One common quality measure is the average loss or expected loss of f, called the risk functional i.e.,

$$\mathcal{R}_{L,\mathbf{P}}(f) := \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x, y).$$

Risk Functional

- Knowing $L(y, f(x))$ to be small for a particular (x, y) is not sufficient. Need to quantify how small the function

$$(x, y) \mapsto L(y, f(x))$$

is.

- One common quality measure is the average loss or expected loss of f, called the risk functional i.e.,

$$\mathcal{R}_{L,\mathbf{P}}(f) := \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x, y).$$

Bayes Risk and Bayes Function

- Note that for each f, we have an associated risk, $\mathcal{R}_{L,\mathbf{P}}(f)$.
- Idea: Choose f that has the smallest risk.

$$f^* := \arg \inf_{f:X \to \mathbb{R}} \mathcal{R}_{L,\mathbf{P}}(f),$$

where the infimum is taken over the set of all measurable functions.

- f^* is called the Bayes function and $\mathcal{R}_{L,\mathbf{P}}(f^*)$ is called the Bayes risk.
- If $\mathbf{P}$ is known, finding f^* is often a relatively easy task and there is nothing to learn.
  - Exercise: Find f^* for $L(y, f(x)) = (y - f(x))^2$ and $L(y, f(x)) = |y - f(x)|$?

Bayes Risk and Bayes Function

- Note that for each f, we have an associated risk, $\mathcal{R}_{L,\mathbf{P}}(f)$.
- Idea: Choose f that has the smallest risk.

$$f^* := \arg \inf_{f:X\to\mathbb{R}} \mathcal{R}_{L,\mathbf{P}}(f),$$

where the infimum is taken over the set of all measurable functions.

- f^* is called the Bayes function and $\mathcal{R}_{L,\mathbf{P}}(f^*)$ is called the Bayes risk.
- If $\mathbf{P}$ is known, finding f^* is often a relatively easy task and there is nothing to learn.
  - Exercise: Find f^* for $L(y, f(x)) = (y - f(x))^2$ and $L(y, f(x)) = |y - f(x)|$?

Universal Consistency

- But **P** is unknown
- Without additional information, it is impossible to find an (approximate) minimizer.
- This additional information comes from the training set, $D := \{(x_1, y_1), \ldots, (x_n, y_n)\} \overset{i.i.d.}{\sim} \mathbf{P}$.
- Given D, the goal is to construct $f_D : X \to \mathbb{R}$ such that

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \approx \mathcal{R}_{L,\mathbf{P}}(f^*)$$

- Universally consistent learning algorithm: for all **P** on $X \times Y$, we have

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \to \mathcal{R}_{L,\mathbf{P}}(f^*), \quad n \to \infty$$

in probability.

Universal Consistency

- But $\mathbf{P}$ is unknown
- Without additional information, it is impossible to find an (approximate) minimizer.
- This additional information comes from the training set, $D := \{(x_1, y_1), \ldots, (x_n, y_n)\} \overset{i.i.d.}{\sim} \mathbf{P}$.
- Given D, the goal is to construct $f_D : X \to \mathbb{R}$ such that

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \approx \mathcal{R}_{L,\mathbf{P}}(f^*)$$

- Universally consistent learning algorithm: for all $\mathbf{P}$ on $X \times Y$, we have

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \to \mathcal{R}_{L,\mathbf{P}}(f^*), \quad n \to \infty$$

in probability.

Universal Consistency

- But **P** is unknown
- Without additional information, it is impossible to find an (approximate) minimizer.
- This additional information comes from the training set, $D := \{(x_1, y_1), \dots, (x_n, y_n)\} \overset{i.i.d.}{\sim} \mathbf{P}$.
- Given D, the goal is to construct $f_D : X \to \mathbb{R}$ such that

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \approx \mathcal{R}_{L,\mathbf{P}}(f^*)$$

- Universally consistent learning algorithm: for all **P** on $X \times Y$, we have

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \to \mathcal{R}_{L,\mathbf{P}}(f^*), \quad n \to \infty$$

in probability.

Universal Consistency

- But $\mathbf{P}$ is unknown
- Without additional information, it is impossible to find an (approximate) minimizer.
- This additional information comes from the training set, $D := \{(x_1, y_1), \ldots, (x_n, y_n)\} \overset{i.i.d.}{\sim} \mathbf{P}$.
- Given D, the goal is to construct $f_D : X \to \mathbb{R}$ such that

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \approx \mathcal{R}_{L,\mathbf{P}}(f^*)$$

- Universally consistent learning algorithm: for all $\mathbf{P}$ on $X \times Y$, we have

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \to \mathcal{R}_{L,\mathbf{P}}(f^*), \;\; n \to \infty$$

in probability.

Empirical Risk Minimization

- Since **P** is unknown but is known through D, it is tempting to replace $\mathcal{R}_{L,\mathbf{P}}(f)$ by

$$\mathcal{R}_{L,D}(f) := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)),$$

called the empirical risk and find f_D by

$$f_D := \arg\min_{f:X\to\mathbb{R}} \mathcal{R}_{L,D}(f)$$

- Is it a good idea?
- No! Choose f_D such that $f_D(x) = y_i,\ x = x_i,\ \forall\, i$ and $f_D(x) = 0,\ \textit{otherwise}$.
- $\mathcal{R}_{L,D}(f_D) = 0$ but can be very far from $\mathcal{R}_{L,\mathbf{P}}(f^*)$

Overfitting!!

Empirical Risk Minimization

- Since $\mathbf{P}$ is unknown but is known through D, it is tempting to replace $\mathcal{R}_{L,\mathbf{P}}(f)$ by

$$\mathcal{R}_{L,D}(f) := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)),$$

called the empirical risk and find f_D by

$$f_D := \arg\min_{f:X\to\mathbb{R}} \mathcal{R}_{L,D}(f)$$

- Is it a good idea?

- No! Choose f_D such that $f_D(x) = y_i,\ x = x_i,\ \forall\, i$ and $f_D(x) = 0,\ \textit{otherwise}$.

- $\mathcal{R}_{L,D}(f_D) = 0$ but can be very far from $\mathcal{R}_{L,\mathbf{P}}(f^*)$

Overfitting!!

Empirical Risk Minimization

- Since **P** is unknown but is known through D, it is tempting to replace $\mathcal{R}_{L,\mathbf{P}}(f)$ by

$$\mathcal{R}_{L,D}(f) := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)),$$

called the empirical risk and find f_D by

$$f_D := \arg\min_{f:X\to\mathbb{R}} \mathcal{R}_{L,D}(f)$$

- Is it a good idea?
- No! Choose f_D such that $f_D(x) = y_i,\ x = x_i,\ \forall\, i$ and $f_D(x) = 0,\ \textit{otherwise}$.
- $\mathcal{R}_{L,D}(f_D) = 0$ but can be very far from $\mathcal{R}_{L,\mathbf{P}}(f^*)$

Overfitting!!

Empirical Risk Minimization

- How to avoid overfitting: Choose a small set $\mathcal{F}$ of functions $f : X \to \mathbb{R}$ that is assumed to contain a reasonably good approximation to f^*.

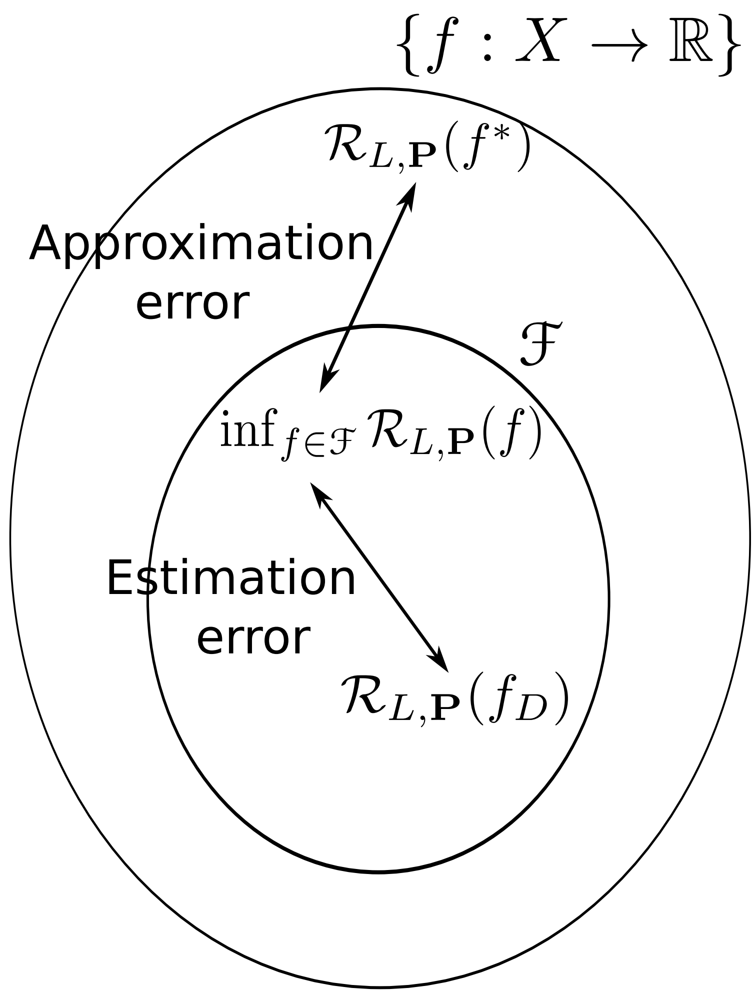
- Do minimization over $\mathcal{F}$:

$$f_D := \arg\min_{f \in \mathcal{F}} \mathcal{R}_{L,D}(f)$$

- Total error: Define $\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} := \inf_{f \in \mathcal{F}} \mathcal{R}_{L,\mathbf{P}}(f)$

$$\mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}_{L,\mathbf{P}}(f^*) = \overbrace{\mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}^*_{L,\mathbf{P},\mathcal{F}}}^{\text{Estimation error}} + \overbrace{\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} - \mathcal{R}_{L,\mathbf{P}}(f^*)}^{\text{Approximation error}}$$

Approximation and Estimation Errors



Regularized Learning

- Let Ω be some functional on $\mathcal{F}$ such that for $c_1 \leq c_2$,

$$\{f \in \mathcal{F} : \Omega(f) \leq c_1\} \subset \{f \in \mathcal{F} : \Omega(f) \leq c_2\}.$$

- Define

$$\begin{aligned} f_D &= \arg \min_{f \in \mathcal{F} : \Omega(f) \leq c} R_{L,D}(f) \\ &= \arg \min_{f \in \mathcal{F} : \Omega(f) \leq c} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \end{aligned}$$

- In the Lagrangian formulation, we have

$$\begin{aligned} f_D &= \arg \min_{f \in \mathcal{F}} R_{L,D}(f) + \lambda \Omega(f) \\ &= \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \Omega(f) \end{aligned}$$

Why RKHS?

- Various choices for $\mathcal{F}$ (with evaluation functional bounded):
  - Lipschitz functions with $\Omega(f) = \|f\|_L$
  - Bounded Lipschitz functions with $\Omega(f) = \|f\|_L + \|f\|_\infty$
  - Bounded measurable functions with $\Omega(f) = \|f\|_\infty$
  - RKHS, $(\mathcal{H}, k)$ with $\Omega(f) = \|f\|_{\mathcal{H}}$
- Advantage with RKHS: For convex L, the regularized objective is a nice convex program.
  - Hinge loss: Support vector machine
  - Squared loss: Kernel regression
- How: By the representer theorem

Can I choose any RKHS?

Why RKHS?

- Various choices for $\mathcal{F}$ (with evaluation functional bounded):
  - Lipschitz functions with $\Omega(f) = \|f\|_L$
  - Bounded Lipschitz functions with $\Omega(f) = \|f\|_L + \|f\|_\infty$
  - Bounded measurable functions with $\Omega(f) = \|f\|_\infty$
  - RKHS, $(\mathcal{H}, k)$ with $\Omega(f) = \|f\|_{\mathcal{H}}$
- Advantage with RKHS: For convex L, the regularized objective is a nice convex program.
  - Hinge loss: Support vector machine
  - Squared loss: Kernel regression
- How: By the representer theorem

Can I choose any RKHS?

Loss Interpretation of Maximum Mean Discrepancy

Suppose $Y = \{-1, 1\}$ and $L(y,t) = -2yt$.

$$\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f\in\mathcal{F}} \int_{X\times Y} L(y, f(x))\, d\mathbf{P}(x,y) \\
&= \inf_{f\in\mathcal{F}} \int_{X\times Y} L(y, f(x))\, d\mathbf{P}(x|y)\, d\mathbf{P}(y) \\
&= \inf_{f\in\mathcal{F}} \int_{X} L(1, f(x))\, \mathbf{P}(y=1)\, d\mathbf{P}(x|y=1) \\
&\qquad + \int_X L(-1, f(x))\, \mathbf{P}(y=-1)\, d\mathbf{P}(x|-1)
\end{aligned}$$

Let $\mathbf{P}(y=1) = \frac{1}{2}$, $\mathbf{P}(x|y=1) = \mathbb{P}(x)$ and $\mathbf{P}(x|y=-1) = \mathbb{Q}(x)$. Therefore,

$$\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f\in\mathcal{F}} \int_X f(x)\, d\mathbb{Q}(x) - \int_X f(x)\, d\mathbb{P}(x) \\
&= -\sup_{f\in\mathcal{F}} \int_X f(x)\, d\mathbb{P}(x) - \int_X f(x)\, d\mathbb{Q}(x)
\end{aligned}$$

Loss Interpretation of Maximum Mean Discrepancy

Suppose $Y = \{-1, 1\}$ and $L(y, t) = -2yt$.

$$\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f \in \mathcal{F}} \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x, y) \\
&= \inf_{f \in \mathcal{F}} \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x|y)\, d\mathbf{P}(y) \\
&= \inf_{f \in \mathcal{F}} \int_X L(1, f(x))\, \mathbf{P}(y = 1)\, d\mathbf{P}(x|y = 1) \\
&\quad + \int_X L(-1, f(x))\, \mathbf{P}(y = -1)\, d\mathbf{P}(x| - 1)
\end{aligned}$$

Let $\mathbf{P}(y = 1) = \frac{1}{2}$, $\mathbf{P}(x|y = 1) = \mathbb{P}(x)$ and $\mathbf{P}(x|y = -1) = \mathbb{Q}(x)$. Therefore,

$$\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f \in \mathcal{F}} \int_X f(x)\, d\mathbb{Q}(x) - \int_X f(x)\, d\mathbb{P}(x) \\
&= -\sup_{f \in \mathcal{F}} \int_X f(x)\, d\mathbb{P}(x) - \int_X f(x)\, d\mathbb{Q}(x)
\end{aligned}$$

Loss Interpretation of Maximum Mean Discrepancy

Suppose $Y = \{-1, 1\}$ and $L(y,t) = -2yt$.

$$\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f\in\mathcal{F}} \int_{X\times Y} L(y, f(x))\, d\mathbf{P}(x,y) \\
&= \inf_{f\in\mathcal{F}} \int_{X\times Y} L(y, f(x))\, d\mathbf{P}(x|y)\, d\mathbf{P}(y) \\
&= \inf_{f\in\mathcal{F}} \int_X L(1, f(x))\, \mathbf{P}(y=1)\, d\mathbf{P}(x|y=1) \\
&\quad + \int_X L(-1, f(x))\, \mathbf{P}(y=-1)\, d\mathbf{P}(x|-1)
\end{aligned}$$

Let $\mathbf{P}(y=1) = \frac{1}{2}$, $\mathbf{P}(x|y=1) = \mathbb{P}(x)$ and $\mathbf{P}(x|y=-1) = \mathbb{Q}(x)$. Therefore,

$$\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f\in\mathcal{F}} \int_X f(x)\, d\mathbb{Q}(x) - \int_X f(x)\, d\mathbb{P}(x) \\
&= -\sup_{f\in\mathcal{F}} \int_X f(x)\, d\mathbb{P}(x) - \int_X f(x)\, d\mathbb{Q}(x)
\end{aligned}$$

Loss Interpretation of Maximum Mean Discrepancy

Suppose $Y = \{-1, 1\}$ and $L(y, t) = -2yt$.

$$
\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f \in \mathcal{F}} \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x, y) \\
&= \inf_{f \in \mathcal{F}} \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x|y)\, d\mathbf{P}(y) \\
&= \inf_{f \in \mathcal{F}} \int_{X} L(1, f(x))\, \mathbf{P}(y = 1)\, d\mathbf{P}(x|y = 1) \\
&\quad + \int_{X} L(-1, f(x))\, \mathbf{P}(y = -1)\, d\mathbf{P}(x| - 1)
\end{aligned}
$$

Let $\mathbf{P}(y = 1) = \frac{1}{2}$, $\mathbf{P}(x|y = 1) = \mathbb{P}(x)$ and $\mathbf{P}(x|y = -1) = \mathbb{Q}(x)$.
Therefore,

$$
\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f \in \mathcal{F}} \int_{X} f(x)\, d\mathbb{Q}(x) - \int_{X} f(x)\, d\mathbb{P}(x) \\
&= -\sup_{f \in \mathcal{F}} \int_{X} f(x)\, d\mathbb{P}(x) - \int_{X} f(x)\, d\mathbb{Q}(x)
\end{aligned}
$$

Loss Interpretation of Maximum Mean Discrepancy

Suppose $Y = \{-1, 1\}$ and $L(y, t) = -2yt$.

$$\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f \in \mathcal{F}} \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x, y) \\
&= \inf_{f \in \mathcal{F}} \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x|y)\, d\mathbf{P}(y) \\
&= \inf_{f \in \mathcal{F}} \int_X L(1, f(x))\, \mathbf{P}(y = 1)\, d\mathbf{P}(x|y = 1) \\
&\quad + \int_X L(-1, f(x))\, \mathbf{P}(y = -1)\, d\mathbf{P}(x| - 1)
\end{aligned}$$

Let $\mathbf{P}(y = 1) = \frac{1}{2}$, $\mathbf{P}(x|y = 1) = \mathbb{P}(x)$ and $\mathbf{P}(x|y = -1) = \mathbb{Q}(x)$. Therefore,

$$\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f \in \mathcal{F}} \int_X f(x)\, d\mathbb{Q}(x) - \int_X f(x)\, d\mathbb{P}(x) \\
&= -\sup_{f \in \mathcal{F}} \int_X f(x)\, d\mathbb{P}(x) - \int_X f(x)\, d\mathbb{Q}(x)
\end{aligned}$$

Loss Interpretation of Maximum Mean Discrepancy

Suppose $Y = \{-1, 1\}$ and $L(y, t) = -2yt$.

$$
\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f\in\mathcal{F}} \int_{X\times Y} L(y, f(x))\, d\mathbf{P}(x, y) \\
&= \inf_{f\in\mathcal{F}} \int_{X\times Y} L(y, f(x))\, d\mathbf{P}(x|y)\, d\mathbf{P}(y) \\
&= \inf_{f\in\mathcal{F}} \int_{X} L(1, f(x))\, \mathbf{P}(y=1)\, d\mathbf{P}(x|y=1) \\
&\quad + \int_{X} L(-1, f(x))\, \mathbf{P}(y=-1)\, d\mathbf{P}(x|-1)
\end{aligned}
$$

Let $\mathbf{P}(y = 1) = \frac{1}{2}$, $\mathbf{P}(x|y = 1) = \mathbb{P}(x)$ and $\mathbf{P}(x|y = -1) = \mathbb{Q}(x)$. Therefore,

$$
\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} &= \inf_{f\in\mathcal{F}} \int_{X} f(x)\, d\mathbb{Q}(x) - \int_{X} f(x)\, d\mathbb{P}(x) \\
&= -\sup_{f\in\mathcal{F}} \int_{X} f(x)\, d\mathbb{P}(x) - \int_{X} f(x)\, d\mathbb{Q}(x)
\end{aligned}
$$

Loss Interpretation of Maximum Mean Discrepancy

Suppose $Y = \{-1, 1\}$ and $L(y, t) = -2yt$.

$$\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathfrak{F}} &= \inf_{f \in \mathfrak{F}} \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x, y) \\
&= \inf_{f \in \mathfrak{F}} \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x|y)\, d\mathbf{P}(y) \\
&= \inf_{f \in \mathfrak{F}} \int_{X} L(1, f(x))\, \mathbf{P}(y = 1)\, d\mathbf{P}(x|y = 1) \\
&\qquad + \int_{X} L(-1, f(x))\, \mathbf{P}(y = -1)\, d\mathbf{P}(x| - 1)
\end{aligned}$$

Let $\mathbf{P}(y = 1) = \frac{1}{2}$, $\mathbf{P}(x|y = 1) = \mathbb{P}(x)$ and $\mathbf{P}(x|y = -1) = \mathbb{Q}(x)$.
Therefore,

$$\begin{aligned}
\mathcal{R}^*_{L,\mathbf{P},\mathfrak{F}} &= \inf_{f \in \mathfrak{F}} \int_X f(x)\, d\mathbb{Q}(x) - \int_X f(x)\, d\mathbb{P}(x) \\
&= -\sup_{f \in \mathfrak{F}} \int_X f(x)\, d\mathbb{P}(x) - \int_X f(x)\, d\mathbb{Q}(x)
\end{aligned}$$

Loss Interpretation of Maximum Mean Discrepancy

$$\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} = -MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$$

- $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$ is a pseudometric on the space of probability measures
  - $MMD(\mathbb{P}, \mathbb{P}, \mathcal{F}) = 0$
  - Symmetry: $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = MMD(\mathbb{Q}, \mathbb{P}, \mathcal{F})$
  - Triangle inequality: $MMD(\mathbb{P}, \mathbb{R}, \mathcal{F}) \leq MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) + MMD(\mathbb{Q}, \mathbb{R}, \mathcal{F})$
- However, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \nRightarrow \mathbb{P} = \mathbb{Q}$
- Only for certain $\mathcal{F}$, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$

Loss Interpretation of Maximum Mean Discrepancy

$$\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} = -MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$$

- $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$ is a pseudometric on the space of probability measures
  - $MMD(\mathbb{P}, \mathbb{P}, \mathcal{F}) = 0$
  - Symmetry: $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = MMD(\mathbb{Q}, \mathbb{P}, \mathcal{F})$
  - Triangle inequality: $MMD(\mathbb{P}, \mathbb{R}, \mathcal{F}) \leq MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) + MMD(\mathbb{Q}, \mathbb{R}, \mathcal{F})$
- However, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \not\Rightarrow \mathbb{P} = \mathbb{Q}$
- Only for certain $\mathcal{F}$, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$

Loss Interpretation of Maximum Mean Discrepancy

$$\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} = -MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$$

- $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$ is a pseudometric on the space of probability measures
  - $MMD(\mathbb{P}, \mathbb{P}, \mathcal{F}) = 0$
  - Symmetry: $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = MMD(\mathbb{Q}, \mathbb{P}, \mathcal{F})$
  - Triangle inequality: $MMD(\mathbb{P}, \mathbb{R}, \mathcal{F}) \leq MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) + MMD(\mathbb{Q}, \mathbb{R}, \mathcal{F})$
- However, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \not\Rightarrow \mathbb{P} = \mathbb{Q}$
- Only for certain $\mathcal{F}$, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$

Loss Interpretation of Maximum Mean Discrepancy

$$\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} = -MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$$

- $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$ is a pseudometric on the space of probability measures
  - $MMD(\mathbb{P}, \mathbb{P}, \mathcal{F}) = 0$
  - Symmetry: $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = MMD(\mathbb{Q}, \mathbb{P}, \mathcal{F})$
  - Triangle inequality: $MMD(\mathbb{P}, \mathbb{R}, \mathcal{F}) \leq MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) + MMD(\mathbb{Q}, \mathbb{R}, \mathcal{F})$
- However, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \not\Rightarrow \mathbb{P} = \mathbb{Q}$
- Only for certain $\mathcal{F}$, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$

Loss Interpretation of Maximum Mean Discrepancy

$$\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} = -MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$$

- $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$ is a pseudometric on the space of probability measures
  - $MMD(\mathbb{P}, \mathbb{P}, \mathcal{F}) = 0$
  - Symmetry: $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = MMD(\mathbb{Q}, \mathbb{P}, \mathcal{F})$
  - Triangle inequality: $MMD(\mathbb{P}, \mathbb{R}, \mathcal{F}) \leq MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) + MMD(\mathbb{Q}, \mathbb{R}, \mathcal{F})$
- However, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \nRightarrow \mathbb{P} = \mathbb{Q}$
- Only for certain $\mathcal{F}$, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$

Loss Interpretation of Maximum Mean Discrepancy

$$\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} = -MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$$

- $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})$ is a pseudometric on the space of probability measures
  - $MMD(\mathbb{P}, \mathbb{P}, \mathcal{F}) = 0$
  - Symmetry: $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = MMD(\mathbb{Q}, \mathbb{P}, \mathcal{F})$
  - Triangle inequality: $MMD(\mathbb{P}, \mathbb{R}, \mathcal{F}) \leq MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) + MMD(\mathbb{Q}, \mathbb{R}, \mathcal{F})$
- However, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \not\Rightarrow \mathbb{P} = \mathbb{Q}$
- Only for certain $\mathcal{F}$, $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$

Choice of $\mathcal{F}$

- Unit Lipschitz ball, $\mathcal{F} = \{\|f\|_L \leq 1\}$: Wasserstein distance
- Unit bounded Lipschitz ball, $\mathcal{F} = \{\|f\|_L + \|f\|_\infty \leq 1\}$: Dudley metric
- Unit sup ball, $\mathcal{F} = \{\|f\|_\infty \leq 1\}$: Total-variation distance

$\mathcal{F}$ is a unit ball in an RKHS?

$\mathcal{F}$ is an RKHS

- When $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$, then

$$
\begin{aligned}
MMD^2(\mathbb{P},\mathbb{Q},\mathcal{F}) &= \left\| \overbrace{\int_X k(\cdot,x)\, d\mathbb{P}(x)}^{\mu_{\mathbb{P}}} - \overbrace{\int_X k(\cdot,x)\, d\mathbb{Q}(x)}^{\mu_{\mathbb{Q}}} \right\|_{\mathcal{H}}^2 \\
&= \overbrace{\int_X \int_X k(x,y)\, d\mathbb{P}(x)\, d\mathbb{P}(y)}^{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}} \\
&\quad + \overbrace{\int_X \int_X k(x,y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y)}^{\langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}} \\
&\quad - 2\overbrace{\int_X \int_X k(x,y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y)}^{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}} \\
&= \int_X \int_X k(x,y) d\mu(x)\, d\mu(y)
\end{aligned}
$$

for $\mu = \mathbb{P} - \mathbb{Q}$.

$\mathcal{F}$ is an RKHS

- When $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$, then

$$
\begin{aligned}
MMD^2(\mathbb{P}, \mathbb{Q}, \mathcal{F}) &= \left\| \overbrace{\int_X k(\cdot, x)\, d\mathbb{P}(x)}^{\mu_{\mathbb{P}}} - \overbrace{\int_X k(\cdot, x)\, d\mathbb{Q}(x)}^{\mu_{\mathbb{Q}}} \right\|_{\mathcal{H}}^2 \\
&= \overbrace{\int_X \int_X k(x, y)\, d\mathbb{P}(x)\, d\mathbb{P}(y)}^{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}} \\
&\quad + \overbrace{\int_X \int_X k(x, y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y)}^{\langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}} \\
&\quad - 2 \overbrace{\int_X \int_X k(x, y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y)}^{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}} \\
&= \int_X \int_X k(x, y) d\mu(x)\, d\mu(y)
\end{aligned}
$$

for $\mu = \mathbb{P} - \mathbb{Q}$.

$\mathcal{F}$ is an RKHS

- When $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$, then

$$
\begin{aligned}
MMD^2(\mathbb{P}, \mathbb{Q}, \mathcal{F}) &= \left\| \overbrace{\int_X k(\cdot, x)\, d\mathbb{P}(x)}^{\mu_{\mathbb{P}}} - \overbrace{\int_X k(\cdot, x)\, d\mathbb{Q}(x)}^{\mu_{\mathbb{Q}}} \right\|_{\mathcal{H}}^2 \\
&= \overbrace{\int_X \int_X k(x, y)\, d\mathbb{P}(x)\, d\mathbb{P}(y)}^{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}} \\
&\quad + \overbrace{\int_X \int_X k(x, y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y)}^{\langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}} \\
&\quad - 2 \overbrace{\int_X \int_X k(x, y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y)}^{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}} \\
&= \int_X \int_X k(x, y) d\mu(x)\, d\mu(y)
\end{aligned}
$$

for $\mu = \mathbb{P} - \mathbb{Q}$.

Not all Kernels are Useful

- $k(x, y) = c$ for all $x, y \in X$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0, \ \forall \, \mathbb{P}, \, \mathbb{Q}.$$

- Another example: $k(x, y) = \langle x, y \rangle_{\mathbb{R}^d}, \ x, y \in \mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|M_{\mathbb{P}} - M_{\mathbb{Q}}\|_{\mathbb{R}^d},$$

where $M_{\mathbb{P}}$ is the mean of $\mathbb{P}$.

- Separable distributions can be made inseparable if the RKHS is not chosen properly.

How to choose $\mathcal{H}$?

Not all Kernels are Useful

- $k(x, y) = c$ for all $x, y \in X$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0, \ \forall \, \mathbb{P}, \, \mathbb{Q}.$$

- Another example: $k(x, y) = \langle x, y \rangle_{\mathbb{R}^d}, \ \ x, y \in \mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|M_{\mathbb{P}} - M_{\mathbb{Q}}\|_{\mathbb{R}^d},$$

where $M_{\mathbb{P}}$ is the mean of $\mathbb{P}$.

- Separable distributions can be made inseparable if the RKHS is not chosen properly.

How to choose $\mathcal{H}$?

Not all Kernels are Useful

- $k(x, y) = c$ for all $x, y \in X$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0, \ \forall \, \mathbb{P}, \ \mathbb{Q}.$$

- Another example: $k(x, y) = \langle x, y \rangle_{\mathbb{R}^d}, \ \ x, y \in \mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|M_{\mathbb{P}} - M_{\mathbb{Q}}\|_{\mathbb{R}^d},$$

where $M_{\mathbb{P}}$ is the mean of $\mathbb{P}$.

- Separable distributions can be made inseparable if the RKHS is not chosen properly.

How to choose $\mathcal{H}$?

Computation: RKHS vs. Other $\mathcal{F}$

- Suppose $\{X_1, \ldots, X_m\} \overset{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} \mathbb{Q}$.
- Define $\mathbb{P}_m := \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$ and $\mathbb{Q}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, where δ_x represents the Dirac measure at x.
- $MMD(\mathbb{P}_m, \mathbb{Q}_n, \{\|f\|_{\mathcal{H}} \leq 1\})$ is obtained in a closed form as:

$$
\begin{aligned}
MMD^2(\mathbb{P}_m, \mathbb{Q}_n, \{\|f\|_{\mathcal{H}} \leq 1\}) \quad = \quad & \frac{1}{m^2}\sum_{i,j=1}^m k(X_i, X_j) + \frac{1}{n^2}\sum_{i,j=1}^n k(Y_i, Y_j) \\
& - \frac{2}{mn}\sum_{i,j} k(X_i, Y_j).
\end{aligned}
$$

Very easy to compute!!

Computation: RKHS vs. Other $\mathcal{F}$

- $MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F})$ is obtained by solving a linear program for $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls. [Sriperumbudur et al., 2010a]

- Define $Z_i = X_i$ for $i = 1, \ldots, m$ and $Z_{m+i} = Y_i$ for $i = 1, \ldots, n$. Let ρ be a metric on X.

- $MMD(\mathbb{P}_m, \mathbb{Q}_n, \{\|f\|_L \leq 1\}) = \frac{1}{m}\sum_{i=1}^{m} a_i^\star - \frac{1}{n}\sum_{i=m+1}^{m+n} a_i^\star$, and $\{a_i^\star\}_{i=1}^{m+n}$ solve the following linear program,

$$\max_{a_1,\ldots,a_{m+n}} \left\{ \frac{1}{m}\sum_{i=1}^{m} a_i - \frac{1}{n}\sum_{i=m+1}^{m+n} a_i \,:\, -\rho(Z_i, Z_j) \leq a_i - a_j \leq \rho(Z_i, Z_j), \forall\, i, j \right\}.$$

More complex than with RKHS!!

Computation: RKHS vs. Other $\mathcal{F}$

- $MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F})$ is obtained by solving a linear program for $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls. [Sriperumbudur et al., 2010a]
- Define $Z_i = X_i$ for $i = 1, \ldots, m$ and $Z_{m+i} = Y_i$ for $i = 1, \ldots, n$. Let ρ be a metric on X.
- $MMD(\mathbb{P}_m, \mathbb{Q}_n, \{\|f\|_L \leq 1\}) = \frac{1}{m}\sum_{i=1}^{m} a_i^\star - \frac{1}{n}\sum_{i=m+1}^{m+n} a_i^\star$, and $\{a_i^\star\}_{i=1}^{m+n}$ solve the following linear program,

$$\max_{a_1,\ldots,a_{m+n}} \left\{ \frac{1}{m}\sum_{i=1}^{m} a_i - \frac{1}{n}\sum_{i=m+1}^{m+n} a_i \;:\; -\rho(Z_i, Z_j) \leq a_i - a_j \leq \rho(Z_i, Z_j), \forall\, i, j \right\}.$$

More complex than with RKHS!!

Computation: RKHS vs. Other $\mathcal{F}$

- $MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F})$ is obtained by solving a linear program for $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls. [Sriperumbudur et al., 2010a]

- Define $Z_i = X_i$ for $i = 1, \ldots, m$ and $Z_{m+i} = Y_i$ for $i = 1, \ldots, n$. Let ρ be a metric on X.

- $MMD(\mathbb{P}_m, \mathbb{Q}_n, \{\|f\|_L + \|f\|_\infty \leq 1\}) = \frac{1}{m}\sum_{i=1}^m a_i^\star - \frac{1}{n}\sum_{i=m+1}^{m+n} a_i^\star$, and $\{a_i^\star\}_{i=1}^{m+n}$ solve the following linear program,

$$
\begin{aligned}
\max_{a_1,\ldots,a_{m+n},b,c} \quad & \frac{1}{m}\sum_{i=1}^{m} a_i - \frac{1}{n}\sum_{i=m+1}^{m+n} a_i \\
\text{s.t.} \quad & -b\,\rho(Z_i, Z_j) \leq a_i - a_j \leq b\,\rho(Z_i, Z_j), \ \forall\, i, j \\
& -c \leq a_i \leq c, \ \forall\, i, \ \ b + c \leq 1.
\end{aligned}
$$

More complex than with RKHS!!

Computation: RKHS vs. Other $\mathcal{F}$

- $MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F})$ is obtained by solving a linear program for $\mathcal{F}=$ Lipschitz and bounded Lipschitz balls. [Sriperumbudur et al., 2010a]

- Define $Z_i = X_i$ for $i = 1, \ldots, m$ and $Z_{m+i} = Y_i$ for $i = 1, \ldots, n$. Let ρ be a metric on X.

- $MMD(\mathbb{P}_m, \mathbb{Q}_n, \{\|f\|_L + \|f\|_\infty \leq 1\}) = \frac{1}{m}\sum_{i=1}^m a_i^\star - \frac{1}{n}\sum_{i=m+1}^{m+n} a_i^\star$, and $\{a_i^\star\}_{i=1}^{m+n}$ solve the following linear program,

$$
\begin{aligned}
\max_{a_1,\ldots,a_{m+n},b,c} \quad & \frac{1}{m}\sum_{i=1}^{m} a_i - \frac{1}{n}\sum_{i=m+1}^{m+n} a_i \\
\text{s.t.} \quad & -b\,\rho(Z_i, Z_j) \leq a_i - a_j \leq b\,\rho(Z_i, Z_j), \forall\, i, j \\
& -c \leq a_i \leq c, \forall\, i, \;\; b + c \leq 1.
\end{aligned}
$$

More complex than with RKHS!!

Error: RKHS vs. Other $\mathcal{F}$

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| =?$$

- RKHS: [Gretton et al., 2007]

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \to 0, \; m, n \to \infty$$

  There exists $C > 0$ (independent of m and n) such that

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \leq C\sqrt{\frac{m+n}{mn}}$$

- Lipschitz and Bounded Lipschitz on $\mathbb{R}^d$: [Sriperumbudur et al., 2010a]

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \to 0, \; m, n \to \infty$$

  There exists $C > 0$ (independent of m and n) such that

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \leq C\left(\frac{m+n}{mn}\right)^{\frac{1}{d+1}}$$

Curse of dimensionality!!

Error: RKHS vs. Other $\mathcal{F}$

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| = ?$$

- RKHS: [Gretton et al., 2007]

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \to 0, \ \ m, n \to \infty$$

There exists $C > 0$ (independent of m and n) such that

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \le C\sqrt{\frac{m+n}{mn}}$$

- Lipschitz and Bounded Lipschitz on $\mathbb{R}^d$: [Sriperumbudur et al., 2010a]

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \to 0, \ \ m, n \to \infty$$

There exists $C > 0$ (independent of m and n) such that

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \le C\left(\frac{m+n}{mn}\right)^{\frac{1}{d+1}}$$

Curse of dimensionality!!

Error: RKHS vs. Other $\mathcal{F}$

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| = ?$$

- RKHS: [Gretton et al., 2007]

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \to 0, \;\; m, n \to \infty$$

There exists $C > 0$ (independent of m and n) such that

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \le C\sqrt{\frac{m+n}{mn}}$$

- Lipschitz and Bounded Lipschitz on $\mathbb{R}^d$: [Sriperumbudur et al., 2010a]

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \to 0, \;\; m, n \to \infty$$

There exists $C > 0$ (independent of m and n) such that

$$|MMD(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F}) - MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F})| \le C\left(\frac{m+n}{mn}\right)^{\frac{1}{d+1}}$$

Curse of dimensionality!!

How to choose $\mathcal{H}$?

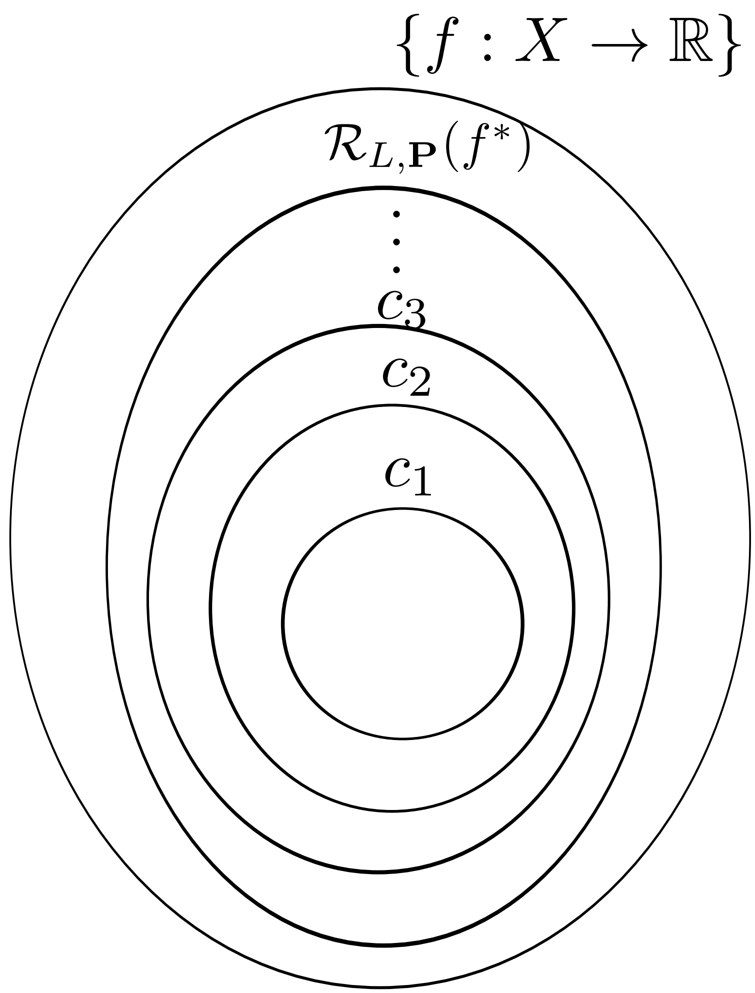
Large RKHS: Universal Kernel/RKHS

- Universal kernel: A kernel k on a compact metric space, X is said to be universal if the RKHS, $\mathcal{H}$ is dense (w.r.t. uniform norm) in $C(X)$.
- [Steinwart and Christmann, 2008]: For certain conditions on L, if k is universal, then

$$\inf_{f \in \mathcal{H}} \mathcal{R}_{L,\mathbf{P}}(f) = \mathcal{R}_{L,\mathbf{P}}(f^*)$$

  - Squared loss, Hinge loss,...

Large RKHS



Strictly Positive Definite Kernels

A symmetric function $k : X \times X \to \mathbb{R}$ is positive definite if $\forall\, n \geq 1$, $\forall\,(a_1, \ldots, a_n) \in \mathbb{R}^n$, $\forall\,(x_1, \ldots, x_n) \in X^n$,

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

k is strictly positive definite if for mutually distinct x_i, the equality holds only when all the a_i are zero.

Stronger than Strictly Positive Definite Kernels

- $M_b(X)$ = set of finite signed measure on X.

[Sriperumbudur et al., 2010b]: k is universal if and only if

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

is injective, i.e.,

$$\int_X k(\cdot, x)\, d\mu(x) = 0 \Rightarrow \mu = 0$$

which is equivalent to

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

Generalization of strictly positive definite kernels

Stronger than Strictly Positive Definite Kernels

- $M_b(X)$ = set of finite signed measure on X.

[Sriperumbudur et al., 2010b]: k is universal if and only if

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

is injective, i.e.,

$$\int_X k(\cdot, x)\, d\mu(x) = 0 \Rightarrow \mu = 0$$

which is equivalent to

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

Generalization of strictly positive definite kernels

Stronger than Strictly Positive Definite Kernels

- $M_b(X)$ = set of finite signed measure on X.

[Sriperumbudur et al., 2010b]: k is universal if and only if

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

is injective, i.e.,

$$\int_X k(\cdot, x)\, d\mu(x) = 0 \Rightarrow \mu = 0$$

which is equivalent to

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

Generalization of strictly positive definite kernels

Why Useful?

- Denseness characterization is not easy to check
- In general, though

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

is also not easy to check, for certain X and for certain families of k, the above condition is easy to check

- Later: Gaussian and Spline kernels are universal; Sinc kernel is not but is strictly positive definite.

MMD: What Kernels are Useful?

- Note that

$$MMD^2(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \int_X \int_X k(x, y)\, d(\mathbb{P} - \mathbb{Q})(x)\, d(\mathbb{P} - \mathbb{Q})(y)$$

- If k is universal, which means

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

  then

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q} \;\; (characteristic)$$

- In other words, universal kernel $\Rightarrow$ characteristic kernel

When is a Kernel Universal?

- [Sriperumbudur et al., 2010b]: The notion of universality can be generalized to non-compact X and we define bounded k to be universal if

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

- Nice characterization can be obtained if k is a bounded continuous translation invariant kernel on $\mathbb{R}^d$, i.e.,

$$k(x,y) = \psi(x-y)$$

  - Examples: Gaussian, $e^{-\|x-y\|_2^2}$, Laplacian, $e^{-\|x-y\|_1}$

- Bochner's Theorem: ψ is positive definite if and only it is the Fourier transform of a non-negative finite Borel measure, Λ,

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

When is a Kernel Universal?

- [Sriperumbudur et al., 2010b]: The notion of universality can be generalized to non-compact X and we define bounded k to be universal if

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

- Nice characterization can be obtained if k is a bounded continuous translation invariant kernel on $\mathbb{R}^d$, i.e.,

$$k(x,y) = \psi(x-y)$$

  - Examples: Gaussian, $e^{-\|x-y\|_2^2}$, Laplacian, $e^{-\|x-y\|_1}$

- Bochner's Theorem: ψ is positive definite if and only it is the Fourier transform of a non-negative finite Borel measure, Λ,

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

When is a Kernel Universal?

- [Sriperumbudur et al., 2010b]: The notion of universality can be generalized to non-compact X and we define bounded k to be universal if

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

- Nice characterization can be obtained if k is a bounded continuous translation invariant kernel on $\mathbb{R}^d$, i.e.,

$$k(x,y) = \psi(x-y)$$

  - Examples: Gaussian, $e^{-\|x-y\|_2^2}$, Laplacian, $e^{-\|x-y\|_1}$

- Bochner's Theorem: ψ is positive definite if and only it is the Fourier transform of a non-negative finite Borel measure, Λ,

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

When is a Kernel Universal?

- [Sriperumbudur et al., 2010b]: The notion of universality can be generalized to non-compact X and we define bounded k to be universal if

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

- Nice characterization can be obtained if k is a bounded continuous translation invariant kernel on $\mathbb{R}^d$, i.e.,

$$k(x,y) = \psi(x-y)$$

  - Examples: Gaussian, $e^{-\|x-y\|_2^2}$, Laplacian, $e^{-\|x-y\|_1}$

- Bochner's Theorem: ψ is positive definite if and only it is the Fourier transform of a non-negative finite Borel measure, Λ,

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

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

[Sriperumbudur et al., 2010c, Sriperumbudur et al., 2010b]:

Result: universal $\Leftrightarrow$ characteristic $\Leftrightarrow$ support of Λ is $\mathbb{R}^d$

- Support of a function, f is $\overline{\{x \in X : f(x) \neq 0\}}$

Proof: support of Λ is $\mathbb{R}^d$ $\Rightarrow$ universal $\Rightarrow$ characteristic

$$\begin{aligned}
\int\int_{\mathbb{R}^d} k(x,y)\, d\mu(x)\, d\mu(y) &= \int\int\int_{\mathbb{R}^d} e^{-\sqrt{-1}(x-y)^T\omega}\, d\Lambda(\omega)\, d\mu(x)\, d\mu(y) \\
&= \int\int_{\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega}\, d\mu(x) \int_{\mathbb{R}^d} e^{\sqrt{-1}y^T\omega}\, d\mu(y)\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\hat{\mu}(\omega)}\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\, d\Lambda(\omega).
\end{aligned}$$

If the support of Λ is $\mathbb{R}^d$, then $\int\int_{\mathbb{R}^d} \psi(x-y)\, d\mu(x)\, d\mu(y) = 0$ implies $\hat{\mu} = 0$ and therefore $\mu = 0$.

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

[Sriperumbudur et al., 2010c, Sriperumbudur et al., 2010b]:

Result: universal $\Leftrightarrow$ characteristic $\Leftrightarrow$ support of Λ is $\mathbb{R}^d$

- Support of a function, f is $\overline{\{x \in X : f(x) \neq 0\}}$

Proof: support of Λ is $\mathbb{R}^d \Rightarrow$ universal $\Rightarrow$ characteristic

$$\begin{aligned}
\int\int_{\mathbb{R}^d} k(x,y)\,d\mu(x)\,d\mu(y) &= \int\int\int_{\mathbb{R}^d} e^{-\sqrt{-1}(x-y)^T\omega}\,d\Lambda(\omega)\,d\mu(x)\,d\mu(y) \\
&= \int\int_{\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega}\,d\mu(x)\int_{\mathbb{R}^d} e^{\sqrt{-1}y^T\omega}\,d\mu(y)\,d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\hat{\mu}(\omega)}\,d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\,d\Lambda(\omega).
\end{aligned}$$

If the support of Λ is $\mathbb{R}^d$, then $\int\int_{\mathbb{R}^d} \psi(x-y)\,d\mu(x)\,d\mu(y) = 0$ implies $\hat{\mu} = 0$ and therefore $\mu = 0$.

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

[Sriperumbudur et al., 2010c, Sriperumbudur et al., 2010b]:

Result: universal $\Leftrightarrow$ characteristic $\Leftrightarrow$ support of Λ is $\mathbb{R}^d$

- Support of a function, f is $\overline{\{x \in X : f(x) \neq 0\}}$

Proof: support of Λ is $\mathbb{R}^d$ $\Rightarrow$ universal $\Rightarrow$ characteristic

$$\begin{aligned}
\int\int_{\mathbb{R}^d} k(x,y)\, d\mu(x)\, d\mu(y) &= \int\int\int_{\mathbb{R}^d} e^{-\sqrt{-1}(x-y)^T\omega}\, d\Lambda(\omega)\, d\mu(x)\, d\mu(y) \\
&= \int\int_{\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega}\, d\mu(x) \int_{\mathbb{R}^d} e^{\sqrt{-1}y^T\omega}\, d\mu(y)\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\hat{\mu}(\omega)}\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\, d\Lambda(\omega).
\end{aligned}$$

If the support of Λ is $\mathbb{R}^d$, then $\int\int_{\mathbb{R}^d} \psi(x-y)\, d\mu(x)\, d\mu(y) = 0$ implies $\hat{\mu} = 0$ and therefore $\mu = 0$.

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

[Sriperumbudur et al., 2010c, Sriperumbudur et al., 2010b]:

Result: universal $\Leftrightarrow$ characteristic $\Leftrightarrow$ support of Λ is $\mathbb{R}^d$

- Support of a function, f is $\overline{\{x \in X : f(x) \neq 0\}}$

Proof: support of Λ is $\mathbb{R}^d$ $\Rightarrow$ universal $\Rightarrow$ characteristic

$$\begin{aligned}
\int\int_{\mathbb{R}^d} k(x,y)\, d\mu(x)\, d\mu(y) &= \int\int\int_{\mathbb{R}^d} e^{-\sqrt{-1}(x-y)^T\omega}\, d\Lambda(\omega)\, d\mu(x)\, d\mu(y) \\
&= \int\int_{\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega}\, d\mu(x) \int_{\mathbb{R}^d} e^{\sqrt{-1}y^T\omega}\, d\mu(y)\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\hat{\mu}(\omega)}\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\, d\Lambda(\omega).
\end{aligned}$$

If the support of Λ is $\mathbb{R}^d$, then $\int\int_{\mathbb{R}^d} \psi(x-y)\, d\mu(x)\, d\mu(y) = 0$ implies $\hat{\mu} = 0$ and therefore $\mu = 0$.

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

[Sriperumbudur et al., 2010c, Sriperumbudur et al., 2010b]:

Result: universal $\Leftrightarrow$ characteristic $\Leftrightarrow$ support of Λ is $\mathbb{R}^d$

- Support of a function, f is $\overline{\{x \in X : f(x) \neq 0\}}$

Proof: support of Λ is $\mathbb{R}^d$ $\Rightarrow$ universal $\Rightarrow$ characteristic

$$\begin{aligned}
\iint_{\mathbb{R}^d} k(x,y)\,d\mu(x)\,d\mu(y) &= \iiint_{\mathbb{R}^d} e^{-\sqrt{-1}(x-y)^T\omega}\,d\Lambda(\omega)\,d\mu(x)\,d\mu(y) \\
&= \iint_{\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega}\,d\mu(x) \int_{\mathbb{R}^d} e^{\sqrt{-1}y^T\omega}\,d\mu(y)\,d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\hat{\mu}(\omega)}\,d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\,d\Lambda(\omega).
\end{aligned}$$

If the support of Λ is $\mathbb{R}^d$, then $\iint_{\mathbb{R}^d} \psi(x-y)\,d\mu(x)\,d\mu(y) = 0$ implies $\hat{\mu} = 0$ and therefore $\mu = 0$.

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

[Sriperumbudur et al., 2010c, Sriperumbudur et al., 2010b]:

Result: universal $\Leftrightarrow$ characteristic $\Leftrightarrow$ support of Λ is $\mathbb{R}^d$

- Support of a function, f is $\overline{\{x \in X : f(x) \neq 0\}}$

Proof: support of Λ is $\mathbb{R}^d$ $\Rightarrow$ universal $\Rightarrow$ characteristic

$$\begin{aligned}
\iint_{\mathbb{R}^d} k(x,y)\,d\mu(x)\,d\mu(y) &= \iiint_{\mathbb{R}^d} e^{-\sqrt{-1}(x-y)^T\omega}\,d\Lambda(\omega)\,d\mu(x)\,d\mu(y) \\
&= \iint_{\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega}\,d\mu(x)\int_{\mathbb{R}^d} e^{\sqrt{-1}y^T\omega}\,d\mu(y)\,d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\hat{\mu}(\omega)}\,d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\,d\Lambda(\omega).
\end{aligned}$$

If the support of Λ is $\mathbb{R}^d$, then $\iint_{\mathbb{R}^d} \psi(x-y)\,d\mu(x)\,d\mu(y) = 0$ implies $\hat{\mu} = 0$ and therefore $\mu = 0$.

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

[Sriperumbudur et al., 2010c, Sriperumbudur et al., 2010b]:

Result: universal $\Leftrightarrow$ characteristic $\Leftrightarrow$ support of Λ is $\mathbb{R}^d$

- Support of a function, f is $\overline{\{x \in X : f(x) \neq 0\}}$

Proof: support of Λ is $\mathbb{R}^d$ $\Rightarrow$ universal $\Rightarrow$ characteristic

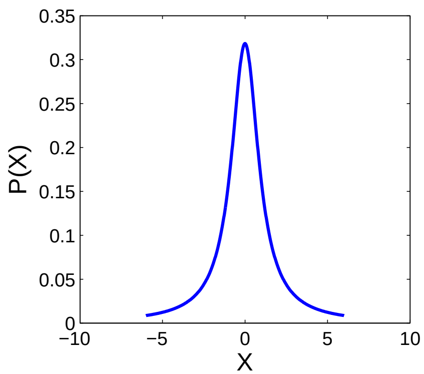
$$
\begin{aligned}
\iint_{\mathbb{R}^d} k(x,y)\, d\mu(x)\, d\mu(y) &= \iiint_{\mathbb{R}^d} e^{-\sqrt{-1}(x-y)^T\omega}\, d\Lambda(\omega)\, d\mu(x)\, d\mu(y) \\
&= \iint_{\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega}\, d\mu(x) \int_{\mathbb{R}^d} e^{\sqrt{-1}y^T\omega}\, d\mu(y)\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\hat{\mu}(\omega)}\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\, d\Lambda(\omega).
\end{aligned}
$$

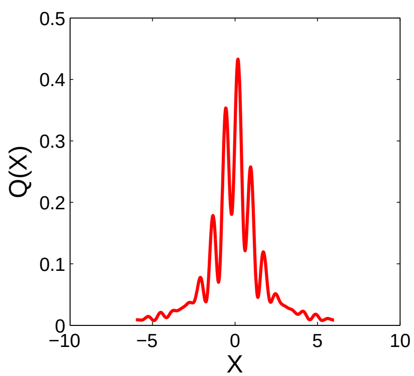
If the support of Λ is $\mathbb{R}^d$, then $\iint_{\mathbb{R}^d} \psi(x-y)\, d\mu(x)\, d\mu(y) = 0$ implies $\hat{\mu} = 0$ and therefore $\mu = 0$.

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

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency

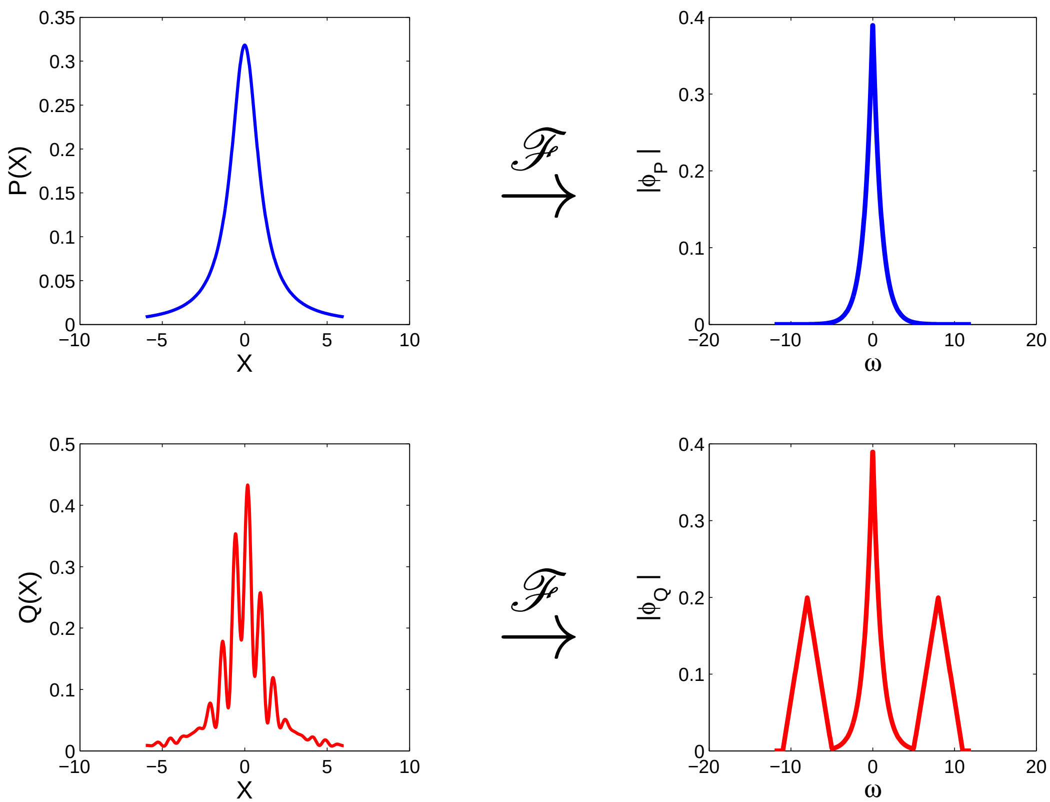




Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathfrak{F}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

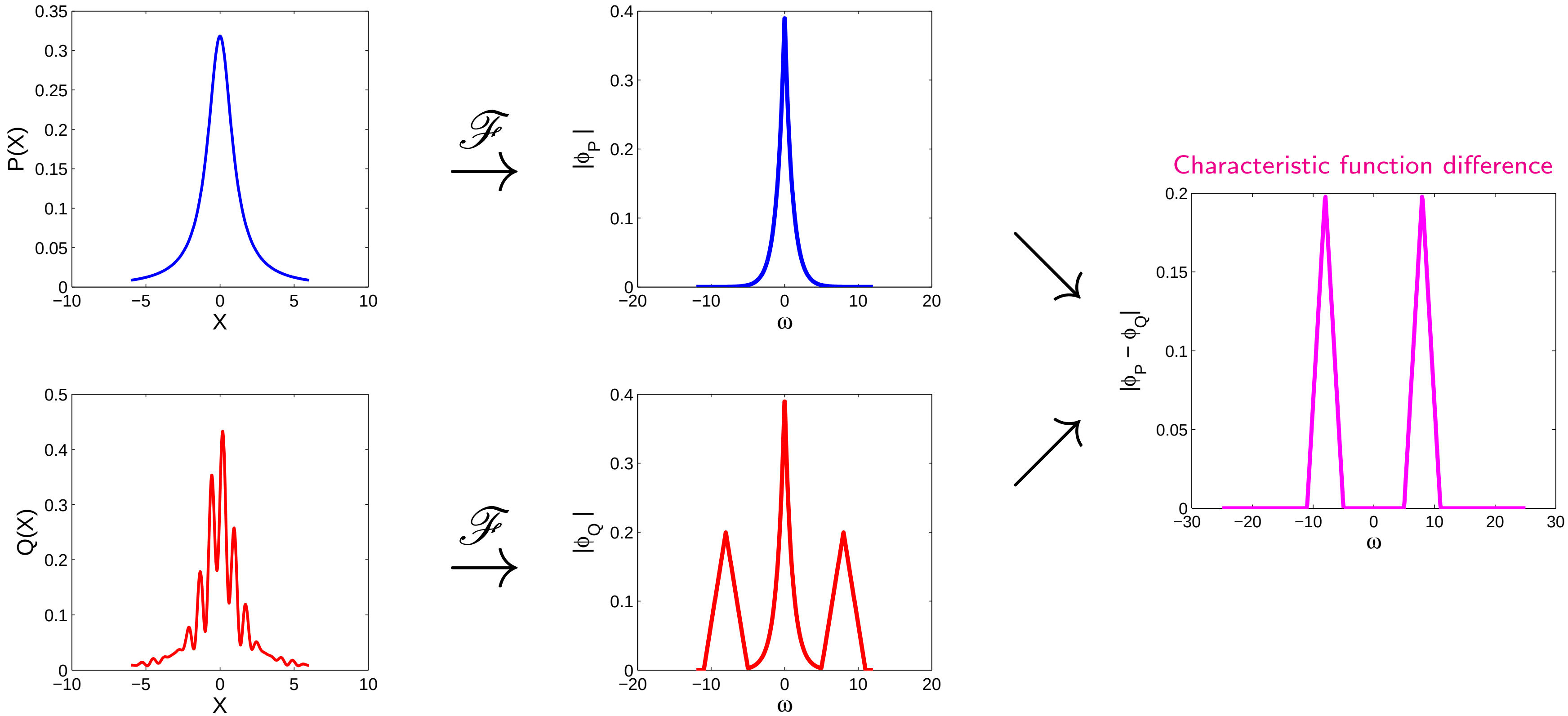
- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency



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

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency



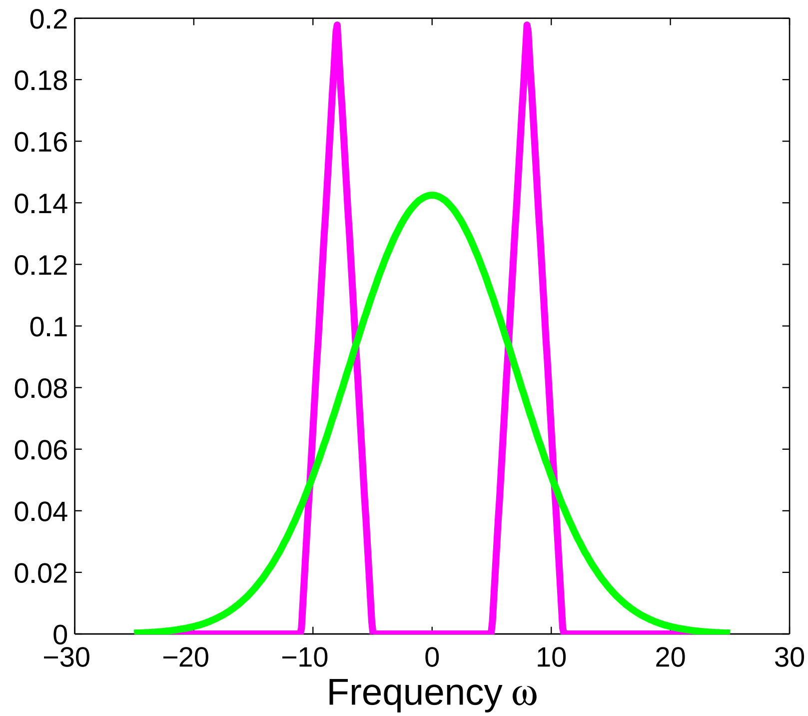
Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency

Gaussian kernel

$|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}|$

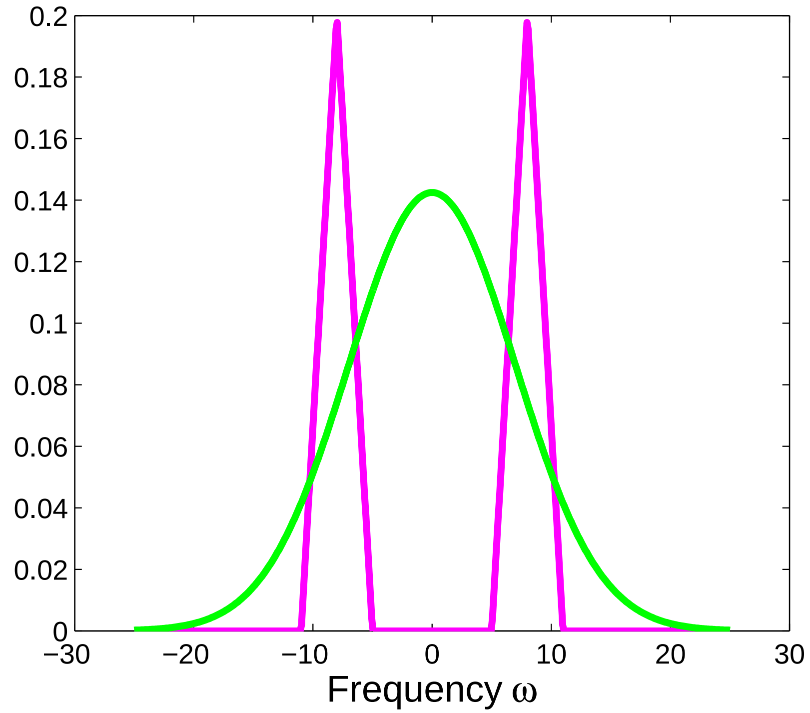


Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency

Universal



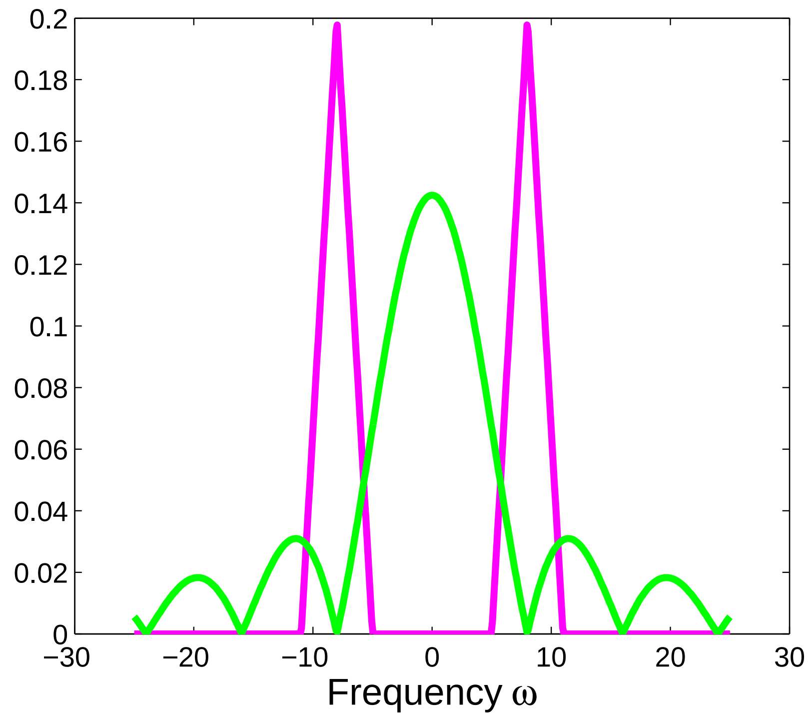
Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency

B-Spline kernel

$|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}|$

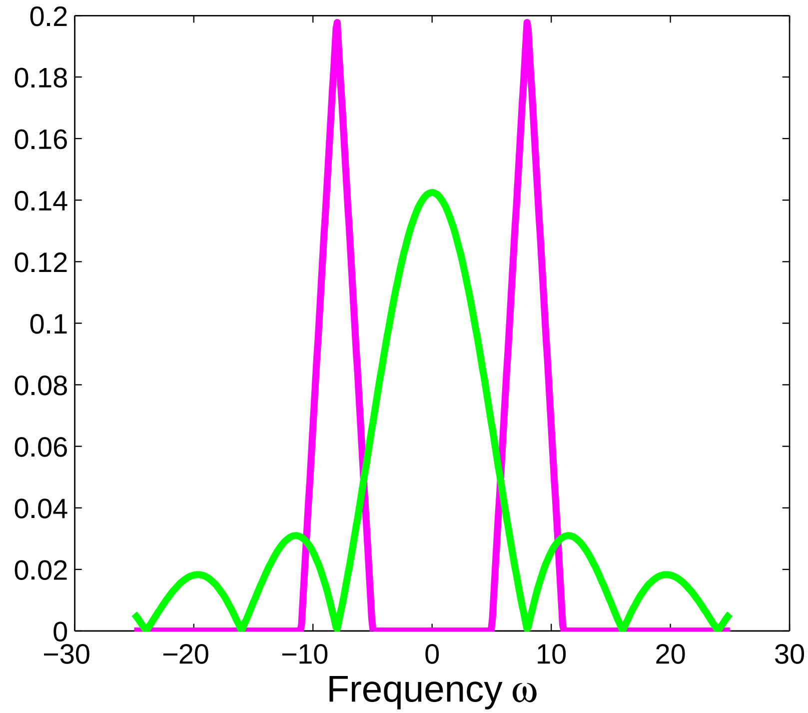


Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency

???

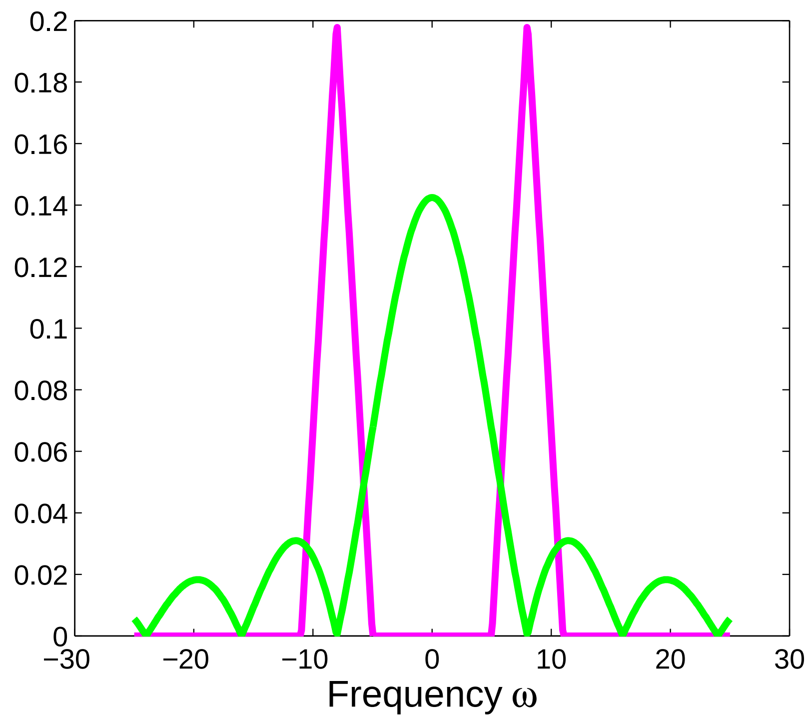


Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency

Universal



Proof Idea of the Converse

- $\text{supp}(\Lambda) = \mathbb{R}^d \Rightarrow$ universal $\Rightarrow$ characteristic
- If we show that characteristic $\Rightarrow \text{supp}(\Lambda) = \mathbb{R}^d$, then we are DONE.
- Equivalently, we need to show that if the support of Λ is NOT $\mathbb{R}^d$, then $\exists\, \mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \{\|f\|_{\mathcal{H}} \leq 1\}) = 0$

Proof Idea of the Converse

- $\text{supp}(\Lambda) = \mathbb{R}^d \Rightarrow$ universal $\Rightarrow$ characteristic
- If we show that characteristic $\Rightarrow \text{supp}(\Lambda) = \mathbb{R}^d$, then we are DONE.
- Equivalently, we need to show that if the support of Λ is NOT $\mathbb{R}^d$, then $\exists\, \mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \{\|f\|_{\mathcal{H}} \leq 1\}) = 0$

Proof Idea of the Converse

- $\text{supp}(\Lambda) = \mathbb{R}^d \Rightarrow$ universal $\Rightarrow$ characteristic
- If we show that characteristic $\Rightarrow \text{supp}(\Lambda) = \mathbb{R}^d$, then we are DONE.
- Equivalently, we need to show that if the support of Λ is NOT $\mathbb{R}^d$, then $\exists\, \mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \{\|f\|_{\mathcal{H}} \leq 1\}) = 0$

Proof

- Suppose support of Λ is NOT $\mathbb{R}^d$.
- Then there exists an open set, $U \subset \mathbb{R}^d \backslash \mathrm{supp}(\Lambda)$.
- Construct a non-zero real-valued symmetric function, θ supported on U with $\theta(0) = 0$.
- Define $d\mu(x) = \hat{\theta}(x)\, dx$ where $\hat{\theta}$ is the Fourier transform of θ.
- Also $\mu(\mathbb{R}^d) = 0$.
- There exists positive measures $\mu^+ \neq \mu^-$ such that $\mu = \mu^+ - \mu^-$ (Jordan decomposition)
- Define $\alpha := \mu^+(\mathbb{R}^d)$, $\mathbb{P} := \alpha^{-1}\mu^+$ and $\mathbb{Q} := \alpha^{-1}\mu^-$
- Clearly $\phi_{\mathbb{P}} - \phi_{\mathbb{Q}} = \alpha^{-1}\theta$ which is NOT supported on $\mathrm{supp}(\Lambda)$
- Therefore, there exits $\mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0$ $\square$

Proof

- Suppose support of Λ is NOT $\mathbb{R}^d$.
- Then there exists an open set, $U \subset \mathbb{R}^d \backslash \mathrm{supp}(\Lambda)$.
- Construct a non-zero real-valued symmetric function, θ supported on U with $\theta(0) = 0$.
- Define $d\mu(x) = \hat{\theta}(x)\,dx$ where $\hat{\theta}$ is the Fourier transform of θ.
- Also $\mu(\mathbb{R}^d) = 0$.
- There exists positive measures $\mu^+ \neq \mu^-$ such that $\mu = \mu^+ - \mu^-$ (Jordan decomposition)
- Define $\alpha := \mu^+(\mathbb{R}^d)$, $\mathbb{P} := \alpha^{-1}\mu^+$ and $\mathbb{Q} := \alpha^{-1}\mu^-$
- Clearly $\phi_{\mathbb{P}} - \phi_{\mathbb{Q}} = \alpha^{-1}\theta$ which is NOT supported on supp(Λ)
- Therefore, there exits $\mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0$ $\square$

Proof

- Suppose support of Λ is NOT $\mathbb{R}^d$.
- Then there exists an open set, $U \subset \mathbb{R}^d \backslash \text{supp}(\Lambda)$.
- Construct a non-zero real-valued symmetric function, θ supported on U with $\theta(0) = 0$.
- Define $d\mu(x) = \hat{\theta}(x)\, dx$ where $\hat{\theta}$ is the Fourier transform of θ.
- Also $\mu(\mathbb{R}^d) = 0$.
- There exists positive measures $\mu^+ \neq \mu^-$ such that $\mu = \mu^+ - \mu^-$ (Jordan decomposition)
- Define $\alpha := \mu^+(\mathbb{R}^d)$, $\mathbb{P} := \alpha^{-1}\mu^+$ and $\mathbb{Q} := \alpha^{-1}\mu^-$
- Clearly $\phi_{\mathbb{P}} - \phi_{\mathbb{Q}} = \alpha^{-1}\theta$ which is NOT supported on supp(Λ)
- Therefore, there exits $\mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0$ $\square$

Proof

- Suppose support of Λ is NOT $\mathbb{R}^d$.
- Then there exists an open set, $U \subset \mathbb{R}^d \backslash \mathrm{supp}(\Lambda)$.
- Construct a non-zero real-valued symmetric function, θ supported on U with $\theta(0) = 0$.
- Define $d\mu(x) = \hat{\theta}(x)\, dx$ where $\hat{\theta}$ is the Fourier transform of θ.
- Also $\mu(\mathbb{R}^d) = 0$.
- There exists positive measures $\mu^+ \neq \mu^-$ such that $\mu = \mu^+ - \mu^-$ (Jordan decomposition)
- Define $\alpha := \mu^+(\mathbb{R}^d)$, $\mathbb{P} := \alpha^{-1}\mu^+$ and $\mathbb{Q} := \alpha^{-1}\mu^-$
- Clearly $\phi_{\mathbb{P}} - \phi_{\mathbb{Q}} = \alpha^{-1}\theta$ which is NOT supported on $\mathrm{supp}(\Lambda)$
- Therefore, there exits $\mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0$ $\square$

Proof

- Suppose support of Λ is NOT $\mathbb{R}^d$.
- Then there exists an open set, $U \subset \mathbb{R}^d \backslash \mathrm{supp}(\Lambda)$.
- Construct a non-zero real-valued symmetric function, θ supported on U with $\theta(0) = 0$.
- Define $d\mu(x) = \hat{\theta}(x)\, dx$ where $\hat{\theta}$ is the Fourier transform of θ.
- Also $\mu(\mathbb{R}^d) = 0$.
- There exists positive measures $\mu^+ \neq \mu^-$ such that $\mu = \mu^+ - \mu^-$ (Jordan decomposition)
- Define $\alpha := \mu^+(\mathbb{R}^d)$, $\mathbb{P} := \alpha^{-1}\mu^+$ and $\mathbb{Q} := \alpha^{-1}\mu^-$
- Clearly $\phi_{\mathbb{P}} - \phi_{\mathbb{Q}} = \alpha^{-1}\theta$ which is NOT supported on $\mathrm{supp}(\Lambda)$
- Therefore, there exits $\mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0$ $\square$

Proof

- Suppose support of Λ is NOT $\mathbb{R}^d$.
- Then there exists an open set, $U \subset \mathbb{R}^d \backslash \mathrm{supp}(\Lambda)$.
- Construct a non-zero real-valued symmetric function, θ supported on U with $\theta(0) = 0$.
- Define $d\mu(x) = \hat{\theta}(x)\, dx$ where $\hat{\theta}$ is the Fourier transform of θ.
- Also $\mu(\mathbb{R}^d) = 0$.
- There exists positive measures $\mu^+ \neq \mu^-$ such that $\mu = \mu^+ - \mu^-$ (Jordan decomposition)
- Define $\alpha := \mu^+(\mathbb{R}^d)$, $\mathbb{P} := \alpha^{-1}\mu^+$ and $\mathbb{Q} := \alpha^{-1}\mu^-$
- Clearly $\phi_{\mathbb{P}} - \phi_{\mathbb{Q}} = \alpha^{-1}\theta$ which is NOT supported on $\mathrm{supp}(\Lambda)$
- Therefore, there exits $\mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0$ $\square$

Proof

- Suppose support of Λ is NOT $\mathbb{R}^d$.
- Then there exists an open set, $U \subset \mathbb{R}^d \backslash \text{supp}(\Lambda)$.
- Construct a non-zero real-valued symmetric function, θ supported on U with $\theta(0) = 0$.
- Define $d\mu(x) = \hat{\theta}(x)\, dx$ where $\hat{\theta}$ is the Fourier transform of θ.
- Also $\mu(\mathbb{R}^d) = 0$.
- There exists positive measures $\mu^+ \neq \mu^-$ such that $\mu = \mu^+ - \mu^-$ (Jordan decomposition)
- Define $\alpha := \mu^+(\mathbb{R}^d)$, $\mathbb{P} := \alpha^{-1}\mu^+$ and $\mathbb{Q} := \alpha^{-1}\mu^-$
- Clearly $\phi_{\mathbb{P}} - \phi_{\mathbb{Q}} = \alpha^{-1}\theta$ which is NOT supported on $\text{supp}(\Lambda)$
- Therefore, there exits $\mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0$ $\square$

Proof

- Suppose support of Λ is NOT $\mathbb{R}^d$.
- Then there exists an open set, $U \subset \mathbb{R}^d \backslash \text{supp}(\Lambda)$.
- Construct a non-zero real-valued symmetric function, θ supported on U with $\theta(0) = 0$.
- Define $d\mu(x) = \hat{\theta}(x)\, dx$ where $\hat{\theta}$ is the Fourier transform of θ.
- Also $\mu(\mathbb{R}^d) = 0$.
- There exists positive measures $\mu^+ \neq \mu^-$ such that $\mu = \mu^+ - \mu^-$ (Jordan decomposition)
- Define $\alpha := \mu^+(\mathbb{R}^d)$, $\mathbb{P} := \alpha^{-1}\mu^+$ and $\mathbb{Q} := \alpha^{-1}\mu^-$
- Clearly $\phi_{\mathbb{P}} - \phi_{\mathbb{Q}} = \alpha^{-1}\theta$ which is NOT supported on $\text{supp}(\Lambda)$
- Therefore, there exits $\mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0$ $\square$

Proof

- Suppose support of Λ is NOT $\mathbb{R}^d$.
- Then there exists an open set, $U \subset \mathbb{R}^d \backslash \text{supp}(\Lambda)$.
- Construct a non-zero real-valued symmetric function, θ supported on U with $\theta(0) = 0$.
- Define $d\mu(x) = \hat{\theta}(x)\, dx$ where $\hat{\theta}$ is the Fourier transform of θ.
- Also $\mu(\mathbb{R}^d) = 0$.
- There exists positive measures $\mu^+ \neq \mu^-$ such that $\mu = \mu^+ - \mu^-$ (Jordan decomposition)
- Define $\alpha := \mu^+(\mathbb{R}^d)$, $\mathbb{P} := \alpha^{-1}\mu^+$ and $\mathbb{Q} := \alpha^{-1}\mu^-$
- Clearly $\phi_{\mathbb{P}} - \phi_{\mathbb{Q}} = \alpha^{-1}\theta$ which is NOT supported on supp(Λ)
- Therefore, there exits $\mathbb{P} \neq \mathbb{Q}$ such that $MMD(\mathbb{P}, \mathbb{Q}, \mathfrak{F}) = 0$ $\square$

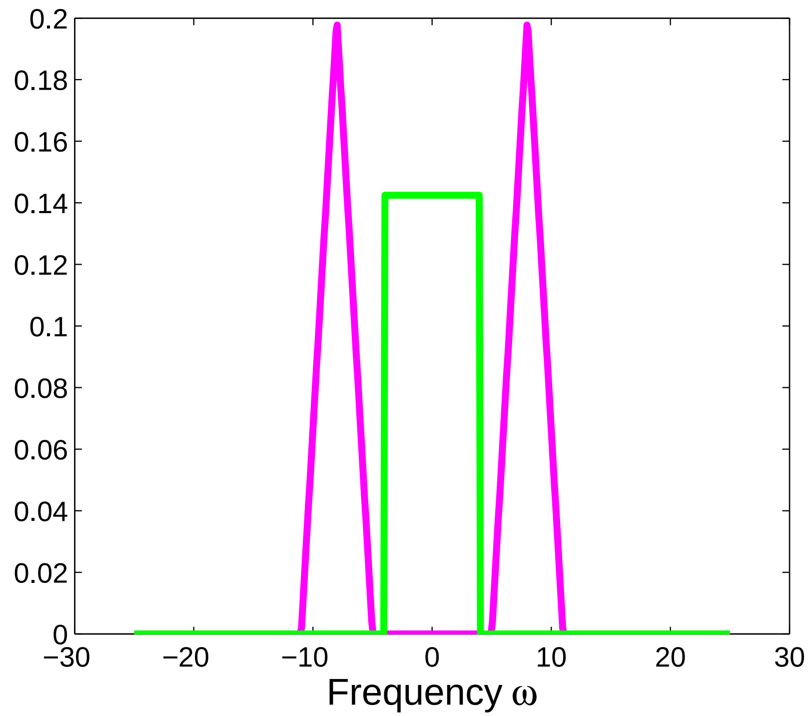
Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency

Sinc kernel

$|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}|$

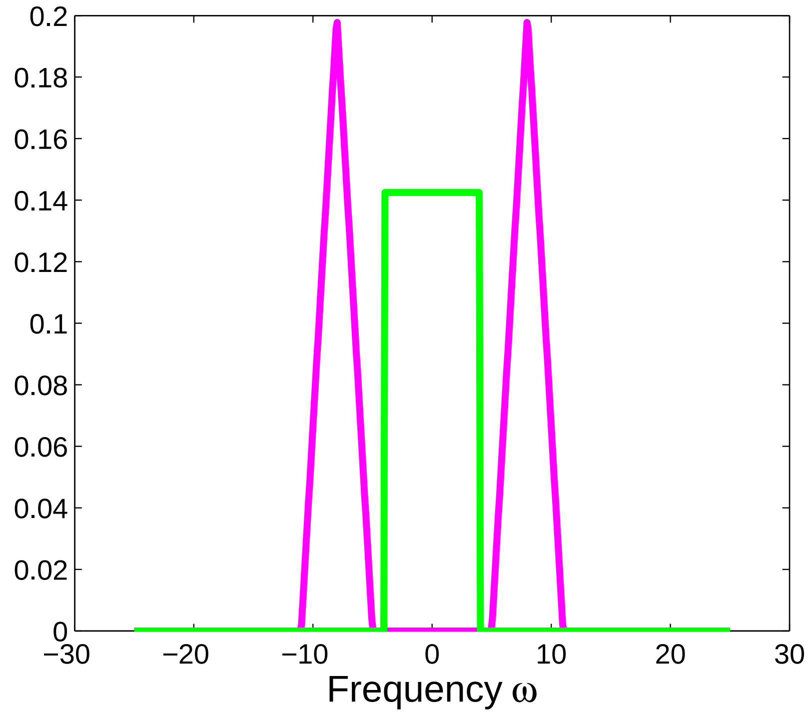


Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency

NOT universal



Summary

- Why RKHS?
  - Problem of learning
  - Loss function, Risk functional
  - Bayes risk and Bayes function
  - Empirical risk minimization
  - Approximation and estimation errors
  - RKHS allows great computational advantage
- How to choose an RKHS?
  - Universal RKHS that makes the approximation error to be zero.
  - Universal kernels generalize strictly positive definite kernels
  - Nice characterization for translation invariant kernels on $\mathbb{R}^d$.

References

- 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.
- Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2010a).
  Non-parametric estimation of integral probability metrics.
  In *Proc. IEEE International Symposium on Information Theory*, pages 1428–1432.
- Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. (2010b).
  On the relation between universality, characteristic kernels and RKHS embedding of measures.
  In Teh, Y. W. and Titterington, M., editors, *Proc. 13^{th} International Conference on Artificial Intelligence and Statistics*, volume 9 of *Workshop and Conference Proceedings*. JMLR.
- Sriperumbudur, B. K., Gretton, A., Fukumizu, K., chölkopf, B. S., and Lanckriet, G. R. G. (2010c).
  Hilbert space embeddings and metrics on probability measures.
  Journal of Machine Learning Research, 11:1517–1561.
- Steinwart, I. and Christmann, A. (2008).
  Support Vector Machines.
  Springer.