Lecture 2

Hilbert Space Embedding of Probability Measures

Bharath K. Sriperumbudur* & Dougal J. Sutherland†

*Department of Statistics, Pennsylvania State University

†Gatsby Unit, University College London

Data Science Summer School
École Polytechnique
June 2019

Recap of Lecture 1

Kernel method provides an elegant approach to achieve non-linear algorithms from linear algorithms.

- Input space, $\mathcal{X}$: the space of observed data on which learning is performed.
- Feature map, Φ: defined through a positive definite kernel function, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$x \mapsto \Phi(x), \qquad x \in \mathcal{X}$$

- Constructing linear algorithms in the feature space $\Phi(\mathcal{X})$ translates as non-linear algorithms in $\mathcal{X}$.
- Elegance: No explicit construction of Φ as $\langle \Phi(x), \Phi(y) \rangle = k(x, y)$.
- Function space view: RKHS; smoothness and generalization

Examples

- Ridge regression. In fact many more (Kernel+SVM/PCA/FDA/CCA/Perceptron/logistic regression, ...)

Outline

- Motivating example: Comparing distributions
- Hilbert space embedding of measures
  - Mean element
  - Distance on probabilities (MMD)
  - Characteristic kernels
  - Cross-covariance operator and measure of independence
- Applications
  - Two-sample testing
- Choice of kernel

Motivating Example: Coin Toss

- Toss 1: $T\ H\ H\ H\ T\ T\ H\ T\ T\ H\ H\ T\ H$
- Toss 2: $H\ T\ T\ H\ T\ H\ T\ T\ H\ H\ H\ T\ T$

Are the coins/tosses statistically similar?

Toss 1 is a sample from $\mathbb{P}$:=Bernoulli(p) and Toss 2 is a sample from $\mathbb{Q}$:=Bernoulli(q).

Is $p = q$ or not?, i.e., compare

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\{0,1\}} x\, d\mathbb{P}(x) \quad \text{and} \quad \mathbb{E}_{\mathbb{Q}}[X] = \int_{\{0,1\}} x\, d\mathbb{Q}(x).$$

Motivating Example: Coin Toss

- Toss 1: $T\ H\ H\ H\ T\ T\ H\ T\ T\ H\ H\ T\ H$
- Toss 2: $H\ T\ T\ H\ T\ H\ T\ T\ H\ H\ H\ T\ T$

Are the coins/tosses statistically similar?

Toss 1 is a sample from $\mathbb{P}$:=Bernoulli(p) and Toss 2 is a sample from $\mathbb{Q}$:=Bernoulli(q).

Is $p = q$ or not?, i.e., compare

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\{0,1\}} x \, d\mathbb{P}(x) \quad \text{and} \quad \mathbb{E}_{\mathbb{Q}}[X] = \int_{\{0,1\}} x \, d\mathbb{Q}(x).$$

Coin Toss Example

In other words, we compare

$$\int_{\mathbb{R}} \Phi(x)\, d\mathbb{P}(x) \quad \text{and} \quad \int_{\mathbb{R}} \Phi(x)\, d\mathbb{Q}(x)$$

where Φ is an identity map,

$$\Phi(x) = x.$$

A positive definite kernel corresponding to Φ is

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_2 = xy,$$

which is a linear kernel on $\{0, 1\}$. Therefore, comparing two Bernoulli is equivalent to

$$\int_{\{0,1\}} k(y, x)\, d\mathbb{P}(x) \stackrel{?}{=} \int_{\{0,1\}} k(y, x)\, d\mathbb{Q}(x)$$

for all $y \in \{0, 1\}$, i.e., compare the expectations of the kernel.

Comparing two Gaussians

$$\mathbb{P} = N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad \mathbb{Q} = N(\mu_2, \sigma_2^2)$$

Comparing $\mathbb{P}$ and $\mathbb{Q}$ is equivalent to comparing μ_1, μ_2 and σ_1^2, σ_2^2, i.e.,

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\mathbb{R}} x \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X]$$

and

$$\mathbb{E}_{\mathbb{P}}[X^2] = \int_{\mathbb{R}} x^2 \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x^2 \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X^2].$$

Concisely

$$\int_{\mathbb{R}} \Phi(x) \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} \Phi(x) \, d\mathbb{Q}(x)$$

where

$$\Phi(x) = (x, x^2).$$

Compare the first moment of the feature map

Comparing two Gaussians

$$\mathbb{P} = N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad \mathbb{Q} = N(\mu_2, \sigma_2^2)$$

Comparing $\mathbb{P}$ and $\mathbb{Q}$ is equivalent to comparing μ_1, μ_2 and σ_1^2, σ_2^2, i.e.,

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\mathbb{R}} x \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X]$$

and

$$\mathbb{E}_{\mathbb{P}}[X^2] = \int_{\mathbb{R}} x^2 \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x^2 \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X^2].$$

Concisely

$$\int_{\mathbb{R}} \Phi(x) \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} \Phi(x) \, d\mathbb{Q}(x)$$

where

$$\Phi(x) = (x, x^2).$$

Compare the first moment of the feature map

Comparing two Gaussians

Using the map Φ, we can construct a positive definite kernel as

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathbb{R}^2} = xy + x^2 y^2$$

which is a polynomial kernel of order 2.

Therefore, comparing two Gaussians is equivalent to

$$\int_{\mathbb{R}} k(y, x)\, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} k(y, x)\, d\mathbb{Q}(x)$$

for all $y \in \mathbb{R}$, i.e., compare the expectations of the kernel.

Comparing general $\mathbb{P}$ and $\mathbb{Q}$

Moment generating function is defined as

$$M_{\mathbb{P}}(y) = \int_{\mathbb{R}} e^{xy} \, d\mathbb{P}(x)$$

and (if it exists) captures the information about a distribution, i.e.,

$$M_{\mathbb{P}} = M_{\mathbb{Q}} \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Choosing

$$\Phi(x) = \left(1, x, \frac{x^2}{\sqrt{2!}}, \ldots, \frac{x^i}{\sqrt{i!}}, \ldots\right) \in \ell_2(\mathbb{N}), \forall\, x \in \mathbb{R}$$

it is easy to verify that

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\ell_2(\mathbb{N})} = e^{xy}$$

and so

$$\int_{\mathbb{R}} k(x, y) \, d\mathbb{P}(x) = \int_{\mathbb{R}} k(x, y) \, d\mathbb{Q}(x), \forall\, y \in \mathbb{R} \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Two-Sample Problem

- Given random samples $\{X_1, \ldots, X_m\} \stackrel{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_1, \ldots, Y_n\} \stackrel{i.i.d.}{\sim} \mathbb{Q}$.
- Determine: $\mathbb{P} = \mathbb{Q}$ or $\mathbb{P} \neq \mathbb{Q}$?

Applications:

- Microarray data (aggregation problem)
- Speaker verification
- Independence Testing: Given random samples $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \stackrel{i.i.d}{\sim} \mathbb{P}_{xy}$. Does $\mathbb{P}_{xy}$ factorize into $\mathbb{P}_x \mathbb{P}_y$?
- Feature selection (microarrays, image and text,...)

Hilbert Space Embedding of Measures

Hilbert Space Embedding of Measures

- Canonical feature map:

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

where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

- Generalization to probabilities:

$$x \mapsto k(\cdot, x) \qquad \equiv \qquad \underbrace{\delta_x}_{\text{point mass at } x} \mapsto \underbrace{k(\cdot, x)}_{\int_{\mathcal{X}} k(\cdot, y)\, d\delta_x(y) = \mathbb{E}_{\delta_x}[k(\cdot, Y)]}$$

Based on the above, the map is extended to probability measures as

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} := \int_{\mathcal{X}} \Phi(x)\, d\mathbb{P}(x) = \underbrace{\int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)}_{\mathbb{E}_{X \sim \mathbb{P}} k(\cdot, X)}$$

(Smola et al., ALT 2007)

Properties

- $\mu_{\mathbb{P}}$ is the mean of the feature map and is called the kernel mean or mean element of $\mathbb{P}$.

- When is $\mu_{\mathbb{P}}$ well defined?

$$\int_{\mathcal{X}} \sqrt{k(x,x)}\, d\mathbb{P}(x) < \infty \quad \Rightarrow \quad \mu_{\mathbb{P}} \in \mathcal{H}$$

Proof:

$$\|\mu_{\mathbb{P}}\|_{\mathcal{H}} = \left\| \int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{P}(x) \right\|_{\mathcal{H}} \stackrel{Jensen's}{\leq} \int_{\mathcal{X}} \|k(\cdot,x)\|_{\mathcal{H}}\, d\mathbb{P}(x).$$

- We know that for any $f \in \mathcal{H}$, $f(x) = \langle f, k(\cdot,x)\rangle_{\mathcal{H}}$. So, for any $f \in \mathcal{H}$,

$$\begin{aligned}\int_{\mathcal{X}} f(x)\, d\mathbb{P}(x) &= \int_{\mathcal{X}} \langle f, k(\cdot,x)\rangle_{\mathcal{H}}\, d\mathbb{P}(x) \stackrel{\bullet}{=} \left\langle f, \int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{P}(x) \right\rangle_{\mathcal{H}} \\ &= \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}.\end{aligned}$$

Properties

- $\mu_{\mathbb{P}}$ is the mean of the feature map and is called the kernel mean or mean element of $\mathbb{P}$.

- When is $\mu_{\mathbb{P}}$ well defined?

$$\int_{\mathcal{X}} \sqrt{k(x,x)}\, d\mathbb{P}(x) < \infty \quad \Rightarrow \quad \mu_{\mathbb{P}} \in \mathcal{H}$$

Proof:

$$\|\mu_{\mathbb{P}}\|_{\mathcal{H}} = \left\| \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) \right\|_{\mathcal{H}} \overset{Jensen's}{\leq} \int_{\mathcal{X}} \|k(\cdot, x)\|_{\mathcal{H}}\, d\mathbb{P}(x).$$

- We know that for any $f \in \mathcal{H}$, $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$. So, for any $f \in \mathcal{H}$,

$$\int_{\mathcal{X}} f(x)\, d\mathbb{P}(x) = \int_{\mathcal{X}} \langle f, k(\cdot, x)\rangle_{\mathcal{H}}\, d\mathbb{P}(x) \doteq \left\langle f, \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) \right\rangle_{\mathcal{H}}$$

$$= \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}.$$

Properties

- $\mu_{\mathbb{P}}$ is the mean of the feature map and is called the kernel mean or mean element of $\mathbb{P}$.

- When is $\mu_{\mathbb{P}}$ well defined?

$$\int_{\mathcal{X}} \sqrt{k(x,x)}\, d\mathbb{P}(x) < \infty \quad \Rightarrow \quad \mu_{\mathbb{P}} \in \mathcal{H}$$

Proof:

$$\|\mu_{\mathbb{P}}\|_{\mathcal{H}} = \left\| \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) \right\|_{\mathcal{H}} \overset{Jensen's}{\leq} \int_{\mathcal{X}} \|k(\cdot, x)\|_{\mathcal{H}}\, d\mathbb{P}(x).$$

- We know that for any $f \in \mathcal{H}$, $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$. So, for any $f \in \mathcal{H}$,

$$\int_{\mathcal{X}} f(x)\, d\mathbb{P}(x) = \int_{\mathcal{X}} \langle f, k(\cdot, x)\rangle_{\mathcal{H}}\, d\mathbb{P}(x) \doteq \left\langle f, \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) \right\rangle_{\mathcal{H}}$$
$$= \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}.$$

Interpretation (S et al., JMLR 2010)

Suppose k is translation invariant on $\mathbb{R}^d$, i.e., $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$. Then

$$\mu_{\mathbb{P}} = \int_{\mathbb{R}^d} \psi(\cdot - x)\, d\mathbb{P}(x) = \psi \star \mathbb{P},$$

where $\star$ is the convolution of ψ and $\mathbb{P}$.

- Convolution is a smoothing operation $\Rightarrow$ $\mu_{\mathbb{P}}$ is a smoothed version of $\mathbb{P}$.
- Example: Suppose $\mathbb{P} = \delta_y$, a point mass at y. Then

$$\mu_{\mathbb{P}} = \psi \star \mathbb{P} = \psi(\cdot - y).$$

- Example: Suppose $\psi \propto N(0, \sigma^2)$ and $\mathbb{P} = N(\mu, \tau^2)$. Then

$$\mu_{\mathbb{P}} = \psi \star \mathbb{P} \propto N(\mu, \sigma^2 + \tau^2).$$

$\mu_{\mathbb{P}}$ is a wider Gaussian than $\mathbb{P}$

Interpretation (S et al., JMLR 2010)

Suppose k is translation invariant on $\mathbb{R}^d$, i.e.,
$k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$. Then

$$\mu_{\mathbb{P}} = \int_{\mathbb{R}^d} \psi(\cdot - x)\, d\mathbb{P}(x) = \psi \star \mathbb{P},$$

where $\star$ is the convolution of ψ and $\mathbb{P}$.

- Convolution is a smoothing operation $\Rightarrow$ $\mu_{\mathbb{P}}$ is a smoothed version of $\mathbb{P}$.

- Example: Suppose $\mathbb{P} = \delta_y$, a point mass at y. Then

$$\mu_{\mathbb{P}} = \psi \star \mathbb{P} = \psi(\cdot - y).$$

- Example: Suppose $\psi \propto N(0, \sigma^2)$ and $\mathbb{P} = N(\mu, \tau^2)$. Then

$$\mu_{\mathbb{P}} = \psi \star \mathbb{P} \propto N(\mu, \sigma^2 + \tau^2).$$

$\mu_{\mathbb{P}}$ is a wider Gaussian than $\mathbb{P}$

Interpretation (S et al., JMLR 2010)

Suppose k is translation invariant on $\mathbb{R}^d$, i.e., $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$. Then

$$\mu_{\mathbb{P}} = \int_{\mathbb{R}^d} \psi(\cdot - x)\, d\mathbb{P}(x) = \psi \star \mathbb{P},$$

where $\star$ is the convolution of ψ and $\mathbb{P}$.

- Convolution is a smoothing operation $\Rightarrow$ $\mu_{\mathbb{P}}$ is a smoothed version of $\mathbb{P}$.
- Example: Suppose $\mathbb{P} = \delta_y$, a point mass at y. Then

$$\mu_{\mathbb{P}} = \psi \star \mathbb{P} = \psi(\cdot - y).$$

- Example: Suppose $\psi \propto N(0, \sigma^2)$ and $\mathbb{P} = N(\mu, \tau^2)$. Then

$$\mu_{\mathbb{P}} = \psi \star \mathbb{P} \propto N(\mu, \sigma^2 + \tau^2).$$

$\mu_{\mathbb{P}}$ is a wider Gaussian than $\mathbb{P}$

Comparing Kernel Means

Define a distance (maximum mean discrepancy) on probabilities

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

(Gretton et al., NIPS 2006; Smola et al., ALT 2007)

$$\begin{aligned}
MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{P}(x) + \int_{\mathcal{X}} \mu_{\mathbb{Q}}(x)\, d\mathbb{Q}(x) - 2\int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{P}(x)\, d\mathbb{P}(y) + \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y) \\
&\qquad - 2\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y) \\
&= \underbrace{\mathbb{E}_{\mathbb{P}} k(X, X')}_{\text{avg. similarity between points from } \mathbb{P}} + \underbrace{\mathbb{E}_{\mathbb{Q}} k(Y, Y')}_{\text{avg. similarity between points from } \mathbb{Q}} \\
&\quad - 2 \cdot \underbrace{\mathbb{E}_{\mathbb{P},\mathbb{Q}} k(X, Y)}_{\text{avg. similarity between points from } \mathbb{P} \text{ and } \mathbb{Q}}.
\end{aligned}$$

Comparing Kernel Means

Define a distance (maximum mean discrepancy) on probabilities

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

(Gretton et al., NIPS 2006; Smola et al., ALT 2007)

$$\begin{aligned}
MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{P}(x) + \int_{\mathcal{X}} \mu_{\mathbb{Q}}(x)\, d\mathbb{Q}(x) - 2 \int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{Q}(x) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y)\, d\mathbb{P}(x)\, d\mathbb{P}(y) + \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y) \\
&\quad - 2 \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y) \\
&= \underbrace{\mathbb{E}_{\mathbb{P}} k(X, X')}_{\text{avg. similarity between points from } \mathbb{P}} + \underbrace{\mathbb{E}_{\mathbb{Q}} k(Y, Y')}_{\text{avg. similarity between points from } \mathbb{Q}} \\
&\quad - 2 \cdot \underbrace{\mathbb{E}_{\mathbb{P},\mathbb{Q}} k(X, Y)}_{\text{avg. similarity between points from } \mathbb{P} \text{ and } \mathbb{Q}}.
\end{aligned}$$

Comparing Kernel Means

Define a distance (maximum mean discrepancy) on probabilities

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

(Gretton et al., NIPS 2006; Smola et al., ALT 2007)

$$\begin{aligned}
MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{P}(x) + \int_{\mathcal{X}} \mu_{\mathbb{Q}}(x)\, d\mathbb{Q}(x) - 2\int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{P}(x)\, d\mathbb{P}(y) + \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y) \\
&\quad - 2\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y) \\
&= \underbrace{\mathbb{E}_{\mathbb{P}} k(X, X')}_{\text{avg. similarity between points from } \mathbb{P}} + \underbrace{\mathbb{E}_{\mathbb{Q}} k(Y, Y')}_{\text{avg. similarity between points from } \mathbb{Q}} \\
&\quad - 2 \cdot \underbrace{\mathbb{E}_{\mathbb{P},\mathbb{Q}} k(X, Y)}_{\text{avg. similarity between points from } \mathbb{P} \text{ and } \mathbb{Q}} .
\end{aligned}$$

Comparing Kernel Means

Define a distance (maximum mean discrepancy) on probabilities

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

(Gretton et al., NIPS 2006; Smola et al., ALT 2007)

$$\begin{aligned}
MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{P}(x) + \int_{\mathcal{X}} \mu_{\mathbb{Q}}(x)\, d\mathbb{Q}(x) - 2\int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{P}(x)\, d\mathbb{P}(y) + \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y) \\
&\qquad - 2\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y) \\
&= \underbrace{\mathbb{E}_{\mathbb{P}} k(X, X')}_{\text{avg. similarity between points from } \mathbb{P}} + \underbrace{\mathbb{E}_{\mathbb{Q}} k(Y, Y')}_{\text{avg. similarity between points from } \mathbb{Q}} \\
&\qquad - 2 \cdot \underbrace{\mathbb{E}_{\mathbb{P},\mathbb{Q}} k(X, Y)}_{\text{avg. similarity between points from } \mathbb{P} \text{ and } \mathbb{Q}} .
\end{aligned}$$

Comparing Kernel Means

Define a distance (maximum mean discrepancy) on probabilities

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

(Gretton et al., NIPS 2006; Smola et al., ALT 2007)

$$\begin{aligned}
MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{P}(x) + \int_{\mathcal{X}} \mu_{\mathbb{Q}}(x)\, d\mathbb{Q}(x) - 2\int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{P}(x)\, d\mathbb{P}(y) + \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y) \\
&\qquad - 2\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y) \\
&= \underbrace{\mathbb{E}_{\mathbb{P}} k(X, X')}_{\text{avg. similarity between points from } \mathbb{P}} + \underbrace{\mathbb{E}_{\mathbb{Q}} k(Y, Y')}_{\text{avg. similarity between points from } \mathbb{Q}} \\
&\qquad - 2 \cdot \underbrace{\mathbb{E}_{\mathbb{P},\mathbb{Q}} k(X, Y)}_{\text{avg. similarity between points from } \mathbb{P} \text{ and } \mathbb{Q}} .
\end{aligned}$$

Comparing Kernel Means

Define a distance (maximum mean discrepancy) on probabilities

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

(Gretton et al., NIPS 2006; Smola et al., ALT 2007)

$$\begin{aligned}
MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{P}(x) + \int_{\mathcal{X}} \mu_{\mathbb{Q}}(x)\, d\mathbb{Q}(x) - 2\int_{\mathcal{X}} \mu_{\mathbb{P}}(x)\, d\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{P}(x)\, d\mathbb{P}(y) + \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y) \\
&\qquad - 2\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y) \\
&= \underbrace{\mathbb{E}_{\mathbb{P}} k(X, X')}_{\text{avg. similarity between points from } \mathbb{P}} + \underbrace{\mathbb{E}_{\mathbb{Q}} k(Y, Y')}_{\text{avg. similarity between points from } \mathbb{Q}} \\
&\quad - 2 \cdot \underbrace{\mathbb{E}_{\mathbb{P},\mathbb{Q}} k(X, Y)}_{\text{avg. similarity between points from } \mathbb{P} \text{ and } \mathbb{Q}}.
\end{aligned}$$

Comparing Kernel Means

In the motivating examples, we compare $\mathbb{P}$ and $\mathbb{Q}$ by comparing

$$\mu_{\mathbb{P}}(y) = \int_{\mathcal{X}} k(y,x)\, d\mathbb{P}(x) \quad \text{and} \quad \mu_{\mathbb{Q}}(y) = \int_{\mathcal{X}} k(y,x)\, d\mathbb{Q}(x), \ \forall\, y \in \mathcal{X}.$$

For any $f \in \mathcal{H}$,

$$\|f\|_\infty = \sup_{y\in\mathcal{X}} |f(y)| = \sup_{y\in\mathcal{X}} |\langle f, k(\cdot,y)\rangle_{\mathcal{H}}| \le \sup_{y\in\mathcal{X}} \sqrt{k(y,y)}\|f\|_{\mathcal{H}}.$$

$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_\infty \le \sup_{y\in\mathcal{X}} \sqrt{k(y,y)}\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

Does $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$? (More on this later)

Comparing Kernel Means

In the motivating examples, we compare $\mathbb{P}$ and $\mathbb{Q}$ by comparing

$$\mu_{\mathbb{P}}(y) = \int_{\mathcal{X}} k(y,x)\, d\mathbb{P}(x) \quad \text{and} \quad \mu_{\mathbb{Q}}(y) = \int_{\mathcal{X}} k(y,x)\, d\mathbb{Q}(x),\ \forall\, y \in \mathcal{X}.$$

For any $f \in \mathcal{H}$,

$$\|f\|_\infty = \sup_{y\in\mathcal{X}} |f(y)| = \sup_{y\in\mathcal{X}} |\langle f, k(\cdot,y)\rangle_{\mathcal{H}}| \leq \sup_{y\in\mathcal{X}} \sqrt{k(y,y)}\|f\|_{\mathcal{H}}.$$

$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_\infty \leq \sup_{y\in\mathcal{X}} \sqrt{k(y,y)}\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

Does $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$? (More on this later)

Comparing Kernel Means

In the motivating examples, we compare $\mathbb{P}$ and $\mathbb{Q}$ by comparing

$$\mu_{\mathbb{P}}(y) = \int_{\mathcal{X}} k(y,x)\, d\mathbb{P}(x) \quad \text{and} \quad \mu_{\mathbb{Q}}(y) = \int_{\mathcal{X}} k(y,x)\, d\mathbb{Q}(x), \ \forall\, y \in \mathcal{X}.$$

For any $f \in \mathcal{H}$,

$$\|f\|_\infty = \sup_{y\in\mathcal{X}} |f(y)| = \sup_{y\in\mathcal{X}} |\langle f, k(\cdot, y)\rangle_{\mathcal{H}}| \leq \sup_{y\in\mathcal{X}} \sqrt{k(y,y)}\|f\|_{\mathcal{H}}.$$

$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_\infty \leq \sup_{y\in\mathcal{X}} \sqrt{k(y,y)}\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

Does $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$? (More on this later)

Integral Probability Metric

The integral probability metric between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$
\begin{aligned}
IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}) &:= \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} f(x)\, d\mathbb{Q}(x) \right| \\
&= \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\mathbb{P}} f(X) - \mathbb{E}_{\mathbb{Q}} f(X) \right| .
\end{aligned}
$$

(Müller, 1997)

- $\mathcal{F}$ controls the degree of distinguishability between $\mathbb{P}$ and $\mathbb{Q}$.
- Related to the Bayes risk of a certain classification problem (S et al., NIPS 2009; EJS 2012)

Integral Probability Metric

The integral probability metric between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\begin{aligned} IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}) &:= \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} f(x)\, d\mathbb{Q}(x) \right| \\ &= \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\mathbb{P}} f(X) - \mathbb{E}_{\mathbb{Q}} f(X) \right|. \end{aligned}$$

(Müller, 1997)

- $\mathcal{F}$ controls the degree of distinguishability between $\mathbb{P}$ and $\mathbb{Q}$.
- Related to the Bayes risk of a certain classification problem (S et al., NIPS 2009; EJS 2012)
- Example: Suppose $\mathcal{F} = \{a \cdot x,\ x \in \mathbb{R} : a \in [-1, 1]\}$. Then

$$IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = \sup_{a \in [-1,1]} |a| \left| \int_{\mathbb{R}} x\, d\mathbb{P}(x) - \int_{\mathbb{R}} x\, d\mathbb{Q}(x) \right|$$

Integral Probability Metric

Example: Suppose $\mathcal{F} = \{a \cdot x + b \cdot x^2,\, x \in \mathbb{R} \,:\, a^2 + b^2 = 1\}$. Then

$$
\begin{aligned}
IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}) &= \sup_{a^2+b^2=1} \left| a \int_{\mathbb{R}} x \, d(\mathbb{P} - \mathbb{Q}) + b \int_{\mathbb{R}} x^2 \, d(\mathbb{P} - \mathbb{Q}) \right| \\
&= \left[\left(\int_{\mathbb{R}} x \, d(\mathbb{P} - \mathbb{Q}) \right)^2 + \left(\int_{\mathbb{R}} x^2 \, d(\mathbb{P} - \mathbb{Q}) \right)^2 \right]^{\frac{1}{2}}.
\end{aligned}
$$

How? Exercise!

- The richer the $\mathcal{F}$ is, the finer is the resolvability of $\mathbb{P}$ and $\mathbb{Q}$.

We will explore the relation of $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})$ to $IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F})$.

Integral Probability Metric

$$IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} f(x)\, d\mathbb{Q}(x) \right|$$

Classical results:

- $\mathcal{F} =$ unit Lipschitz ball (Wasserstein distance) (Dudley, 2002)
- $\mathcal{F} =$ unit bounded-Lipschitz ball (Dudley metric) (Dudley, 2002)
- $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\}$ (Kolmogorov metric) (Müller, 1997)
- $\mathcal{F} =$ unit ball in bounded measurable functions (Total variation distance) (Dudley, 2002)

For all these $\mathcal{F}$, $IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$.

(Gretton et al., NIPS 2006, JMLR 2012; S et al., COLT 2008): $\mathcal{F} =$ unit ball in an RKHS, $\mathcal{H}$ with bounded kernel, k. Then

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}).$$

Proof: $\int_{\mathcal{X}} f(x)\, d(\mathbb{P} - \mathbb{Q})(x) = \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$

Integral Probability Metric

$$IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} f(x)\, d\mathbb{Q}(x) \right|$$

Classical results:

- $\mathcal{F} =$ unit Lipschitz ball (Wasserstein distance) (Dudley, 2002)
- $\mathcal{F} =$ unit bounded-Lipschitz ball (Dudley metric) (Dudley, 2002)
- $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\}$ (Kolmogorov metric) (Müller, 1997)
- $\mathcal{F} =$ unit ball in bounded measurable functions (Total variation distance) (Dudley, 2002)

For all these $\mathcal{F}$, $IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$.

(Gretton et al., NIPS 2006, JMLR 2012; S et al., COLT 2008): $\mathcal{F} =$ unit ball in an RKHS, $\mathcal{H}$ with bounded kernel, k. Then

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}).$$

Proof: $\int_{\mathcal{X}} f(x)\, d(\mathbb{P} - \mathbb{Q})(x) = \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \le \|f\|_{\mathcal{H}} \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$

Integral Probability Metric

$$IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} f(x)\, d\mathbb{Q}(x) \right|$$

Classical results:

- $\mathcal{F} =$ unit Lipschitz ball (Wasserstein distance) (Dudley, 2002)
- $\mathcal{F} =$ unit bounded-Lipschitz ball (Dudley metric) (Dudley, 2002)
- $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\}$ (Kolmogorov metric) (Müller, 1997)
- $\mathcal{F} =$ unit ball in bounded measurable functions (Total variation distance) (Dudley, 2002)

For all these $\mathcal{F}$, $IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$.

(Gretton et al., NIPS 2006, JMLR 2012; S et al., COLT 2008): $\mathcal{F} =$ unit ball in an RKHS, $\mathcal{H}$ with bounded kernel, k. Then

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F}).$$

Proof: $\int_{\mathcal{X}} f(x)\, d(\mathbb{P} - \mathbb{Q})(x) = \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$.

Two-Sample Problem

- Given random samples $\{X_1, \ldots, X_m\} \overset{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} \mathbb{Q}$.
- Determine: $\mathbb{P} = \mathbb{Q}$ or $\mathbb{P} \neq \mathbb{Q}$?
- Approach: Define ρ to be a distance on probabilities

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

- If empirical ρ is
  - far from zero: reject H_0
  - close to zero: accept H_0

Two-Sample Problem

- Given random samples $\{X_1, \ldots, X_m\} \overset{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} \mathbb{Q}$.
- Determine: $\mathbb{P} = \mathbb{Q}$ or $\mathbb{P} \neq \mathbb{Q}$?
- Approach: Define ρ to be a distance on probabilities

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

- If empirical ρ is
  - far from zero: reject H_0
  - close to zero: accept H_0

Two-Sample Problem

- Given random samples $\{X_1, \ldots, X_m\} \overset{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} \mathbb{Q}$.
- Determine: $\mathbb{P} = \mathbb{Q}$ or $\mathbb{P} \neq \mathbb{Q}$?
- Approach: Define ρ to be a distance on probabilities

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

- If empirical ρ is
  - far from zero: reject H_0
  - close to zero: accept H_0

Why $MMD_{\mathcal{H}}$?

- Related to the estimation of $IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F})$.
- Recall

$$MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \left\| \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{Q}(x) \right\|^2_{\mathcal{H}}.$$

- A trivial approximation: $\mathbb{P}_m := \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$ and $\mathbb{Q}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, where δ_x represents the Dirac measure at x.

$$MMD^2_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_n) = \left\| \frac{1}{m}\sum_{i=1}^m k(\cdot, X_i) - \frac{1}{n}\sum_{i=1}^n k(\cdot, Y_i) \right\|^2_{\mathcal{H}}$$

$$= \frac{1}{m^2}\sum_{i,j=1}^m k(X_i, X_j) + \frac{1}{n^2}\sum_{i,j=1}^n k(Y_i, Y_j) - 2\sum_{i,j} k(X_i, Y_j)$$

V-statistic; biased estimator of $MMD^2_{\mathcal{H}}$

Why $MMD_{\mathcal{H}}$?

- Related to the estimation of $IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F})$.
- Recall

$$MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \left\| \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{Q}(x) \right\|^2_{\mathcal{H}}.$$

- A trivial approximation: $\mathbb{P}_m := \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$ and $\mathbb{Q}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, where δ_x represents the Dirac measure at x.

$$MMD^2_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_n) = \left\| \frac{1}{m}\sum_{i=1}^m k(\cdot, X_i) - \frac{1}{n}\sum_{i=1}^n k(\cdot, Y_i) \right\|^2_{\mathcal{H}}$$

$$= \frac{1}{m^2}\sum_{i,j=1}^m k(X_i, X_j) + \frac{1}{n^2}\sum_{i,j=1}^n k(Y_i, Y_j) - 2\sum_{i,j} k(X_i, Y_j)$$

V-statistic; biased estimator of $MMD^2_{\mathcal{H}}$

Why $MMD_{\mathcal{H}}$?

- $IPM(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F})$ is obtained by solving a linear program for $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls. (S et al., EJS 2012)
- Quality of approximation (S et al., EJS 2012)
  - For $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls,

$$|IPM(\mathbb{P}_m, \mathbb{Q}_m, \mathcal{F}) - IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F})| = O_p\left(m^{-\frac{1}{d+1}}\right), \quad d > 2$$

  - For $\mathcal{F} =$ unit RKHS ball,

$$|MMD_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_m) - MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})| = O_p\left(m^{-\frac{1}{2}}\right)$$

- Are there any other estimators of $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})$ that are statistically better than $MMD_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_m)$? NO!! (Tolstikhin et al., 2016)
- In practice? YES!! (Krikamol et al., JMLR 2016; S, Bernoulli 2016)

Why $MMD_{\mathcal{H}}$?

- $IPM(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F})$ is obtained by solving a linear program for $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls. (S et al., EJS 2012)
- Quality of approximation (S et al., EJS 2012)
  - For $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls,

$$|IPM(\mathbb{P}_m, \mathbb{Q}_m, \mathcal{F}) - IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F})| = O_p\left(m^{-\frac{1}{d+1}}\right), \quad d > 2$$

  - For $\mathcal{F} =$ unit RKHS ball,

$$|MMD_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_m) - MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})| = O_p\left(m^{-\frac{1}{2}}\right)$$

- Are there any other estimators of $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})$ that are statistically better than $MMD_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_m)$? NO!! (Tolstikhin et al., 2016)
- In practice? YES!! (Krikamol et al., JMLR 2016; S, Bernoulli 2016)

Why $MMD_{\mathcal{H}}$?

- $IPM(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F})$ is obtained by solving a linear program for $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls. (S et al., EJS 2012)
- Quality of approximation (S et al., EJS 2012)
  - For $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls,

$$|IPM(\mathbb{P}_m, \mathbb{Q}_m, \mathcal{F}) - IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F})| = O_p\left(m^{-\frac{1}{d+1}}\right), \quad d > 2$$

  - For $\mathcal{F} =$ unit RKHS ball,

$$|MMD_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_m) - MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})| = O_p\left(m^{-\frac{1}{2}}\right)$$

- Are there any other estimators of $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})$ that are statistically better than $MMD_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_m)$? NO!! (Tolstikhin et al., 2016)
- In practice? YES!! (Krikamol et al., JMLR 2016; S, Bernoulli 2016)

Why $MMD_{\mathcal{H}}$?

- $IPM(\mathbb{P}_m, \mathbb{Q}_n, \mathcal{F})$ is obtained by solving a linear program for $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls. (S et al., EJS 2012)

- Quality of approximation (S et al., EJS 2012)

  - For $\mathcal{F} =$ Lipschitz and bounded Lipschitz balls,

$$|IPM(\mathbb{P}_m, \mathbb{Q}_m, \mathcal{F}) - IPM(\mathbb{P}, \mathbb{Q}, \mathcal{F})| = O_p\left(m^{-\frac{1}{d+1}}\right), \ \ d > 2$$

  - For $\mathcal{F} =$ unit RKHS ball,

$$|MMD_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_m) - MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})| = O_p\left(m^{-\frac{1}{2}}\right)$$

- Are there any other estimators of $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})$ that are statistically better than $MMD_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_m)$? NO!! (Tolstikhin et al., 2016)

- In practice? YES!! (Krikamol et al., JMLR 2016; S, Bernoulli 2016)

Beware of Pitfalls

- There are many other distances on probabilities:
  - Total variation distance
  - Hellinger distance
  - Kullback-Leibler divergence and its variants
  - Fisher divergence ...
- Estimating these distances is both computationally and statistically difficult.
- $MMD_{\mathcal{H}}$ is computationally simpler and appears statistically powerful with no curse of dimensionality. In fact, it is NOT statistically powerful. (Ramdas et al., AAAI 2015; S, Bernoulli, 2016)
- Recall: $MMD_{\mathcal{H}}$ is based on $\mu_{\mathbb{P}}$ which is a smoothed version of $\mathbb{P}$. Even though $\mathbb{P}$ and $\mathbb{Q}$ can be distinguished (coming up!!) based on $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$, the distinguishability is weak compared to that of the above distances. (S et al., JMLR 2010; S, Bernoulli, 2016)

NO FREE LUNCH!!

Beware of Pitfalls

- There are many other distances on probabilities:
  - Total variation distance
  - Hellinger distance
  - Kullback-Leibler divergence and its variants
  - Fisher divergence ...
- Estimating these distances is both computationally and statistically difficult.
- $MMD_{\mathcal{H}}$ is computationally simpler and appears statistically powerful with no curse of dimensionality. In fact, it is NOT statistically powerful. (Ramdas et al., AAAI 2015; S, Bernoulli, 2016)
- Recall: $MMD_{\mathcal{H}}$ is based on $\mu_{\mathbb{P}}$ which is a smoothed version of $\mathbb{P}$. Even though $\mathbb{P}$ and $\mathbb{Q}$ can be distinguished (coming up!!) based on $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$, the distinguishability is weak compared to that of the above distances. (S et al., JMLR 2010; S, Bernoulli, 2016)

NO FREE LUNCH!!

Beware of Pitfalls

- There are many other distances on probabilities:
  - Total variation distance
  - Hellinger distance
  - Kullback-Leibler divergence and its variants
  - Fisher divergence ...
- Estimating these distances is both computationally and statistically difficult.
- $MMD_{\mathcal{H}}$ is computationally simpler and appears statistically powerful with no curse of dimensionality. In fact, it is NOT statistically powerful. (Ramdas et al., AAAI 2015; S, Bernoulli, 2016)
- Recall: $MMD_{\mathcal{H}}$ is based on $\mu_{\mathbb{P}}$ which is a smoothed version of $\mathbb{P}$. Even though $\mathbb{P}$ and $\mathbb{Q}$ can be distinguished (coming up!!) based on $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$, the distinguishability is weak compared to that of the above distances. (S et al., JMLR 2010; S, Bernoulli, 2016)

NO FREE LUNCH!!

So far . . .

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$$

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

- Computation
- Estimation

When is $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ one-to-one?, i.e., $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 0 \quad \Rightarrow \quad \mathbb{P} = \mathbb{Q}$?

Characteristic Kernel

k is said to be characteristic if

$$MMD_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$$

for any $\mathbb{P}$ and $\mathbb{Q}$.

Not all kernels are characteristic.

- Example: If $k(x,y) = c > 0, \forall\, x,y \in \mathcal{X}$, then

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{P}(x) = c, \quad \mu_{\mathbb{Q}} = c$$

  and $MMD_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = 0, \forall\, \mathbb{P},\mathbb{Q}$.

- Example: Let $k(x,y) = xy,\ x,y \in \mathbb{R}$. Then

$$MMD_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = |\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X]|.$$

  Characteristic for Bernoulli's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

- Example: Let $k(x,y) = (1+xy)^2,\ x,y \in \mathbb{R}$. Then

$$MMD^2_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = 2(\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X])^2 + (\mathbb{E}_{\mathbb{P}}[X^2] - \mathbb{E}_{\mathbb{Q}}[X^2]).$$

  Characteristic for Gaussian's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

Characteristic Kernel

k is said to be characteristic if

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$$

for any $\mathbb{P}$ and $\mathbb{Q}$.

Not all kernels are characteristic.

- Example: If $k(x, y) = c > 0, \forall x, y \in \mathcal{X}$, then

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) = c, \quad \mu_{\mathbb{Q}} = c$$

  and $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 0, \forall \mathbb{P}, \mathbb{Q}$.

- Example: Let $k(x, y) = xy$, $x, y \in \mathbb{R}$. Then

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = |\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X]|.$$

  Characteristic for Bernoulli's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

- Example: Let $k(x, y) = (1 + xy)^2$, $x, y \in \mathbb{R}$. Then

$$MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 2(\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X])^2 + (\mathbb{E}_{\mathbb{P}}[X^2] - \mathbb{E}_{\mathbb{Q}}[X^2]).$$

  Characteristic for Gaussian's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

Characteristic Kernel

k is said to be characteristic if

$$MMD_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$$

for any $\mathbb{P}$ and $\mathbb{Q}$.

Not all kernels are characteristic.

- Example: If $k(x,y) = c > 0$, $\forall\, x, y \in \mathcal{X}$, then

  $$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{P}(x) = c, \quad \mu_{\mathbb{Q}} = c$$

  and $MMD_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = 0$, $\forall\, \mathbb{P},\mathbb{Q}$.

- Example: Let $k(x,y) = xy$, $x, y \in \mathbb{R}$. Then

  $$MMD_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = |\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X]|.$$

  Characteristic for Bernoulli's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

- Example: Let $k(x,y) = (1+xy)^2$, $x, y \in \mathbb{R}$. Then

  $$MMD^2_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = 2(\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X])^2 + (\mathbb{E}_{\mathbb{P}}[X^2] - \mathbb{E}_{\mathbb{Q}}[X^2]).$$

  Characteristic for Gaussian's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

Characteristic Kernel

k is said to be characteristic if

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$$

for any $\mathbb{P}$ and $\mathbb{Q}$.

Not all kernels are characteristic.

- Example: If $k(x, y) = c > 0, \forall\, x, y \in \mathcal{X}$, then

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) = c, \quad \mu_{\mathbb{Q}} = c$$

  and $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 0, \forall\, \mathbb{P}, \mathbb{Q}$.

- Example: Let $k(x, y) = xy$, $x, y \in \mathbb{R}$. Then

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = |\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X]|.$$

  Characteristic for Bernoulli's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

- Example: Let $k(x, y) = (1 + xy)^2$, $x, y \in \mathbb{R}$. Then

$$MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 2(\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X])^2 + (\mathbb{E}_{\mathbb{P}}[X^2] - \mathbb{E}_{\mathbb{Q}}[X^2]).$$

  Characteristic for Gaussian's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

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

- Translation invariant kernel: $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$; bounded and continuous.
- Bochner's theorem:

$$\psi(x) = \int_{\mathbb{R}^d} e^{\sqrt{-1}\langle x,\omega\rangle_2}\, d\Lambda(\omega),\ \ x \in \mathbb{R}^d,$$

  where Λ is a non-negative finite Borel measure on $\mathbb{R}^d$.

Then, k is characteristic $\Leftrightarrow$ $\text{supp}(\Lambda) = \mathbb{R}^d$. (S et al., COLT 2008; JMLR, 2010)

- Corollary: Compactly supported ψ are characteristic (S et al., COLT 2008; JMLR, 2010).

Key Idea: Fourier representation of $MMD_{\mathcal{H}}$

Fourier Representation of $MMD^2_{\mathcal{H}}$

$$MMD^2_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}^d} |\varphi_{\mathbb{P}}(\omega) - \varphi_{\mathbb{Q}}(\omega)|^2 \, d\Lambda(\omega)$$

where $\varphi_{\mathbb{P}}$ is the characteristic function of $\mathbb{P}$.

Proof:

$$\begin{aligned}
MMD^2_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \psi(x-y)\, d(\mathbb{P}-\mathbb{Q})(x)\, d(\mathbb{P}-\mathbb{Q})(y) \\
&\overset{(*)}{=} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-\sqrt{-1}\langle x-y,\omega\rangle}\, d\Lambda(\omega)\, d(\mathbb{P}-\mathbb{Q})(x)\, d(\mathbb{P}-\mathbb{Q})(y) \\
&\overset{(\dagger)}{=} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-\sqrt{-1}\langle x,\omega\rangle}\, d(\mathbb{P}-\mathbb{Q})(x) \int_{\mathbb{R}^d} e^{\sqrt{-1}\langle y,\omega\rangle}\, d(\mathbb{P}-\mathbb{Q})(y)\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\varphi_{\mathbb{P}}(\omega) - \varphi_{\mathbb{Q}}(\omega)|^2 \, d\Lambda(\omega),
\end{aligned}$$

where Bochner's theorem is used in $(*)$ and Fubini's theorem in $(\dagger)$.

- Suppose $\Lambda = 1$, i.e., uniform on $\mathbb{R}^d$ (!!). Then $MMD_{\mathcal{H}}(\mathbb{P},\mathbb{Q})$ is the L^2 distance between the densities (if they exist) of $\mathbb{P}$ and $\mathbb{Q}$.

Fourier Representation of $MMD^2_{\mathcal{H}}$

$$MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} |\varphi_{\mathbb{P}}(\omega) - \varphi_{\mathbb{Q}}(\omega)|^2 \, d\Lambda(\omega)$$

where $\varphi_{\mathbb{P}}$ is the characteristic function of $\mathbb{P}$.

Proof:

$$\begin{aligned} MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \psi(x-y) \, d(\mathbb{P}-\mathbb{Q})(x) \, d(\mathbb{P}-\mathbb{Q})(y) \\ &\overset{(*)}{=} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-\sqrt{-1}\langle x-y, \omega \rangle} \, d\Lambda(\omega) \, d(\mathbb{P}-\mathbb{Q})(x) \, d(\mathbb{P}-\mathbb{Q})(y) \\ &\overset{(\dagger)}{=} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-\sqrt{-1}\langle x, \omega \rangle} \, d(\mathbb{P}-\mathbb{Q})(x) \int_{\mathbb{R}^d} e^{\sqrt{-1}\langle y, \omega \rangle} \, d(\mathbb{P}-\mathbb{Q})(y) \, d\Lambda(\omega) \\ &= \int_{\mathbb{R}^d} |\varphi_{\mathbb{P}}(\omega) - \varphi_{\mathbb{Q}}(\omega)|^2 \, d\Lambda(\omega), \end{aligned}$$

where Bochner's theorem is used in $(*)$ and Fubini's theorem in $(\dagger)$.

- Suppose $\Lambda = 1$, i.e., uniform on $\mathbb{R}^d$ (!!). Then $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})$ is the L^2 distance between the densities (if they exist) of $\mathbb{P}$ and $\mathbb{Q}$.

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

Proof:

- Suppose $\text{supp}(\Lambda) = \mathbb{R}^d$. Then

$$MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow \int_{\mathbb{R}^d} \left|\varphi_{\mathbb{P}}(\omega) - \varphi_{\mathbb{Q}}(\omega)\right|^2 \, d\Lambda(\omega) = 0 \Rightarrow \varphi_{\mathbb{P}} = \varphi_{\mathbb{Q}} \text{ a.e.}$$

  But characteristic functions are uniformly continuous and so $\varphi_{\mathbb{P}} = \varphi_{\mathbb{Q}}$ which implies $\mathbb{P} = \mathbb{Q}$.

- Suppose $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. Then there exists an open set $U \subsetneq \mathbb{R}^d$ such that $\Lambda(U) = 0$. Construct $\mathbb{P}$ and $\mathbb{Q}$ such that $\varphi_{\mathbb{P}}$ and $\varphi_{\mathbb{Q}}$ differ only in U, i.e., $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) > 0$.

- If ψ is compactly supported, its Fourier transform is analytic, i.e., cannot vanish on an interval.

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

Proof:

- Suppose $\text{supp}(\Lambda) = \mathbb{R}^d$. Then

$$MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow \int_{\mathbb{R}^d} |\varphi_{\mathbb{P}}(\omega) - \varphi_{\mathbb{Q}}(\omega)|^2 \, d\Lambda(\omega) = 0 \Rightarrow \varphi_{\mathbb{P}} = \varphi_{\mathbb{Q}} \text{ a.e.}$$

  But characteristic functions are uniformly continuous and so $\varphi_{\mathbb{P}} = \varphi_{\mathbb{Q}}$ which implies $\mathbb{P} = \mathbb{Q}$.

- Suppose $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. Then there exists an open set $U \subsetneq \mathbb{R}^d$ such that $\Lambda(U) = 0$. Construct $\mathbb{P}$ and $\mathbb{Q}$ such that $\varphi_{\mathbb{P}}$ and $\varphi_{\mathbb{Q}}$ differ only in U, i.e., $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) > 0$.

- If ψ is compactly supported, its Fourier transform is analytic, i.e., cannot vanish on an interval.

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

Proof:

- Suppose $\text{supp}(\Lambda) = \mathbb{R}^d$. Then

  $$MMD^2_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = 0 \Rightarrow \int_{\mathbb{R}^d} \left|\varphi_{\mathbb{P}}(\omega) - \varphi_{\mathbb{Q}}(\omega)\right|^2 \, d\Lambda(\omega) = 0 \Rightarrow \varphi_{\mathbb{P}} = \varphi_{\mathbb{Q}} \text{ a.e.}$$

  But characteristic functions are uniformly continuous and so $\varphi_{\mathbb{P}} = \varphi_{\mathbb{Q}}$ which implies $\mathbb{P} = \mathbb{Q}$.

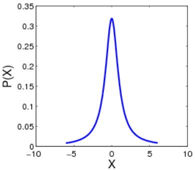
- Suppose $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. Then there exists an open set $U \subsetneq \mathbb{R}^d$ such that $\Lambda(U) = 0$. Construct $\mathbb{P}$ and $\mathbb{Q}$ such that $\varphi_{\mathbb{P}}$ and $\varphi_{\mathbb{Q}}$ differ only in U, i.e., $MMD_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) > 0$.

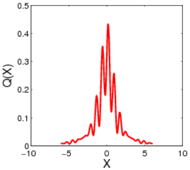
- If ψ is compactly supported, its Fourier transform is <u>analytic</u>, i.e., cannot vanish on an interval.

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

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\varphi_{\mathbb{P}} - \varphi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

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

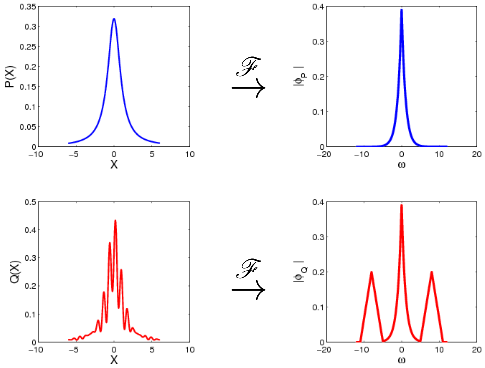




Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\varphi_{\mathbb{P}} - \varphi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

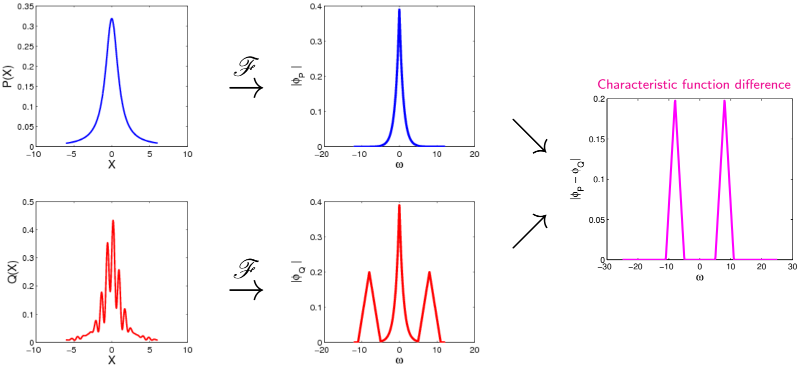
- Example: $\mathbb{P}$ differs from $\mathbb{Q}$ at (roughly) one frequency



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

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\varphi_{\mathbb{P}} - \varphi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

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



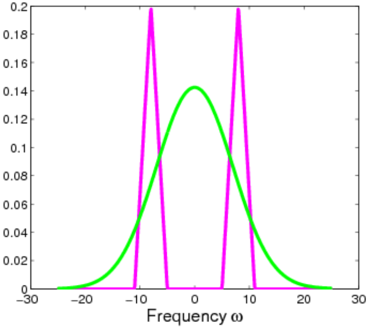
Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\varphi_{\mathbb{P}} - \varphi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

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

Gaussian kernel

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

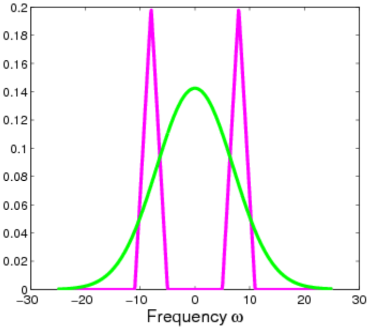


Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\varphi_{\mathbb{P}} - \varphi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

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

Characteristic



Picture credit: A. Gretton

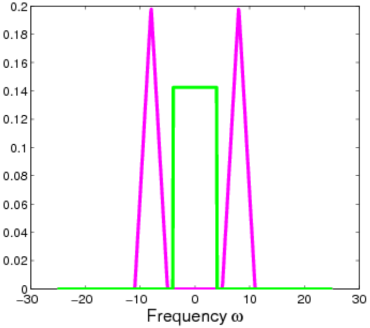
Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\varphi_{\mathbb{P}} - \varphi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

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

Sinc kernel

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

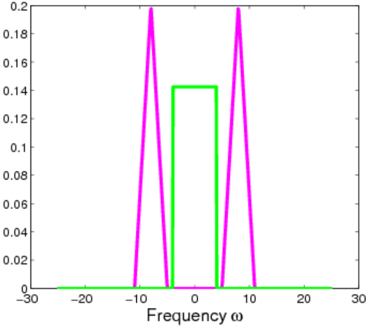


Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\varphi_{\mathbb{P}} - \varphi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d,\Lambda)}$$

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

NOT characteristic



Picture credit: A. Gretton

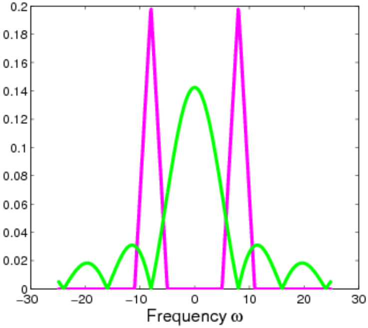
Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\varphi_{\mathbb{P}} - \varphi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

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

B-Spline kernel

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

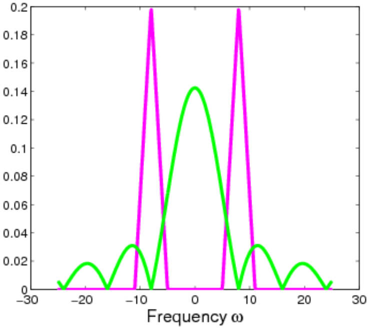


Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\varphi_{\mathbb{P}} - \varphi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

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

???

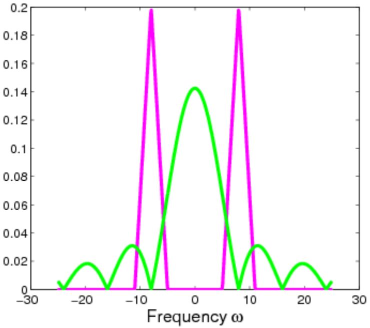


Translation Invariant Kernels on $\mathbb{R}^d$

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\varphi_{\mathbb{P}} - \varphi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d, \Lambda)}$$

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

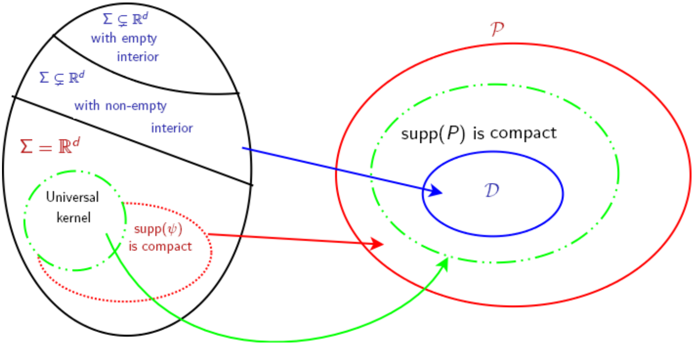
Characteristic



Picture credit: A. Gretton

Caution

Chararacteristic property relates class of kernels and class of probabilities.



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

(S et al., COLT 2008; JMLR 2010)

Characteristic Kernels

Similar reasoning hold wherever extensions of Bochner's theorem exist (Fukumizu et al., NIPS 2009):

- Locally compact Abelian groups (periodic domains, e.g., circle, d-Torus)
- Compact, non-Abelian groups (Orthogonal matrices)
  - Represent and compare distributions over matrices
- The semigroup $\mathbb{R}_+^d$ (histograms)
  - Compare distributions over distributions

- Characteristic property is related to the richness of $\mathcal{H}$ in approximating certain class of functions. Characteristic property is in general a weaker notion than universality. But for translation invariant kernels on $\mathbb{R}^d$, these notions are equivalent. (Gretton et al., NIPS 2006; Fukumizu et al., NIPS 2008, 2009; Steinwart and Christmann, 2008; S et al., JMLR 2010, JMLR 2011; Simon-Gabriel and Schölkopf, 2016)

Characteristic Kernels

Similar reasoning hold wherever extensions of Bochner's theorem exist (Fukumizu et al., NIPS 2009):

- Locally compact Abelian groups (periodic domains, e.g., circle, d-Torus)
- Compact, non-Abelian groups (Orthogonal matrices)
  - Represent and compare distributions over matrices
- The semigroup $\mathbb{R}_+^d$ (histograms)
  - Compare distributions over distributions

- Characteristic property is related to the richness of $\mathcal{H}$ in approximating certain class of functions. Characteristic property is in general a weaker notion than universality. But for translation invariant kernels on $\mathbb{R}^d$, these notions are equivalent. (Gretton et al., NIPS 2006; Fukumizu et al., NIPS 2008, 2009; Steinwart and Christmann, 2008; S et al., JMLR 2010, JMLR 2011; Simon-Gabriel and Schölkopf, 2016)

So far . . .

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$$

$$MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

- Computation
- Estimation
- $MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 0 \quad \Rightarrow \quad \mathbb{P} = \mathbb{Q}$ for characteristic kernels.

Measuring (In)Dependence

- X and Y are random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$.
- $(X, Y) \sim \mathbb{P}_{XY}$ with marginals $X \sim \mathbb{P}_X$ and $Y \sim \mathbb{P}_Y$.
- Dependency measure using MMD: Using k_X and k_Y defined on $\mathcal{X}$ and $\mathcal{Y}$,

$$MMD_{\mathcal{H}}(\mathbb{P}_{XY}, \mathbb{P}_X \times \mathbb{P}_Y) := \left\| \int \underbrace{k_X(\cdot, x) k_Y(\cdot, y)}_{k(\cdot,(x,y))} \, d(\mathbb{P}_{XY} - \mathbb{P}_X \times \mathbb{P}_Y)(x, y) \right\|_{\mathcal{H}}$$

  where $\mathcal{H} := \mathcal{H}_X \otimes \mathcal{H}_Y$.
- If k is characteristic on $\mathcal{X} \times \mathcal{Y}$, then $MMD_{\mathcal{H}}$ captures independence.

Measuring (In)Dependence

- X and Y are random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$.
- $(X, Y) \sim \mathbb{P}_{XY}$ with marginals $X \sim \mathbb{P}_X$ and $Y \sim \mathbb{P}_Y$.
- Dependency measure using MMD: Using k_X and k_Y defined on $\mathcal{X}$ and $\mathcal{Y}$,

$$MMD_{\mathcal{H}}(\mathbb{P}_{XY}, \mathbb{P}_X \times \mathbb{P}_Y) := \left\| \int \underbrace{k_X(\cdot, x) k_Y(\cdot, y)}_{k(\cdot,(x,y))} \, d(\mathbb{P}_{XY} - \mathbb{P}_X \times \mathbb{P}_Y)(x, y) \right\|_{\mathcal{H}}$$

  where $\mathcal{H} := \mathcal{H}_X \otimes \mathcal{H}_Y$.
- If k is characteristic on $\mathcal{X} \times \mathcal{Y}$, then $MMD_{\mathcal{H}}$ captures independence.

Measuring (In)Dependence

- Let X and Y be Gaussian random variables on $\mathbb{R}$. Then

  X and Y are independent $\Leftrightarrow \text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$

- In general, $\text{Cov}(X, Y) = 0 \not\Rightarrow X \perp Y$.
- Covariance captures the linear relationship between X and Y.
- Feature space view point: How about $\text{Cov}(\Phi(X), \Psi(Y))$?
- Suppose

  $$\Phi(X) = (1, X, X^2) \text{ and } \Psi(Y) = (1, Y, Y^2, Y^3).$$

  Then $\text{Cov}(\Phi(X), \Phi(Y))$ captures $\text{Cov}(X^i, Y^j)$ for $i \in \{0, 1, 2\}$ and $j \in \{0, 1, 2, 3\}$.

Measuring (In)Dependence

- Let X and Y be Gaussian random variables on $\mathbb{R}$. Then

  X and Y are independent $\Leftrightarrow \mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$

- In general, $\mathrm{Cov}(X, Y) = 0 \not\Rightarrow X \perp Y$.
- Covariance captures the linear relationship between X and Y.
- Feature space view point: How about $\mathrm{Cov}(\Phi(X), \Psi(Y))$?
- Suppose

$$\Phi(X) = (1, X, X^2) \text{ and } \Psi(Y) = (1, Y, Y^2, Y^3).$$

  Then $\mathrm{Cov}(\Phi(X), \Phi(Y))$ captures $\mathrm{Cov}(X^i, Y^j)$ for $i \in \{0, 1, 2\}$ and $j \in \{0, 1, 2, 3\}$.

Measuring (In)Dependence

- Characterization of independence:

$$X \perp Y \Leftrightarrow \text{Cov}(f(X), g(Y)) = 0, \forall \text{ measurable functions } f \text{ and } g.$$

- Dependence measure:

$$\sup_{f,g} |\text{Cov}(f(X), g(Y))| = \sup_{f,g} |\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]|$$

Similar to the IPM between $\mathbb{P}_{XY}$ and $\mathbb{P}_X\mathbb{P}_Y$.

- Restricting functions in RKHS: (constrained covariance)

$$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) := \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} |\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]|.$$

(Gretton et al., AISTATS 2005, JMLR 2005)

Measuring (In)Dependence

- Characterization of independence:

$$X \perp Y \Leftrightarrow \text{Cov}(f(X), g(Y)) = 0, \forall \text{ measurable functions } f \text{ and } g.$$

- Dependence measure:

$$\sup_{f,g} |\text{Cov}(f(X), g(Y))| = \sup_{f,g} |\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]|$$

Similar to the IPM between $\mathbb{P}_{XY}$ and $\mathbb{P}_X\mathbb{P}_Y$.

- Restricting functions in RKHS: (constrained covariance)

$$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) := \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} |\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]|.$$

(Gretton et al., AISTATS 2005, JMLR 2005)

Covariance Operator

Let k_X and k_Y be the r.k.'s of $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively. Then

- $\mathbb{E}[f(X)] = \langle f, \mu_{\mathbb{P}_X} \rangle_{\mathcal{H}_X}$ and $\mathbb{E}[g(Y)] = \langle g, \mu_{\mathbb{P}_Y} \rangle_{\mathcal{H}_Y}$
- $$\begin{aligned} \mathbb{E}[f(X)]\mathbb{E}[g(Y)] &= \langle f, \mu_{\mathbb{P}_X} \rangle_{\mathcal{H}_X} \langle g, \mu_{\mathbb{P}_Y} \rangle_{\mathcal{H}_Y} \\ &= \langle f \otimes g, \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_Y} \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} \\ &= \langle f, (\mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_Y}) g \rangle_{\mathcal{H}_X} \\ &= \langle g, (\mu_{\mathbb{P}_Y} \otimes \mu_{\mathbb{P}_X}) f \rangle_{\mathcal{H}_Y} \end{aligned}$$
- $$\begin{aligned} \mathbb{E}[f(X)g(Y)] &= \mathbb{E}[\langle f, k_X(\cdot, X) \rangle_{\mathcal{H}_X} \langle g, k_Y(\cdot, Y) \rangle_{\mathcal{H}_Y}] \\ &= \mathbb{E}[\langle f \otimes g, k_X(\cdot, X) \otimes k_Y(\cdot, Y) \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y}] \\ &= \mathbb{E}[\langle f, (k_X(\cdot, X) \otimes k_Y(\cdot, Y)) g \rangle_{\mathcal{H}_X}] \\ &= \mathbb{E}[\langle g, (k_Y(\cdot, Y) \otimes k_X(\cdot, X)) f \rangle_{\mathcal{H}_Y}] \end{aligned}$$

Covariance Operator

Let k_X and k_Y be the r.k.'s of $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively. Then

- $\mathbb{E}[f(X)] = \langle f, \mu_{\mathbb{P}_X} \rangle_{\mathcal{H}_X}$ and $\mathbb{E}[g(Y)] = \langle g, \mu_{\mathbb{P}_Y} \rangle_{\mathcal{H}_Y}$
- $$\begin{aligned} \mathbb{E}[f(X)]\mathbb{E}[g(Y)] &= \langle f, \mu_{\mathbb{P}_X} \rangle_{\mathcal{H}_X} \langle g, \mu_{\mathbb{P}_Y} \rangle_{\mathcal{H}_Y} \\ &= \langle f \otimes g, \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_Y} \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} \\ &= \langle f, (\mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_Y}) g \rangle_{\mathcal{H}_X} \\ &= \langle g, (\mu_{\mathbb{P}_Y} \otimes \mu_{\mathbb{P}_X}) f \rangle_{\mathcal{H}_Y} \end{aligned}$$
- $$\begin{aligned} \mathbb{E}[f(X)g(Y)] &= \mathbb{E}[\langle f, k_X(\cdot, X) \rangle_{\mathcal{H}_X} \langle g, k_Y(\cdot, Y) \rangle_{\mathcal{H}_Y}] \\ &= \mathbb{E}[\langle f \otimes g, k_X(\cdot, X) \otimes k_Y(\cdot, Y) \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y}] \\ &= \mathbb{E}[\langle f, (k_X(\cdot, X) \otimes k_Y(\cdot, Y)) g \rangle_{\mathcal{H}_X}] \\ &= \mathbb{E}[\langle g, (k_Y(\cdot, Y) \otimes k_X(\cdot, X)) f \rangle_{\mathcal{H}_Y}] \end{aligned}$$

Covariance Operator

- Assuming $\mathbb{E}\sqrt{k_X(X,X)k_Y(Y,Y)} < \infty$, we obtain

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \langle f, \mathbb{E}[k_X(\cdot,X) \otimes k_Y(\cdot,Y)]g\rangle_{\mathcal{H}_X} \\ &= \langle g, \mathbb{E}[k_Y(\cdot,Y) \otimes k_X(\cdot,X)]f\rangle_{\mathcal{H}_Y}\end{aligned}$$

-

$$\mathrm{Cov}(f(X), g(Y)) = \langle f, C_{XY} g\rangle_{\mathcal{H}_X} = \langle g, C_{YX} f\rangle_{\mathcal{H}_Y}$$

where

$$C_{XY} := \mathbb{E}[k_X(\cdot,X) \otimes k_Y(\cdot,Y)] - \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_Y}$$

is a cross-covariance operator from $\mathcal{H}_Y$ to $\mathcal{H}_X$ and $C_{YX} = C^*_{XY}$.

Compare to the feature space view point with canonical feature maps

Dependence Measures

- $$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) = \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} |\langle f, C_{XY} g\rangle_{\mathcal{H}_X}|$$
  $$= \|C_{XY}\|_{\text{op}} = \|C_{YX}\|_{\text{op}},$$
  which is the maximum singular value of C_{XY}.
- Choosing $k_X(\cdot, X) = \langle \cdot, X\rangle_2$ and $k_Y(\cdot, Y) = \langle \cdot, Y\rangle_2$, for Gaussian distributions,
  $$X \perp Y \Leftrightarrow C_{YX} = 0$$
- In general,
  $$X \perp Y \overset{?}{\Leftrightarrow} C_{YX} = 0.$$

Dependence Measures

- $$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) = \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} |\langle f, C_{XY} g\rangle_{\mathcal{H}_X}| = \|C_{XY}\|_{op} = \|C_{YX}\|_{op},$$

  which is the maximum singular value of C_{XY}.

- Choosing $k_X(\cdot, X) = \langle \cdot, X\rangle_2$ and $k_Y(\cdot, Y) = \langle \cdot, Y\rangle_2$, for Gaussian distributions,

  $$X \perp Y \Leftrightarrow C_{YX} = 0$$

- In general,

  $$X \perp Y \overset{?}{\Leftrightarrow} C_{YX} = 0.$$

Dependence Measures

- How about we consider other singular values?
- How about $\|C_{YX}\|^2_{HS}$, which is the sum of squared singular values of C_{YX}?

  Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., ALT 2005, JMLR 2005)
- $\|C_{YX}\|_{\text{op}} \leq \|C_{YX}\|_{HS}$

Dependence Measures

- $$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) := \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} \left|\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]\right|.$$

- How about we use different constraint, i.e., $\|f \otimes g\|_{\mathcal{H}_X \otimes \mathcal{H}_Y} \leq 1$?

$$\begin{aligned}
\sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \mathrm{Cov}(f(X), g(Y)) &= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \langle f, C_{XY} g\rangle_{\mathcal{H}_X} \\
&= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \langle f \otimes g, C_{XY}\rangle_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} = \|C_{XY}\|_{HS}
\end{aligned}$$

- $$\begin{aligned}
\|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} &= \|\mathbb{E}[k_X(\cdot, X) \otimes k_Y(\cdot, Y)] - \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_X}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \left\| \int k_X(\cdot, X) \otimes k_Y(\cdot, Y)\, d(\mathbb{P}_{XY} - \mathbb{P}_X \times \mathbb{P}_Y) \right\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= MMD_{\mathcal{H}_X\otimes\mathcal{H}_Y}(\mathbb{P}_{XY}, \mathbb{P}_X \times \mathbb{P}_Y)
\end{aligned}$$

Dependence Measures

- $$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) := \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} \left|\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]\right|.$$

- How about we use different constraint, i.e., $\|f \otimes g\|_{\mathcal{H}_X \otimes \mathcal{H}_Y} \leq 1$?

$$\begin{aligned}
\sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \mathrm{Cov}(f(X), g(Y)) &= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \langle f, C_{XY} g\rangle_{\mathcal{H}_X} \\
&= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \langle f \otimes g, C_{XY}\rangle_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} = \|C_{XY}\|_{HS}
\end{aligned}$$

- $$\begin{aligned}
\|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} &= \|\mathbb{E}[k_X(\cdot, X) \otimes k_Y(\cdot, Y)] - \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_X}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \left\| \int k_X(\cdot, X) \otimes k_Y(\cdot, Y)\, d(\mathbb{P}_{XY} - \mathbb{P}_X \times \mathbb{P}_Y) \right\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= MMD_{\mathcal{H}_X\otimes\mathcal{H}_Y}(\mathbb{P}_{XY}, \mathbb{P}_X \times \mathbb{P}_Y)
\end{aligned}$$

Dependence Measures

- $$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) := \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} \left|\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]\right|.$$

- How about we use different constraint, i.e., $\|f \otimes g\|_{\mathcal{H}_X \otimes \mathcal{H}_Y} \leq 1$?

$$\begin{aligned}\sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \operatorname{Cov}(f(X), g(Y)) &= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \langle f, C_{XY} g\rangle_{\mathcal{H}_X} \\ &= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \langle f \otimes g, C_{XY}\rangle_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\ &= \|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} = \|C_{XY}\|_{HS}\end{aligned}$$

- $$\begin{aligned}\|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} &= \|\mathbb{E}[k_X(\cdot, X) \otimes k_Y(\cdot, Y)] - \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_X}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\ &= \left\| \int k_X(\cdot, X) \otimes k_Y(\cdot, Y)\, d(\mathbb{P}_{XY} - \mathbb{P}_X \times \mathbb{P}_Y)\right\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\ &= MMD_{\mathcal{H}_X\otimes\mathcal{H}_Y}(\mathbb{P}_{XY}, \mathbb{P}_X \times \mathbb{P}_Y)\end{aligned}$$

Dependence Measures

- $\mathcal{H}_X \otimes \mathcal{H}_Y$ is an RKHS with kernel $k_X k_Y$.
- If $k_X k_Y$ is characteristic, then

$$\|C_{XY}\|_{\mathcal{H}_X \otimes \mathcal{H}_Y} = 0 \Leftrightarrow \mathbb{P}_{XY} = \mathbb{P}_X \times \mathbb{P}_Y \Leftrightarrow X \perp Y$$

- If k_X and k_Y are characteristic, then

$$\|C_{XY}\|_{HS} = 0 \Leftrightarrow X \perp Y.$$

(Zoltan & S., 2018)

- Using the reproducing property,

$$\begin{aligned}\|C_{XY}\|_{HS}^2 = \mathbb{E}_{XY}\mathbb{E}_{X'Y'} k_X(X,X') k_Y(Y,Y') \\ + \mathbb{E}_{XX'} k_X(X,X') \mathbb{E}_{YY'} k_Y(Y,Y') \\ -2 \cdot \mathbb{E}_{X'Y'} \left[\mathbb{E}_X k_X(X,X') \mathbb{E}_Y k_Y(Y,Y')\right]\end{aligned}$$

- Can be estimated using a V-statistic (empirical sums).

Applications

- Two-sample testing (Gretton et al., NIPS 2006, JMLR 2012; Harchaoui et al., NIPS 2008)
- Goodness-of-fit testing (Balasubramanian et al., 2017)
- Independence testing (Gretton et al., NIPS 2008)
- Conditional independence testing (Fukumizu et al., NIPS 2008)
- Supervised dimensionality reduction (Fukumizu et al., JMLR 2004)
- Kernel Bayes rule (filtering, prediction and smoothing) (Fukumizu et al., JMLR 2013)
- Distribution regression (Szabó et al., JMLR 2016)
- Kernel CCA (Fukumizu et al., JMLR 2007),....

Review paper (Muandet et al., 2017)

Application: Two-Sample Testing

Two-Sample Problem

- Given random samples $\{X_1, \ldots, X_m\} \stackrel{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_1, \ldots, Y_n\} \stackrel{i.i.d.}{\sim} \mathbb{Q}$.
- Determine: $\mathbb{P} = \mathbb{Q}$ or $\mathbb{P} \neq \mathbb{Q}$?
- Approach:

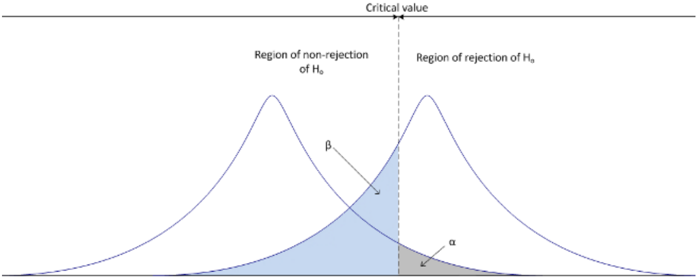
$$\begin{array}{c} H_0 : \mathbb{P} = \mathbb{Q} \\ H_1 : \mathbb{P} \neq \mathbb{Q} \end{array} \equiv \begin{array}{c} H_0 : MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = 0 \\ H_1 : MMD_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) > 0 \end{array}$$

- If $MMD^2_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_n)$ is
  - far from zero: reject H_0
  - close to zero: accept H_0

Type-I and Type-II Errors

Statistical decision	Truth: Null hypothesis true	Truth: Null hypothesis false
Reject null hypothesis	Type I error	Correct (power)
Do not reject null hypothesis	Correct	Type II error

- Given $\mathbb{P} = \mathbb{Q}$, want threshold or critical value $t_{1-\alpha}$ such that $\Pr_{H_0}(MMD^2_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_n) > t_{1-\alpha}) \leq \alpha$.



Statistical Test: Large Deviation Bounds

- Given $\mathbb{P} = \mathbb{Q}$, want threshold t such that $\Pr_{H_0}(MMD^2_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_n) > t) \leq \alpha$.
- We showed that (S et al., EJS 2012)

$$\Pr\Big(\big|MMD^2_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_n) - MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})\big| \geq \sqrt{\tfrac{2(m+n)}{mn}}\Big(1 + \sqrt{2\log\tfrac{1}{\alpha}}\Big)\Big) \leq \alpha.$$

- α-level test: Accept H_0 if

$$MMD^2_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_n) < \sqrt{\frac{2(m+n)}{mn}}\left(1 + \sqrt{2\log\frac{1}{\alpha}}\right)$$

Otherwise reject.

Too conservative!!

Statistical Test: Large Deviation Bounds

- Given $\mathbb{P} = \mathbb{Q}$, want threshold t such that $\Pr_{H_0}(MMD^2_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_n) > t) \leq \alpha$.
- We showed that (S et al., EJS 2012)

$$\Pr\left(\left| MMD^2_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_n) - MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) \right| \geq \sqrt{\tfrac{2(m+n)}{mn}} \left(1 + \sqrt{2 \log \tfrac{1}{\alpha}}\right)\right) \leq \alpha.$$

- α-level test: Accept H_0 if

$$MMD^2_{\mathcal{H}}(\mathbb{P}_m, \mathbb{Q}_n) < \sqrt{\frac{2(m+n)}{mn}} \left(1 + \sqrt{2 \log \frac{1}{\alpha}}\right)$$

Otherwise reject.

Too conservative!!

Statistical Test: Asymptotic Distribution (Gretton et al., NIPS 2006, JMLR 2012)

Unbiased estimator of $MMD^2_{\mathcal{H}}(\mathbb{P},\mathbb{Q})$: U-statistic

$$\widehat{MMD}^2_{\mathcal{H}} := \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \underbrace{k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(X_j, Y_i)}_{h((X_i,Y_i),(X_j,Y_j))}$$

- Under H_0,

$$m\,\widehat{MMD}^2_{\mathcal{H}} \xrightarrow{w} \sum_{i=1}^{\infty} \lambda_i \left(\theta_i^2 - 2\right) \quad \text{as } n \to \infty,$$

where $\theta_i \sim \mathcal{N}(0,2)$ i.i.d., and λ_i are solutions to

$$\int_{\mathcal{X}} \underbrace{\widetilde{k}(x,y)}_{\text{centered}} \psi_i(x)\, d\mathbb{P}(x) = \lambda_i \psi_i(y)$$

- Consistent (Type-II error goes to zero): Under H_1,

$$\sqrt{m}\left(\widehat{MMD}^2_{\mathcal{H}} - MMD^2_{\mathcal{H}}(\mathbb{P},\mathbb{Q})\right) \xrightarrow{w} \mathcal{N}(0, \sigma_h^2) \quad \text{as } n \to \infty.$$

Statistical Test: Asymptotic Distribution (Gretton et al., NIPS 2006, JMLR 2012)

Unbiased estimator of $MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})$: U-statistic

$$\widehat{MMD^2_{\mathcal{H}}} := \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \underbrace{k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(X_j, Y_i)}_{h((X_i, Y_i), (X_j, Y_j))}$$

- Under H_0,

$$m\,\widehat{MMD^2_{\mathcal{H}}} \xrightarrow{w} \sum_{i=1}^{\infty} \lambda_i \left(\theta_i^2 - 2\right) \quad \text{as } n \to \infty,$$

  where $\theta_i \sim \mathcal{N}(0, 2)$ i.i.d., and λ_i are solutions to

$$\int_{\mathcal{X}} \underbrace{\widetilde{k}(x, y)}_{\text{centered}} \psi_i(x)\, d\mathbb{P}(x) = \lambda_i \psi_i(y)$$

- Consistent (Type-II error goes to zero): Under H_1,

$$\sqrt{m}\left(\widehat{MMD^2_{\mathcal{H}}} - MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})\right) \xrightarrow{w} \mathcal{N}(0, \sigma_h^2) \quad \text{as } n \to \infty.$$

Statistical Test: Asymptotic Distribution (Gretton et al., NIPS 2006, JMLR 2012)

Unbiased estimator of $MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q})$: U-statistic

$$\widehat{MMD}^2_{\mathcal{H}} := \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \underbrace{k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(X_j, Y_i)}_{h((X_i, Y_i), (X_j, Y_j))}$$

- Under H_0,

$$m \, \widehat{MMD}^2_{\mathcal{H}} \xrightarrow{w} \sum_{i=1}^{\infty} \lambda_i \left(\theta_i^2 - 2\right) \quad \text{as } n \to \infty,$$

  where $\theta_i \sim \mathcal{N}(0, 2)$ i.i.d., and λ_i are solutions to

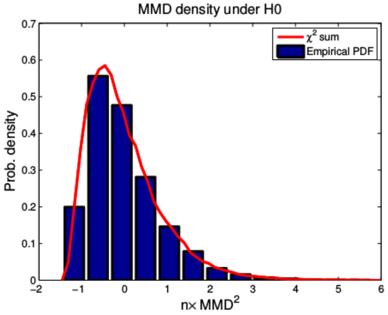
$$\int_{\mathcal{X}} \underbrace{\widetilde{k}(x, y)}_{\text{centered}} \psi_i(x) \, d\mathbb{P}(x) = \lambda_i \psi_i(y)$$

- Consistent (Type-II error goes to zero): Under H_1,

$$\sqrt{m} \left(\widehat{MMD}^2_{\mathcal{H}} - MMD^2_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) \right) \xrightarrow{w} \mathcal{N}(0, \sigma_h^2) \quad \text{as } n \to \infty.$$

Statistical Test: Asymptotic Distribution (Gretton et al., NIPS 2006, JMLR 2012)

- α-level test: Estimate $1 - \alpha$ quantile of the null distribution using bootstrap.



Computationally intensive!!

Picture credit: A. Gretton

Statistical Test Without Bootstrap (Gretton et al., NIPS 2009)

- Estimate the eigenvalues, λ_i from combined samples
  - Define $Z := (X_1, \ldots, X_m, Y_1, \ldots, Y_m)$
  - $K_{ij} := k(Z_i, Z_j)$
  - Compute the eigenvalues, $\widehat{\lambda}_i$ of

$$\widetilde{K} = HKH$$

  where $H = I - \frac{1}{2m} 1_{2m} 1_{2m}^T$

- α-level test: Compute the $1 - \alpha$ quantile of the distribution associated with

$$\sum_{i=1}^{2m} \widehat{\lambda}_i \left(\theta_i^2 - 2\right)$$

- Test is asymptotically α-level consistent

Experiments (Gretton et al., NIPS 2009)

- Comparison example: Canadian Hansard corpus (agriculture, fisheries and immigration)
- Samples: 5-line extracts
- Kernel: k-spectrum kernel with $k = 10$
- Sample size: 10
- Repetitions: 300
- Compute $\widehat{MMD}^2_{\mathcal{H}}$

k-spectrum kernel: average Type II error 0 ($\alpha = 0.05$)

Bag of words kernel: average Type II error 0.18

First ever test on structured data

Choice of Characteristic Kernel

Choice of Characteristic Kernels

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

- $MMD_{\mathcal{H}_\sigma}$ is a function of σ.
- So $MMD_{\mathcal{H}_\sigma}$ is a family of metrics. Which one should we use in practice?
- Note that $MMD_{\mathcal{H}_\sigma} \to 0$ as $\sigma \to 0$ or $\sigma \to \infty$.

Therefore, the kernel choice is very critical in applications.

Heuristics:

- Median: $\sigma = \text{median}\left(\|X_i^* - X_j^*\|_2 : i \neq j,\ i, j = 1, \ldots, m\right)$ where $X^* = ((X_i)_i, (Y_i)_i)$ (Gretton et al., NIPS 2006, NIPS 2009, JMLR 2012).
- Choose the test statistic to be $MMD_{\mathcal{H}_{\sigma^*}}(\mathbb{P}_m, \mathbb{Q}_m)$ where

$$\sigma^* = \arg \max_{\sigma \in (0,\infty)} MMD_{\mathcal{H}_\sigma}(\mathbb{P}_m, \mathbb{Q}_m)$$

(S et al., NIPS 2009)

Choice of Characteristic Kernels

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

- $MMD_{\mathcal{H}_\sigma}$ is a function of σ.
- So $MMD_{\mathcal{H}_\sigma}$ is a family of metrics. Which one should we use in practice?
- Note that $MMD_{\mathcal{H}_\sigma} \to 0$ as $\sigma \to 0$ or $\sigma \to \infty$.

Therefore, the kernel choice is very critical in applications.

Heuristics:

- Median: $\sigma = \text{median}\left(\|X_i^* - X_j^*\|_2 : i \neq j,\ i, j = 1, \ldots, m\right)$ where $X^* = ((X_i)_i, (Y_i)_i)$ (Gretton et al., NIPS 2006, NIPS 2009, JMLR 2012).
- Choose the test statistic to be $MMD_{\mathcal{H}_{\sigma^*}}(\mathbb{P}_m, \mathbb{Q}_m)$ where

$$\sigma^* = \arg \max_{\sigma \in (0,\infty)} MMD_{\mathcal{H}_\sigma}(\mathbb{P}_m, \mathbb{Q}_m)$$

(S et al., NIPS 2009)

Choice of Characteristic Kernels

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

- $MMD_{\mathcal{H}_\sigma}$ is a function of σ.
- So $MMD_{\mathcal{H}_\sigma}$ is a family of metrics. Which one should we use in practice?
- Note that $MMD_{\mathcal{H}_\sigma} \to 0$ as $\sigma \to 0$ or $\sigma \to \infty$.

Therefore, the kernel choice is very critical in applications.

Heuristics:

- Median: $\sigma = \text{median}\left(\|X_i^* - X_j^*\|_2 : i \neq j,\ i, j = 1, \ldots, m\right)$ where $X^* = ((X_i)_i, (Y_i)_i)$ (Gretton et al., NIPS 2006, NIPS 2009, JMLR 2012).
- Choose the test statistic to be $MMD_{\mathcal{H}_{\sigma^*}}(\mathbb{P}_m, \mathbb{Q}_m)$ where

$$\sigma^* = \arg \max_{\sigma \in (0,\infty)} MMD_{\mathcal{H}_\sigma}(\mathbb{P}_m, \mathbb{Q}_m)$$

(S et al., NIPS 2009)

Choice of Characteristic Kernels

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

- $MMD_{\mathcal{H}_\sigma}$ is a function of σ.
- So $MMD_{\mathcal{H}_\sigma}$ is a family of metrics. Which one should we use in practice?
- Note that $MMD_{\mathcal{H}_\sigma} \to 0$ as $\sigma \to 0$ or $\sigma \to \infty$.

Therefore, the kernel choice is very critical in applications.

Heuristics:

- Median: $\sigma = \text{median}\left(\|X_i^* - X_j^*\|_2 : i \neq j,\ i, j = 1, \ldots, m\right)$ where $X^* = ((X_i)_i, (Y_i)_i)$ (Gretton et al., NIPS 2006, NIPS 2009, JMLR 2012).
- Choose the test statistic to be $MMD_{\mathcal{H}_{\sigma^*}}(\mathbb{P}_m, \mathbb{Q}_m)$ where

$$\sigma^* = \arg \max_{\sigma \in (0,\infty)} MMD_{\mathcal{H}_\sigma}(\mathbb{P}_m, \mathbb{Q}_m)$$

(S et al., NIPS 2009)

Classes of Characteristic Kernels (S et al., NIPS 2009)

More generally, we use

$$MMD(\mathbb{P},\mathbb{Q}) := \sup_{k\in\mathcal{K}} MMD_{\mathcal{H}_k}(\mathbb{P},\mathbb{Q}).$$

Examples for $\mathcal{K}$:

- $\mathcal{K}_g := \{e^{-\sigma\|x-y\|_2^2}, x,y\in\mathbb{R}^d \,:\, \sigma\in\mathbb{R}_+\}$.
- $\mathcal{K}_{lin} := \{k_\lambda = \sum_{i=1}^{\ell}\lambda_i k_i | k_\lambda \text{ is pd}, \sum_{i=1}^{\ell}\lambda_i = 1\}$.
- $\mathcal{K}_{con} := \{k_\lambda = \sum_{i=1}^{\ell}\lambda_i k_i | \lambda_i \geq 0, \sum_{i=1}^{\ell}\lambda_i = 1\}$.

Test:

- α-level test: Estimate $1-\alpha$ quantile of the null distribution of $MMD(\mathbb{P}_m,\mathbb{Q}_m)$ using bootstrap.
- Test consistency: Based on the functional central limit theorem for U-processes indexed by VC-subgraph $\mathcal{K}$.

Computational disadvantage!!

Classes of Characteristic Kernels (S et al., NIPS 2009)

More generally, we use

$$MMD(\mathbb{P},\mathbb{Q}) := \sup_{k\in\mathcal{K}} MMD_{\mathcal{H}_k}(\mathbb{P},\mathbb{Q}).$$

Examples for $\mathcal{K}$:

- $\mathcal{K}_g := \{e^{-\sigma\|x-y\|_2^2},\ x,y \in \mathbb{R}^d\ :\ \sigma \in \mathbb{R}_+\}$.
- $\mathcal{K}_{lin} := \{k_\lambda = \sum_{i=1}^{\ell} \lambda_i k_i | k_\lambda \text{ is pd}, \sum_{i=1}^{\ell} \lambda_i = 1\}$.
- $\mathcal{K}_{con} := \{k_\lambda = \sum_{i=1}^{\ell} \lambda_i k_i | \lambda_i \geq 0, \sum_{i=1}^{\ell} \lambda_i = 1\}$.

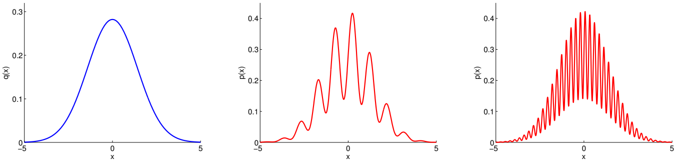
Test:

- α-level test: Estimate $1-\alpha$ quantile of the null distribution of $MMD(\mathbb{P}_m,\mathbb{Q}_m)$ using bootstrap.
- Test consistency: Based on the functional central limit theorem for U-processes indexed by VC-subgraph $\mathcal{K}$.

Computational disadvantage!!

Experiments

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

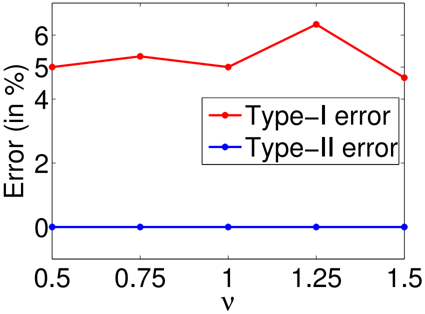


$\nu = 0$ $\quad$ $\nu = 2$ $\quad$ $\nu = 7.5$

- $k(x, y) = \exp(-(x - y)^2/\sigma)$.
- Test statistics: $MMD(\mathbb{P}_m, \mathbb{Q}_m)$ and $MMD_{\mathcal{H}_\sigma}(\mathbb{P}_m, \mathbb{Q}_m)$ for various σ.

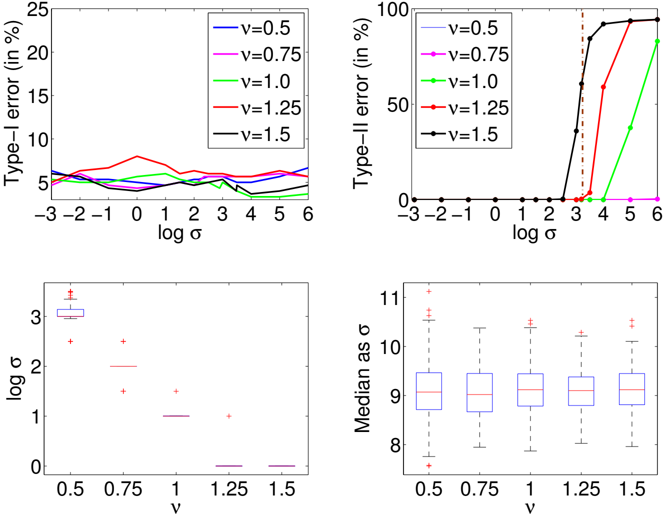
Experiments

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



Experiments

$$MMD_{\mathcal{H}_\sigma}(\mathbb{P}, \mathbb{Q})$$



Choice of Characteristic Kernels (Gretton et al., NIPS 2012)

- Choose a kernel that minimizes the Type-II error for a given Type-I error:

$$k^* \in \arg \inf_{k \in \mathcal{K}:\, Type_I(k) \leq \alpha} Type_{II}(k).$$

- Not easy to compute with the asymptotic distributions of the U-statistic, $\widehat{MMD}^2_{\mathcal{H}_k}(\mathbb{P}_m, \mathbb{Q}_m)$.

- Modified statistic: Average of U-statistics computed on independent blocks of size 2.

$$\widetilde{MMD}^2_{\mathcal{H}_k}(\mathbb{P}_m, \mathbb{Q}_m) = \frac{2}{m} \sum_{i=1}^{m/2} \underbrace{k(X_{2i-1}, X_{2i}) + k(Y_{2i-1}, Y_{2i}) - k(X_{2i-1}, Y_{2i}) - k(Y_{2i-1}, X_{2i})}_{h_k(Z_i)},$$

where $Z_i = (X_{2i-1}, X_{2i}, Y_{2i-1}, Y_{2i})$.

- Recall

$$\widehat{MMD}^2_{\mathcal{H}} := \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \underbrace{k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(X_j, Y_i)}_{h((X_i, Y_i), (X_j, Y_j))}$$

Choice of Characteristic Kernels (Gretton et al., NIPS 2012)

- Choose a kernel that minimizes the Type-II error for a given Type-I error:

$$k^* \in \arg \inf_{k \in \mathcal{K}: Type_I(k) \leq \alpha} Type_{II}(k).$$

- Not easy to compute with the asymptotic distributions of the U-statistic, $\widehat{MMD}^2_{\mathcal{H}_k}(\mathbb{P}_m, \mathbb{Q}_m)$.

- Modified statistic: Average of U-statistics computed on independent blocks of size 2.

$$\widetilde{MMD}^2_{\mathcal{H}_k}(\mathbb{P}_m, \mathbb{Q}_m) = \frac{2}{m} \sum_{i=1}^{m/2} \underbrace{k(X_{2i-1}, X_{2i}) + k(Y_{2i-1}, Y_{2i}) - k(X_{2i-1}, Y_{2i}) - k(Y_{2i-1}, X_{2i})}_{h_k(Z_i)},$$

where $Z_i = (X_{2i-1}, X_{2i}, Y_{2i-1}, Y_{2i})$.

- Recall

$$\widehat{MMD}^2_{\mathcal{H}} := \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \underbrace{k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(X_j, Y_i)}_{h((X_i, Y_i), (X_j, Y_j))}$$

Modified Statistic

Advantages:

- $\widetilde{MMD}^2_{\mathcal{H}}$ is computable in $O(m)$ while $\widehat{MMD}^2_{\mathcal{H}}$ requires $O(m^2)$ computations.
- Under H_0,
$$\sqrt{m}\,\widetilde{MMD}^2_{\mathcal{H}_k}(\mathbb{P}_m, \mathbb{Q}_m) \xrightarrow{w} \mathcal{N}(0, 2\sigma^2_{h_k}),$$
where $\sigma^2_{h_k} = \mathbb{E}_Z h^2_k(Z) - (\mathbb{E}_Z h_k(Z))^2$ assuming $0 < \mathbb{E}_Z h^2_k(Z) < \infty$.
- The asymptotic distribution is normal as against weighted sum of infinite χ^2. Therefore, the test threshold is easy to compute.

Disadvantages:

- Larger variance
- Smaller power

Modified Statistic

Advantages:

- $\widetilde{MMD}^2_{\mathcal{H}}$ is computable in $O(m)$ while $\widehat{MMD}^2_{\mathcal{H}}$ requires $O(m^2)$ computations.
- Under H_0,

$$\sqrt{m}\,\widetilde{MMD}^2_{\mathcal{H}_k}(\mathbb{P}_m, \mathbb{Q}_m) \xrightarrow{w} \mathcal{N}(0, 2\sigma^2_{h_k}),$$

  where $\sigma^2_{h_k} = \mathbb{E}_Z h_k^2(Z) - (\mathbb{E}_Z h_k(Z))^2$ assuming $0 < \mathbb{E}_Z h_k^2(Z) < \infty$.

- The asymptotic distribution is normal as against weighted sum of infinite χ^2. Therefore, the test threshold is easy to compute.

Disadvantages:

- Larger variance
- Smaller power

Type-I and Type-II Errors

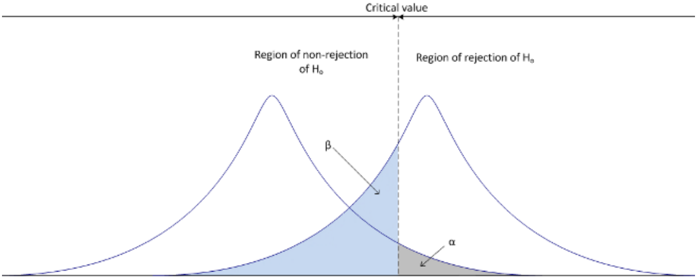
- Test threshold: For a given k and α,

$$t_{k,1-\alpha} = \sqrt{2}\sigma_{h_k}\Phi_N^{-1}(1-\alpha)$$

where Φ_N is the cdf of $\mathcal{N}(0,1)$.

- Type-II error:

$$\Phi_N\left(\Phi_N^{-1}(1-\alpha) - \frac{MMD^2_{\mathcal{H}_k}(\mathbb{P},\mathbb{Q})\sqrt{m}}{\sqrt{2}\sigma_{h_k}}\right)$$



Best Kernel: Minimizes Type-II Error

- Since Φ_N is a strictly increasing function, the Type-II error is minimized by maximizing $\frac{MMD^2_{\mathcal{H}_k}(\mathbb{P},\mathbb{Q})}{\sigma_{h_k}}$.
- Optimal kernel:

$$k^* \in \arg\sup_{k \in \mathcal{K}} \frac{MMD^2_{\mathcal{H}_k}(\mathbb{P},\mathbb{Q})}{\sigma_{h_k}}.$$

- Since $MMD^2_{\mathcal{H}_k}$ and σ_{h_k} depend on unknown $\mathbb{P}$ and $\mathbb{Q}$, we split the data into train and test data to estimate k^* on the train data as $\hat{k}^*$ and evaluate the threshold $t_{\hat{k}^*,1-\alpha}$ on the test data.

Data-Dependent Kernel

- Train data: $\widetilde{MMD}^2_{\mathcal{H}_k}$ and $\hat{\sigma}_{h_k}$.
- Define

$$\hat{k}^* \in \arg\sup_{k \in \mathcal{K}} \frac{\widetilde{MMD}^2_{\mathcal{H}_k}}{\hat{\sigma}_{h_k} + \lambda_m}$$

  for some $\lambda_m \to 0$ as $m \to \infty$.
- Test data: $\widetilde{MMD}^2_{\mathcal{H}_{\hat{k}^*}}$, $\hat{\sigma}_{h_{\hat{k}^*}}$ and $t_{\hat{k}^*, 1-\alpha}$.
- If $\widetilde{MMD}^2_{\mathcal{H}_{\hat{k}^*}} > t_{\hat{k}^*, 1-\alpha}$, reject H_0, else accept.

Similar results are recently obtained for $\widehat{MMD}^2_{\mathcal{H}_k}$ (Sutherland et al., ICLR 2017)

Data-Dependent Kernel

- Train data: $\widetilde{MMD}^2_{\mathcal{H}_k}$ and $\hat{\sigma}_{h_k}$.
- Define
$$\hat{k}^* \in \arg\sup_{k \in \mathcal{K}} \frac{\widetilde{MMD}^2_{\mathcal{H}_k}}{\hat{\sigma}_{h_k} + \lambda_m}$$
  for some $\lambda_m \to 0$ as $m \to \infty$.
- Test data: $\widetilde{MMD}^2_{\mathcal{H}_{\hat{k}^*}}$, $\hat{\sigma}_{h_{\hat{k}^*}}$ and $t_{\hat{k}^*,1-\alpha}$.
- If $\widetilde{MMD}^2_{\mathcal{H}_{\hat{k}^*}} > t_{\hat{k}^*,1-\alpha}$, reject H_0, else accept.

Similar results are recently obtained for $\widehat{MMD}^2_{\mathcal{H}_k}$ (Sutherland et al., ICLR 2017)

Advanced Topics

- Consistency of kernel CCA (Fukumizu et al., JMLR 2007)
- Convergence rates for kernel-based hypothesis tests (Balasubramanian et al., 2017)
- Conditional covariance operators and applications

Questions

References I

Balasubramanian, K., Li, T., and Yuan, M. (2017).
On the optimality of the kernel-embedding based goodness-of-fit tests.
http://arxiv.org/abs/1709.08148.

Dudley, R. M. (2002).
Real Analysis and Probability.
Cambridge University Press, Cambridge, UK.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007).
Statistical consistency of kernel canonical correlation analysis.
Journal of Machine Learning Research, 8:361–383.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004).
Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces.
Journal of Machine Learning Research, 5:73–99.

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., Song, L., and Gretton, A. (2013).
Kernel Bayes' rule: Bayesian inference with positive definite kernels.
Journal of Machine Learning Research, 14:3753–3783.

Fukumizu, K., Sriperumbudur, B. K., Gretton, A., and Schölkopf, B. (2009).
Characteristic kernels on groups and semigroups.
In *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 *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. (2012a).
A kernel two-sample test.
Journal of Machine Learning Research, 13:723–773.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a).
Measuring statistical dependence with Hilbert-Schmidt norms.
In Jain, S., Simon, H. U., and Tomita, E., editors, *Proceedings of Algorithmic Learning Theory*, pages 63–77, Berlin. Springer-Verlag.

References II

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2010).
A fast, consistent kernel two-sample test.
In *Advances in Neural Information Processing Systems 22*, Cambridge, MA. MIT Press.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005b).
Kernel methods for measuring independence.
Journal of Machine Learning Research, 6:2075–2129.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. (2005c).
Kernel constrained covariance for dependence measurement.
In Ghahramani, Z. and Cowell, R., editors, *Proc. 10^{th} International Workshop on Artificial Intelligence and Statistics*, pages 1–8.

Gretton, A., Sriperumbudur, B., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., and Fukumizu, K. (2012b).
Optimal kernel choice for large-scale two-sample tests.
In *Advances in Neural Information Processing Systems 24*, Cambridge, MA. MIT Press.

Harchaoui, Z., Bach, F. R., and Moulines, E. (2008).
Testing for homogeneity with kernel Fisher discriminant analysis.
In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 609–616. Curran Associates, Inc.

Muandet, K., Fukumizu, K., Sriperumbudur, B. K., and Schölkopf, B. (2017).
Kernel mean embedding of distributions: A review and beyond, volume 10.
Foundations and Trends in Machine Learning.

Muandet, K., Sriperumbudur, B. K., Fukumizu, K., Gretton, A., and Schölkopf, B. (2016).
Kernel mean shrinkage estimators.
Journal of Machine Learning Research, 17(48):1–41.

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

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. (2015).
On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions.
In *Proc. of 29th AAAI Conference on Artificial Intelligence*, pages 3571–3577.

Simon-Gabriel, C. and Schölkopf, B. (2016).
Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions.
arXiv:1604.05251.

References III

Smola, A. J., Gretton, A., Song, L., and Schölkopf, B. (2007).
A Hilbert space embedding for distributions.
In *Proc. 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, Berlin, Germany.

Sriperumbudur, B. K. (2016).
On the optimal estimation of probability measures in weak and strong topologies.
Bernoulli, 22(3):1839–1893.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2012).
On the empirical estimation of integral probability metrics.
Electronic Journal of Statistics, 6:1550–1599.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. (2011).
Universality, characteristic kernels and RKHS embedding of measures.
Journal of Machine Learning Research, 12:2389–2410.

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.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. (2010).
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.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017).
Generative models and model criticism via optimized maximum mean discrepancy.
In *International Conference on Learning Representations*.

Szabo, Z. and Sriperumbudur, B. K. (2018).
Characteristic and universal tensor product kernels.
Journal of Machine Learning Research, 18(233):1–29.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016).
Learning theory for distribution regression.
Journal of Machine Learning Research, 17:1–40.

References IV

Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. (2016).
Minimax estimation of kernel mean embeddings.
arXiv:1602.04361.