Machine Learning with Kernel Methods



Michael Arbel, Julien Mairal & Jean-Philippe Vert

firstname.lastname@m4x.org

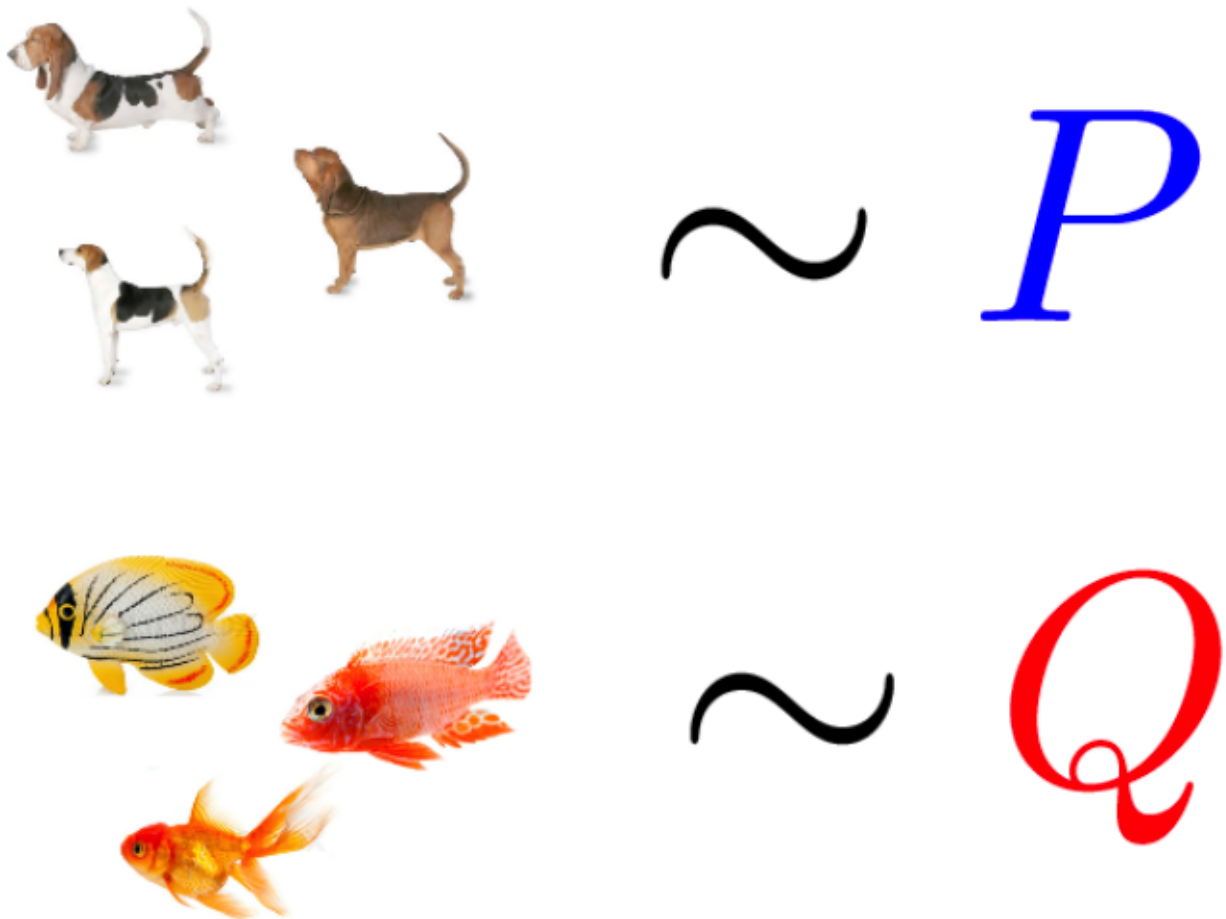
Last update: Jan 2022

Introduction

- We have seen how to represent each individual data-point by an embedding in some feature space.
- This allows to compare data points by evaluating the kernel.
- Now we are interested in comparing two or more *sets* of data-points, or more generally *distributions* of data points.

Motivation I: Comparing two distributions

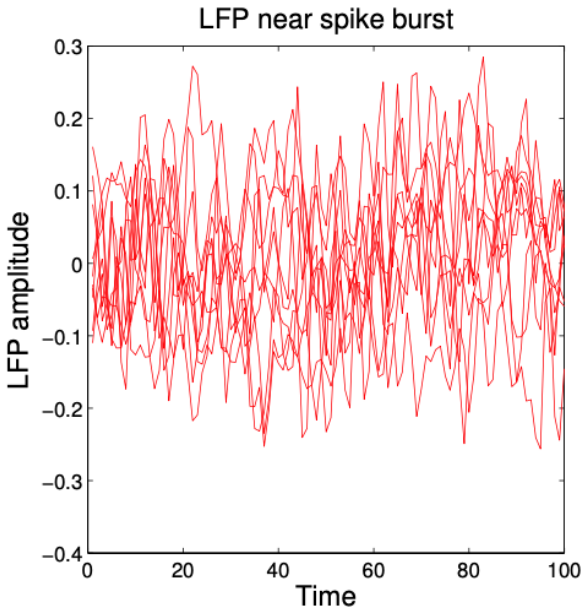
- Data: Samples from unknown distributions $\mathbb{P}$ and $\mathbb{Q}$.
- Goal: do $\mathbb{P}$ and $\mathbb{Q}$ differ?

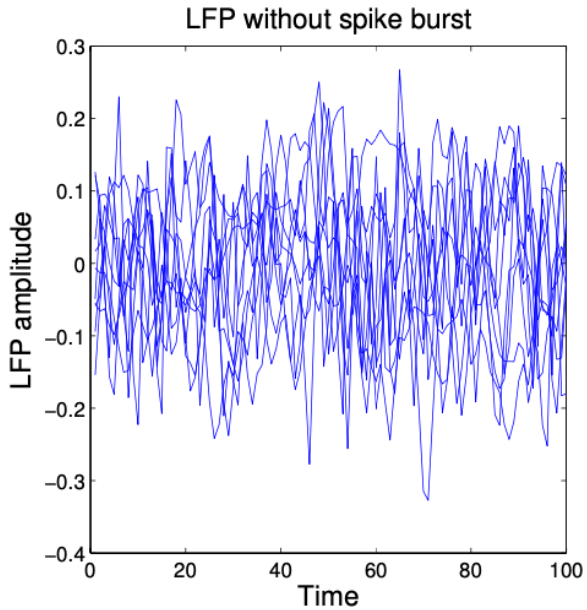


Differences between dogs and fish.

Motivation I: Comparing two distributions

- Data: Samples from unknown distributions $\mathbb{P}$ and $\mathbb{Q}$.
- Goal: do $\mathbb{P}$ and $\mathbb{Q}$ differ?

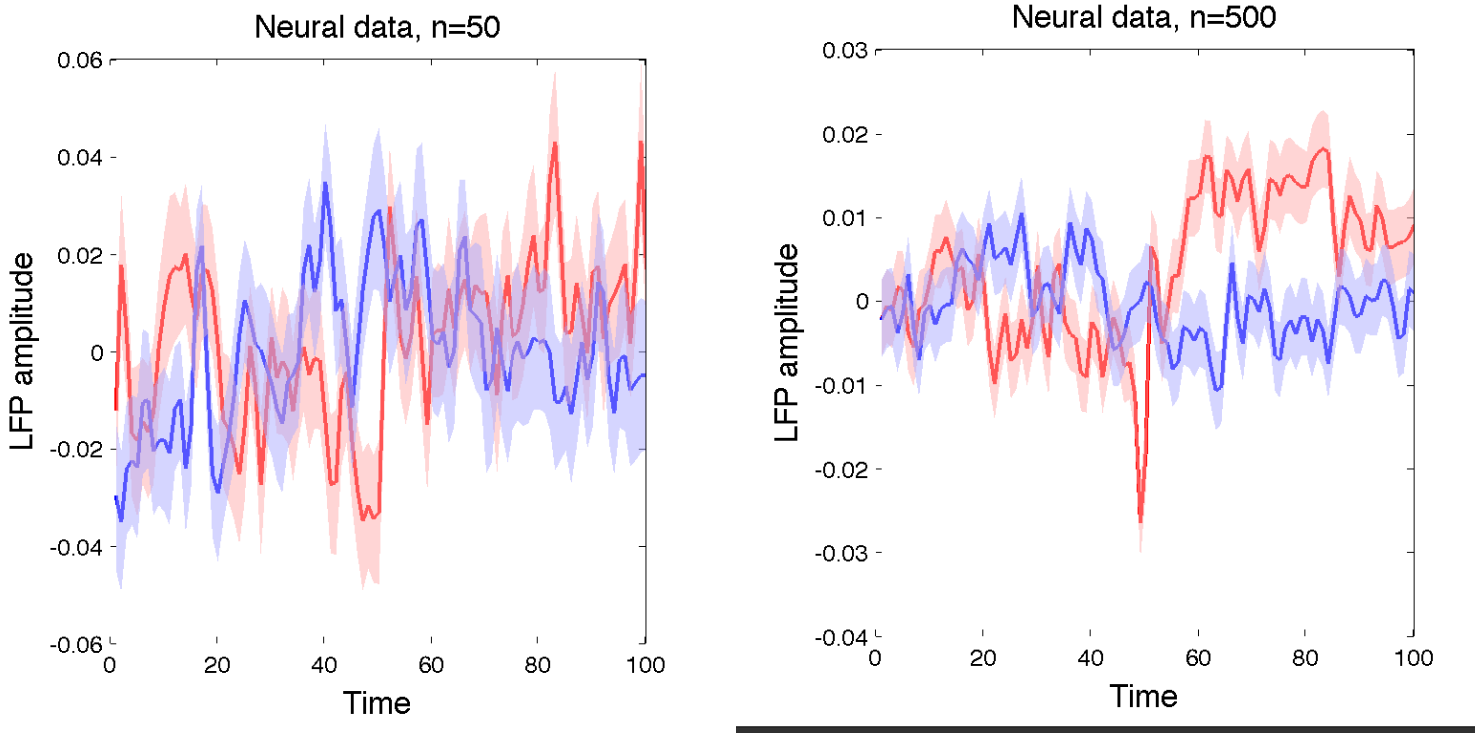




Difference in brain signals: Do local field potential (LFP) signals change when measured near a spike burst?

Motivation I: Comparing two distributions

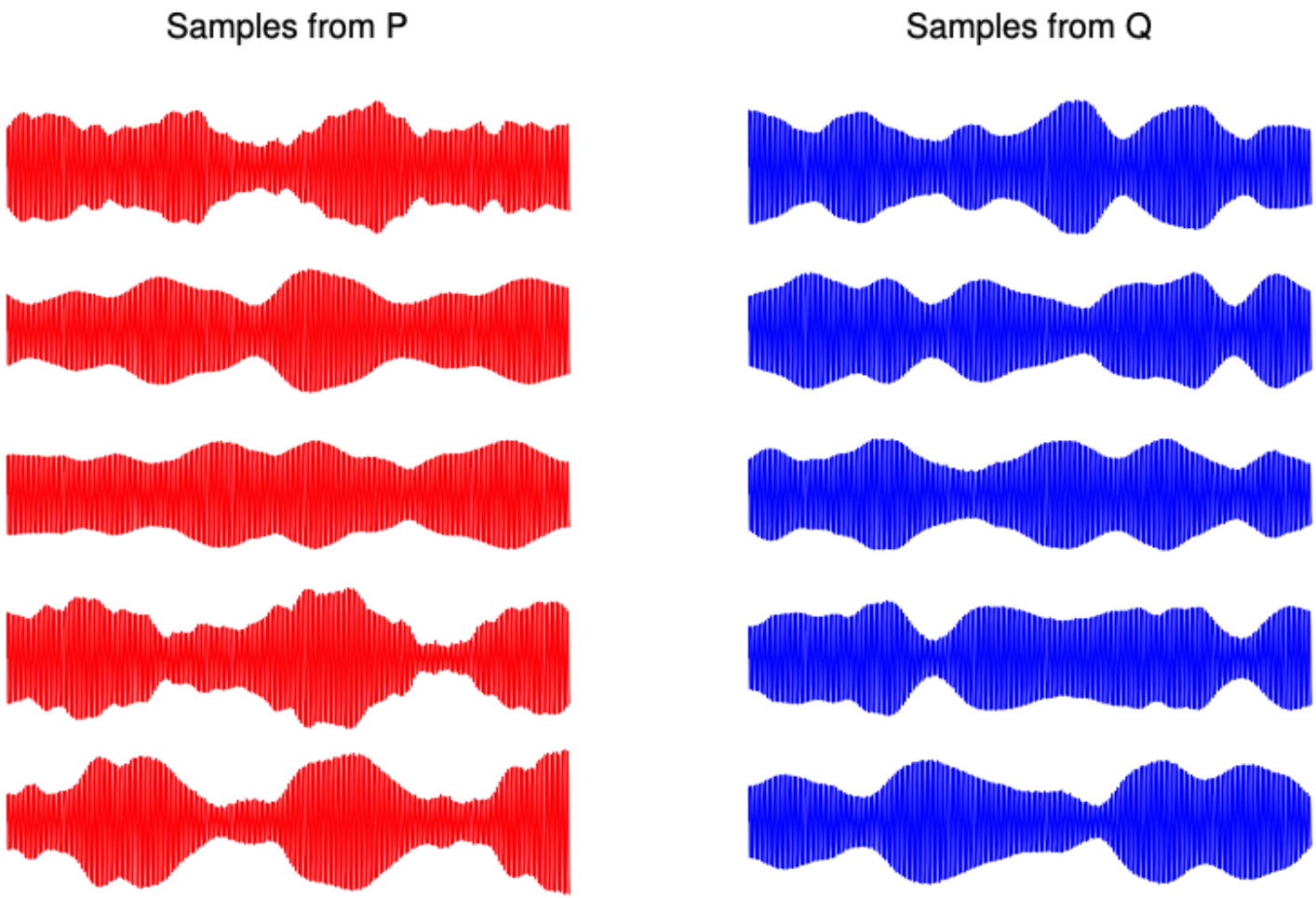
- Data: Samples from unknown distributions $\mathbb{P}$ and $\mathbb{Q}$.
- Goal: do $\mathbb{P}$ and $\mathbb{Q}$ differ?



Difference in brain signals: Do local field potential (LFP) signals change when measured near a spike burst?

Motivation I: Comparing two distributions

- Data: Samples from unknown distributions $\mathbb{P}$ and $\mathbb{Q}$.
- Goal: do $\mathbb{P}$ and $\mathbb{Q}$ differ?



Motivation II: Detecting dependence

X_1: Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

X_2: No doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development.

. . .

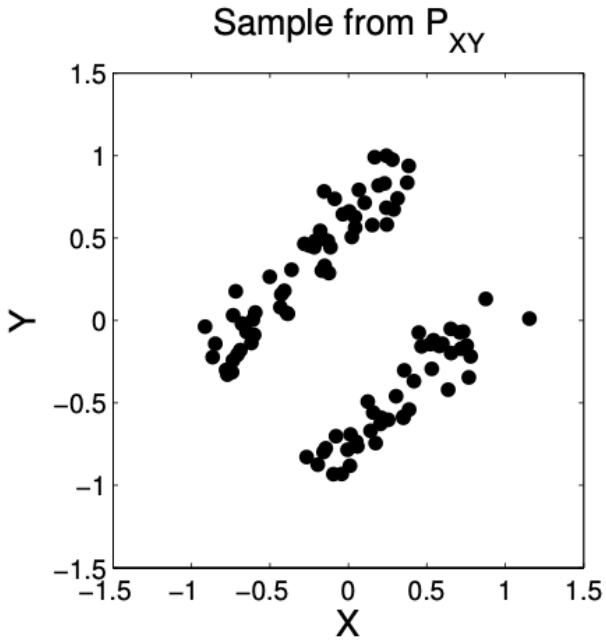
Y_1: Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat et concerne l'aide financiére qu'on a annoncée pour les agriculteurs. La plupart des agriculteurs n'ont encore rien reu de cet argent.

Y_2: Il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes enfants.

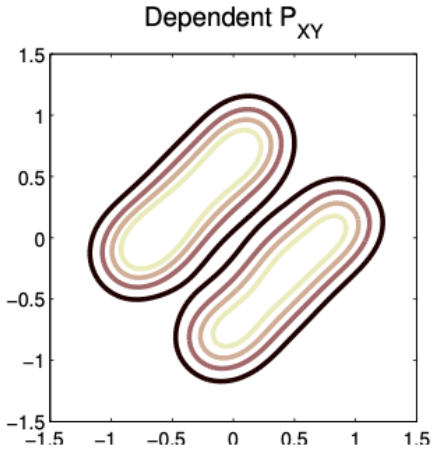
. . .

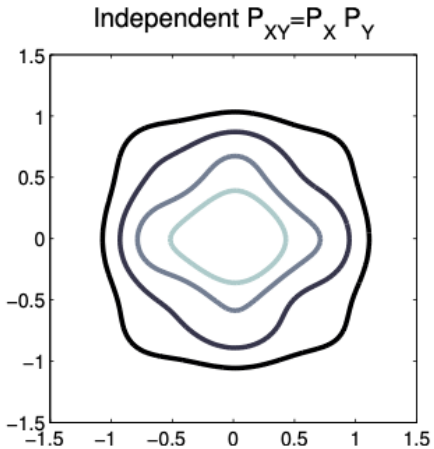
Motivation II: Detecting dependence

ntinuous domain?



?





Motivating questions

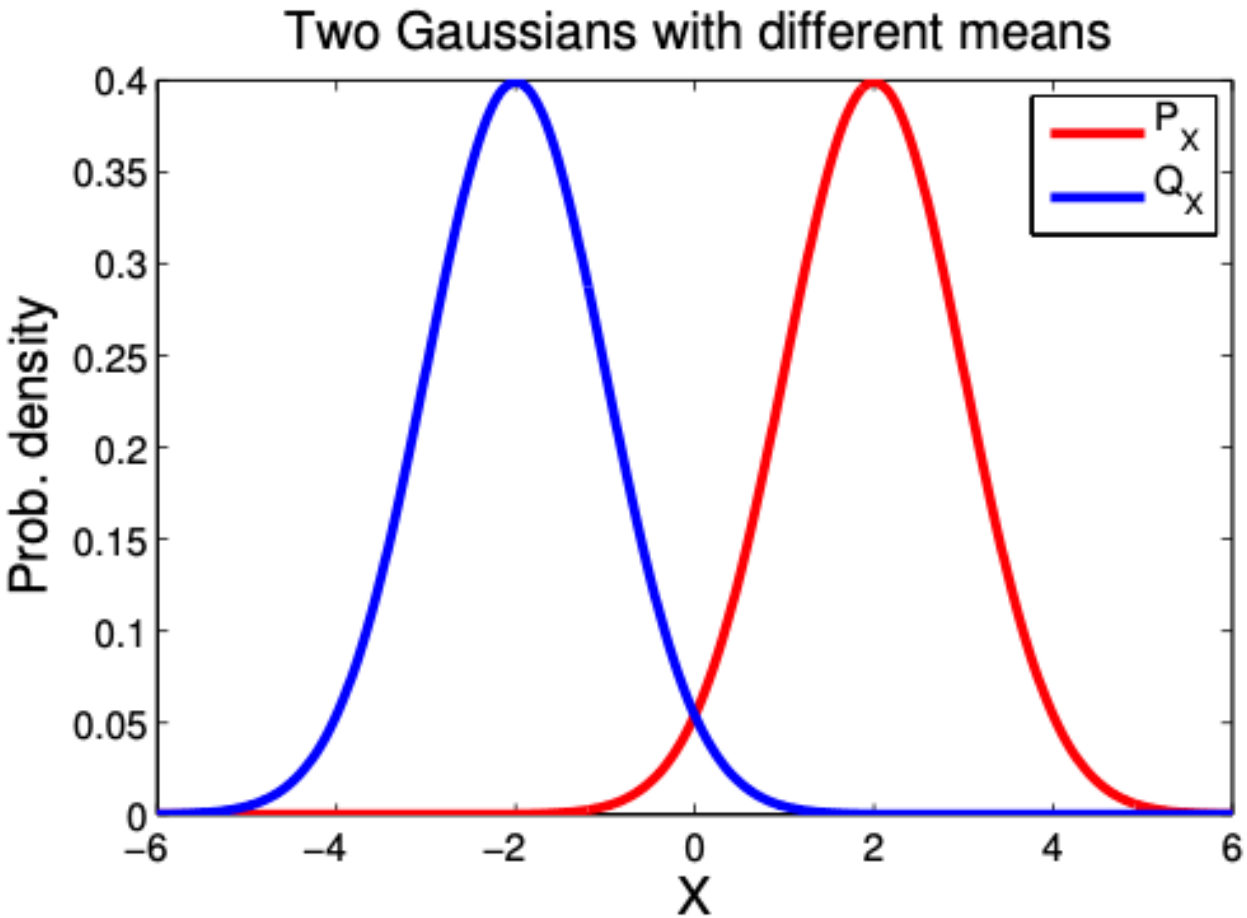
- Comparing distributions in high dimensions, low sample size and "complex" structure.
- Detecting dependence in high dimensional data
  - Feature selection
  - Blind source separation.

Outline

1. Characterizing probabilities with kernels
   - Kernel mean embedding
   - The Maximum Mean Discrepancy
   - Characteristic kernels

Feature mean difference

- Simple example: Samples from 2 Gaussians with same means but different variance.
- Idea: Look at difference in *means of features* of the samples.



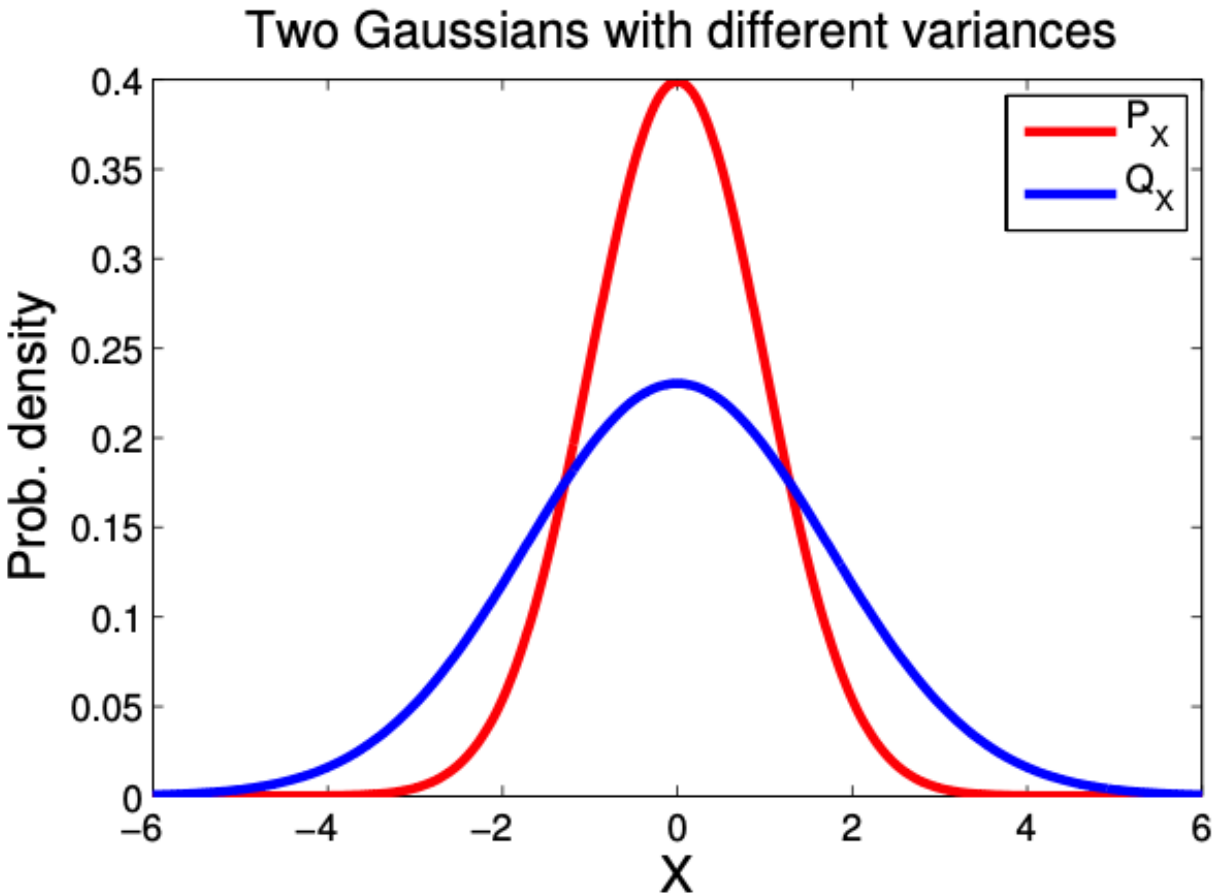
Compare

$$\hat{\mu}_{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

$$\hat{\mu}_{\mathbb{Q}} = \frac{1}{M}\sum_{j=1}^{N} y_j$$

Feature mean difference

- Simple example: Samples from 2 Gaussians with same means but different variance.
- Idea: Look at difference in *means of features* of the samples. Here $\varphi(x) = (x, x^2)$.



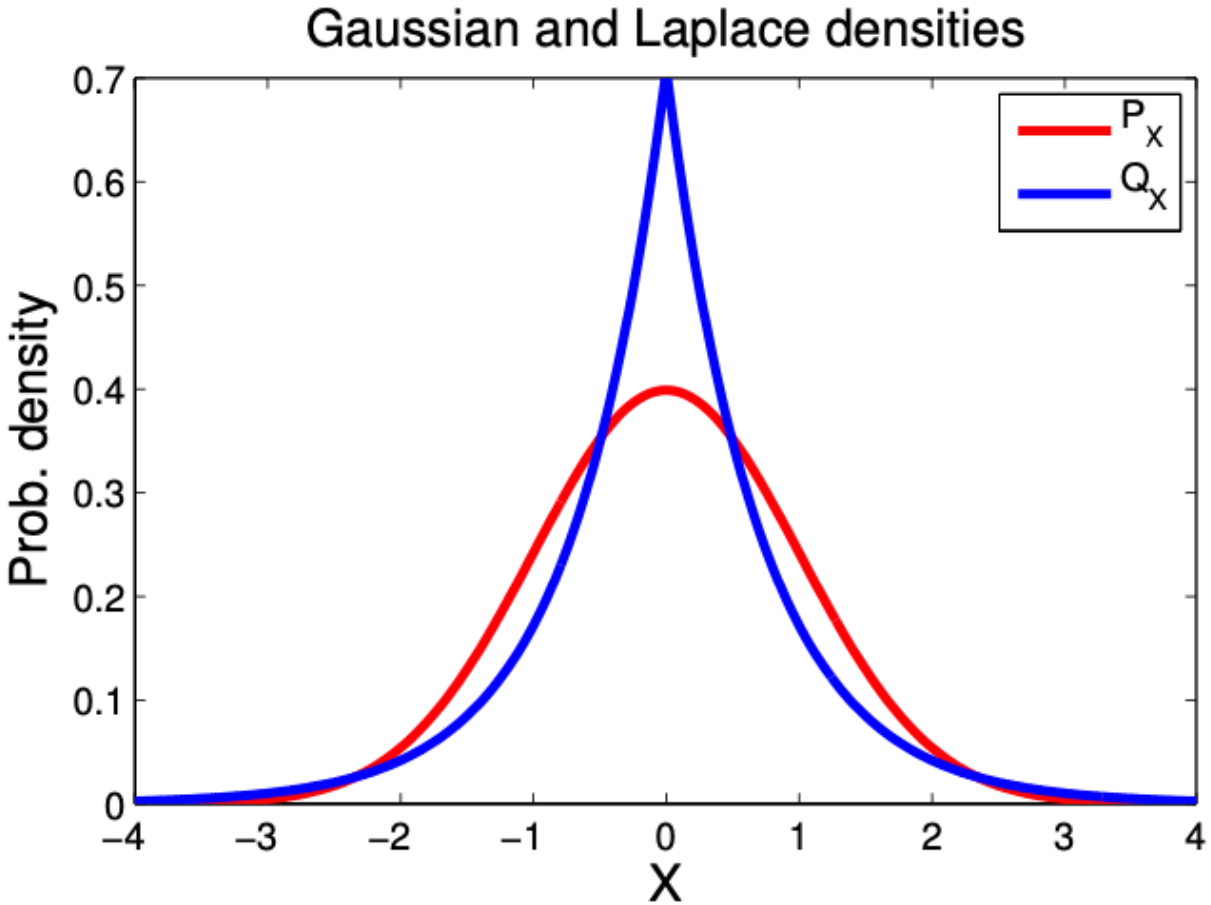
Compare

$$\hat{\mu}_{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^{N} \varphi(x_i),$$

$$\hat{\mu}_{\mathbb{Q}} = \frac{1}{M}\sum_{j=1}^{N} \varphi(y_j)$$

Feature mean difference

- Simple example: Centered Gaussian and Laplace distributions: same mean and variance.
- Idea: Look at difference in *means of high order features* of the samples: $\varphi(x) = (x, x^2, ...)$ (*RKHS*).



Compare

$$\hat{\mu}_{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i),$$

$$\hat{\mu}_{\mathbb{Q}} = \frac{1}{M}\sum_{j=1}^{N}\varphi(y_j)$$

Kernel Mean Embedding

Definition

Given a kernel K defined on a topological set $\mathcal{X}$ with corresponding RKHS $\mathcal{H}$, the mean embedding of a *Borel* probability distribution $\mathbb{P}$ on $\mathcal{X}$ is the function $\mu_{\mathbb{P}} : \mathcal{X} \to \mathbb{R}$ in $\mathcal{H}$ defined as

$$\mu_{\mathbb{P}}(y) := \mathbb{E}_{X \sim \mathbb{P}}[K(X, y)]$$

Kernel Mean Embedding

Definition

Given a kernel K defined on a topological set $\mathcal{X}$ with corresponding RKHS $\mathcal{H}$, the mean embedding of a *Borel* probability distribution $\mathbb{P}$ on $\mathcal{X}$ is the function $\mu_{\mathbb{P}} : \mathcal{X} \to \mathbb{R}$ in $\mathcal{H}$ defined as

$$\mu_{\mathbb{P}}(y) := \mathbb{E}_{X \sim \mathbb{P}}[K(X, y)]$$

- For any x, x' in $\mathcal{X}$,

$$K(x, x') = \langle K_x, K_{x'} \rangle_{\mathcal{H}},$$

- The kernel trick:
  For any $f \in \mathcal{H}$ and $x \in \mathcal{X}$,

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}}$$

Kernel Mean Embedding

Definition

Given a kernel K defined on a topological set $\mathcal{X}$ with corresponding RKHS $\mathcal{H}$, the mean embedding of a *Borel* probability distribution $\mathbb{P}$ on $\mathcal{X}$ is the function $\mu_{\mathbb{P}} : \mathcal{X} \to \mathbb{R}$ in $\mathcal{H}$ defined as

$$\mu_{\mathbb{P}}(y) := \mathbb{E}_{X \sim \mathbb{P}}[K(X, y)]$$

- For any x, x' in $\mathcal{X}$,

$$K(x, x') = \langle K_x, K_{x'} \rangle_{\mathcal{H}},$$

- The kernel trick:
  For any $f \in \mathcal{H}$ and $x \in \mathcal{X}$,

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}}$$

- For any Borel measure $\mathbb{P}$ and $\mathbb{Q}$,

$$\mathbb{E}_{(X,Y) \sim \mathbb{P},\mathbb{Q}} K(X, Y) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}},$$

- The generalized kernel trick:
  For any $f \in \mathcal{H}$ and Borel measure $\mathbb{P}$,

$$\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$$

Kernel Mean Embedding

Kernel Mean Embedding

The kernel mean embedding: $\mu_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}[K_X]$
The generalized kernel trick: $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

Kernel Mean Embedding

Kernel Mean Embedding

The kernel mean embedding: $\mu_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}[K_X]$
The generalized kernel trick: $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

- Mean embedding $\mu_{\mathbb{P}}$ summarizes $\mathbb{P}$:
  Can compute expectations under $\mathbb{P}$ of all functions in $\mathcal{H}$ using $\mu_{\mathbb{P}}$.
- In practice, you can estimate $\mu_{\mathbb{P}}$ using N i.i.d. samples from $\mathbb{P}$:

$$\hat{\mu}_{\mathbb{P}}(x) = \frac{1}{N} \sum_{i=1}^{N} K(X_i, x), \qquad X_i \overset{i.i.d.}{\sim} \mathbb{P}$$

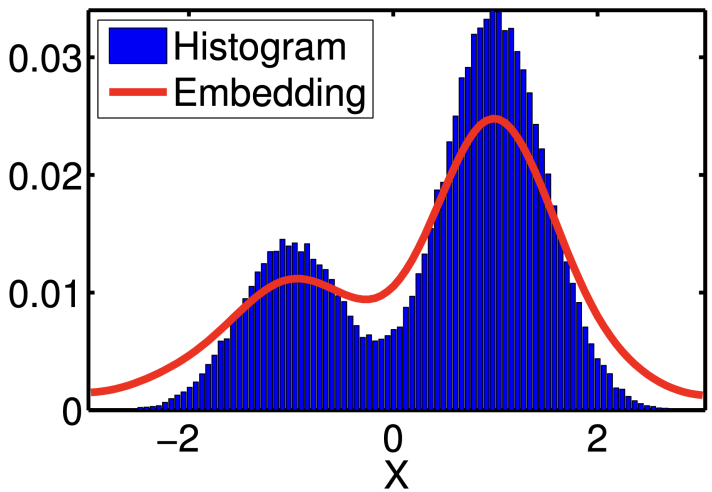
Kernel Mean Embedding

Kernel Mean Embedding

The kernel mean embedding: $\mu_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}[K_X]$
The generalized kernel trick: $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

- Mean embedding $\mu_{\mathbb{P}}$ summarizes $\mathbb{P}$:
  Can compute expectations under $\mathbb{P}$ of all functions in $\mathcal{H}$ using $\mu_{\mathbb{P}}$.
- In practice, you can estimate $\mu_{\mathbb{P}}$ using N i.i.d. samples from $\mathbb{P}$:

$$\hat{\mu}_{\mathbb{P}}(x) = \frac{1}{N} \sum_{i=1}^{N} K(X_i, x), \qquad X_i \overset{i.i.d.}{\sim} \mathbb{P}$$

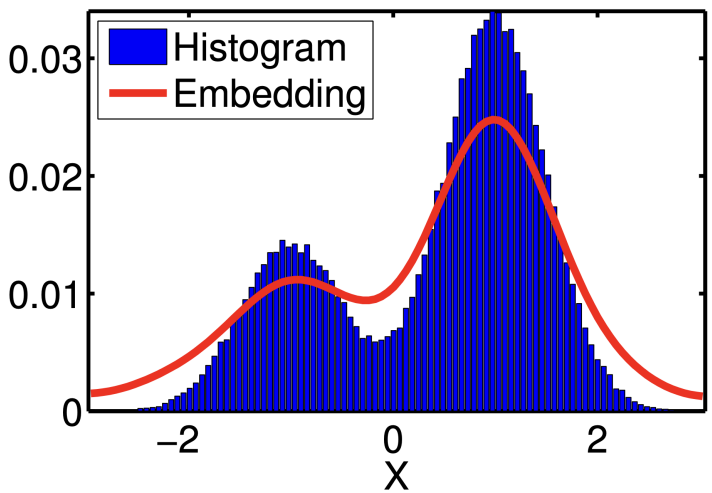


Kernel Mean Embedding

Kernel Mean Embedding

The kernel mean embedding: $\mu_{\mathbb{P}} = \mathbb{E}_{X\sim\mathbb{P}}[K_X]$
The generalized kernel trick: $\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

- Mean embedding $\mu_{\mathbb{P}}$ summarizes $\mathbb{P}$:
  Can compute expectations under $\mathbb{P}$ of all functions in $\mathcal{H}$ using $\mu_{\mathbb{P}}$.
- In practice, you can estimate $\mu_{\mathbb{P}}$ using N i.i.d. samples from $\mathbb{P}$:



$$\hat{\mu}_{\mathbb{P}}(x) = \frac{1}{N}\sum_{i=1}^{N} K(X_i, x), \qquad X_i \overset{i.i.d.}{\sim} \mathbb{P}$$

Does the mean embedding $\mu_{\mathbb{P}}$ exist? i.e. an element $\mu_{\mathbb{P}} \in \mathcal{H}$ such that

$$\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}, \forall f \in \mathcal{H}.$$

Existence of mean embeddings

Proposition

Let $\mathbb{P}$ be a Borel probability distribution on a set $\mathcal{X}$ endowed with its Borel sigma algebra. Let K be a p.d. kernel defined on $\mathcal{X}$ with corresponding RKHS $\mathcal{H}$. Assume that $\mathbb{E}_{X\sim\mathbb{P}}[\sqrt{K(X,X)}] < \infty$. Then there exits a unique element $\mu_{\mathbb{P}} \in \mathcal{H}$ such that

$$\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}.$$

In particular, for any $y \in \mathcal{X}$, it holds that:

$$\mu_{\mathbb{P}}(y) = \langle K_y, \mu_{\mathbb{P}} \rangle = \mathbb{E}_{X\sim\mathbb{P}}[K(X,y)].$$

Existence of mean embeddings

Proposition

Let $\mathbb{P}$ be a Borel probability distribution on a set $\mathcal{X}$ endowed with its Borel sigma algebra. Let K be a p.d. kernel defined on $\mathcal{X}$ with corresponding RKHS $\mathcal{H}$. Assume that $\mathbb{E}_{X\sim\mathbb{P}}[\sqrt{K(X,X)}] < \infty$. Then there exits a unique element $\mu_{\mathbb{P}} \in \mathcal{H}$ such that

$$\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}.$$

In particular, for any $y \in \mathcal{X}$, it holds that:

$$\mu_{\mathbb{P}}(y) = \langle K_y, \mu_{\mathbb{P}} \rangle = \mathbb{E}_{X\sim\mathbb{P}}[K(X,y)].$$

Proof:

The linear form on $\mathcal{H}$: $T_{\mathbb{P}} f = \mathbb{E}_{X\sim\mathbb{P}}[f(X)]$ is bounded by assumption:

$$|T_{\mathbb{P}} f| \leq \mathbb{E}_{X\sim\mathbb{P}}[|f(X)|] = \mathbb{E}_{X\sim\mathbb{P}}[|\langle f, K_X \rangle_{\mathcal{H}}|] \leq \mathbb{E}_{X\sim\mathbb{P}}[\sqrt{K(X,X)}\|f\|_{\mathcal{H}}].$$

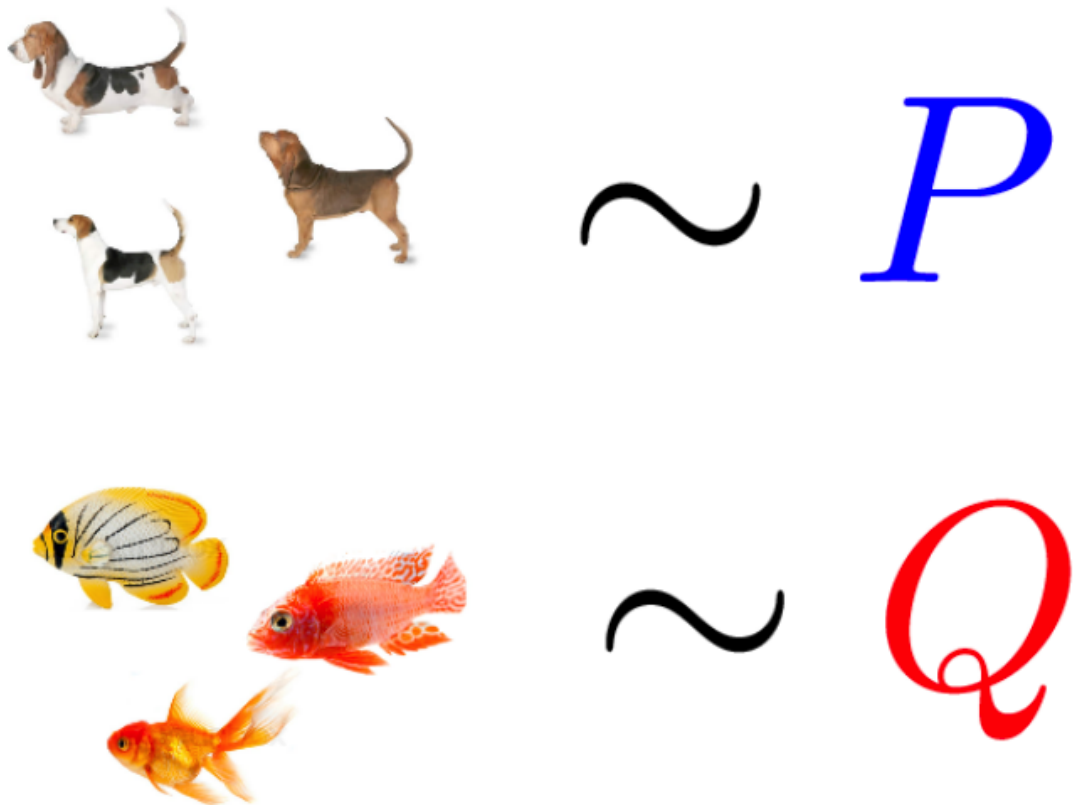
Hence, by Riesz's theorem, there exists $\mu_{\mathbb{P}} \in \mathcal{H}$ such that $T_{\mathbb{P}} f = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$.

Outline

1. Characterizing probabilities with kernels
   - Kernel mean embedding
   - The Maximum Mean Discrepancy
   - Characteristic kernels

Motivation: Comparing two distributions

- Data: Samples from unknown distributions $\mathbb{P}$ and $\mathbb{Q}$.
- Goal: do $\mathbb{P}$ and $\mathbb{Q}$ differ?



Differences between dogs and fish.

The Maximum Mean Discrepancy

The maximum mean discrepancy (MMD) is the RKHS distance between mean embeddings:

$$MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}}$$

The Maximum Mean Discrepancy

The maximum mean discrepancy (MMD) is the RKHS distance between mean embeddings:

$$\begin{aligned} MMD^2(\mathbb{P}, \mathbb{Q}) =& \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}} \\ =& \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \end{aligned}$$

The Maximum Mean Discrepancy

The maximum mean discrepancy (MMD) is the RKHS distance between mean embeddings:

$$\begin{aligned} MMD^2(\mathbb{P}, \mathbb{Q}) =& \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}} \\ =& \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ =& \mathbb{E}_{X,X' \sim \mathbb{P}\otimes\mathbb{P}}[k(X, X')] + \mathbb{E}_{Y,Y' \sim \mathbb{Q}\otimes\mathbb{Q}}[k(Y, Y')] \\ & - 2\mathbb{E}_{X,Y \sim \mathbb{P}\otimes\mathbb{Q}}[k(X, Y)] \end{aligned}$$

The Maximum Mean Discrepancy

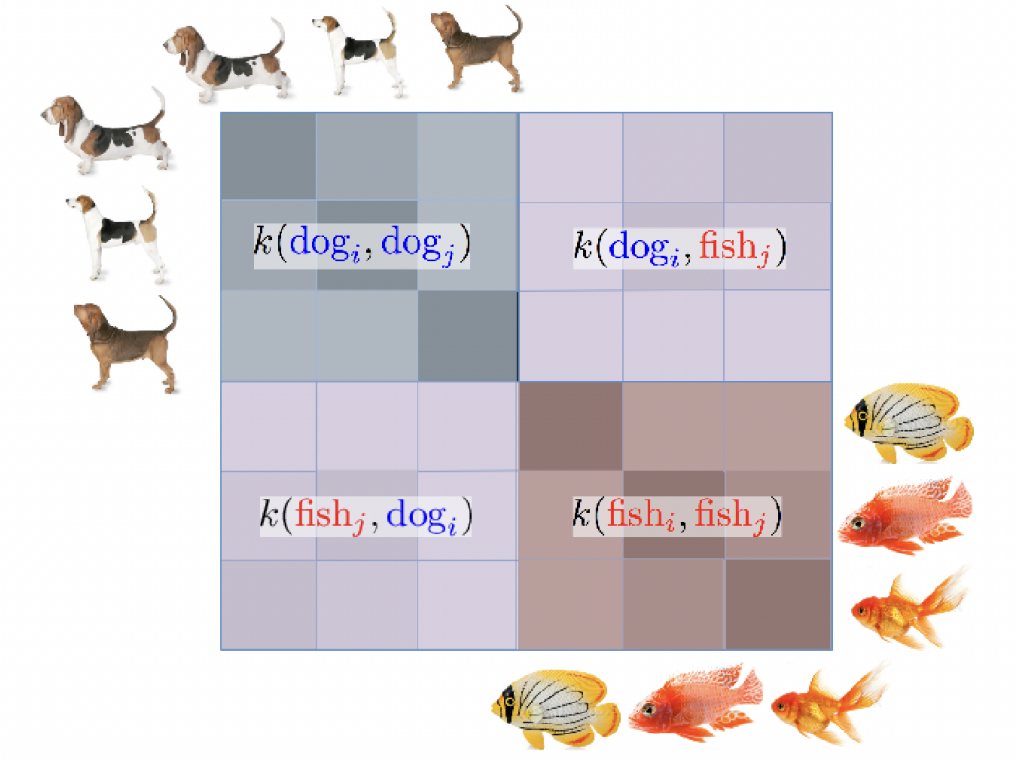
The maximum mean discrepancy (MMD) is the RKHS distance between mean embeddings:

$$\begin{aligned} MMD^2(\mathbb{P}, \mathbb{Q}) =& \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}} \\ =& \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ =& \mathbb{E}_{X,X' \sim \mathbb{P} \otimes \mathbb{P}}[k(X, X')] + \mathbb{E}_{Y,Y' \sim \mathbb{Q} \otimes \mathbb{Q}}[k(Y, Y')] \\ & - 2\mathbb{E}_{X,Y \sim \mathbb{P} \otimes \mathbb{Q}}[k(X, Y)] \end{aligned}$$

- **Intra-similarity** terms : $\mathbb{E}_{X,X' \sim \mathbb{P} \otimes \mathbb{P}}[k(X, X')]$ and $\mathbb{E}_{Y,Y' \sim \mathbb{Q} \otimes \mathbb{Q}}[k(Y, Y')]$.
- **Inter-similarity** term: $\mathbb{E}_{X,Y \sim \mathbb{P} \otimes \mathbb{Q}}[k(X, Y)]$.
- In general, MMD is a semi-metric: $(MMD(\mathbb{P}, \mathbb{Q}) = 0 \nRightarrow \mathbb{P} = \mathbb{Q})$.
- For some kernels (called characteristic kernels), MMD is a metric $(MMD(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q})$.
- From now on, we assume MMD is a metric. Later, we'll say more about characteristic kernels.

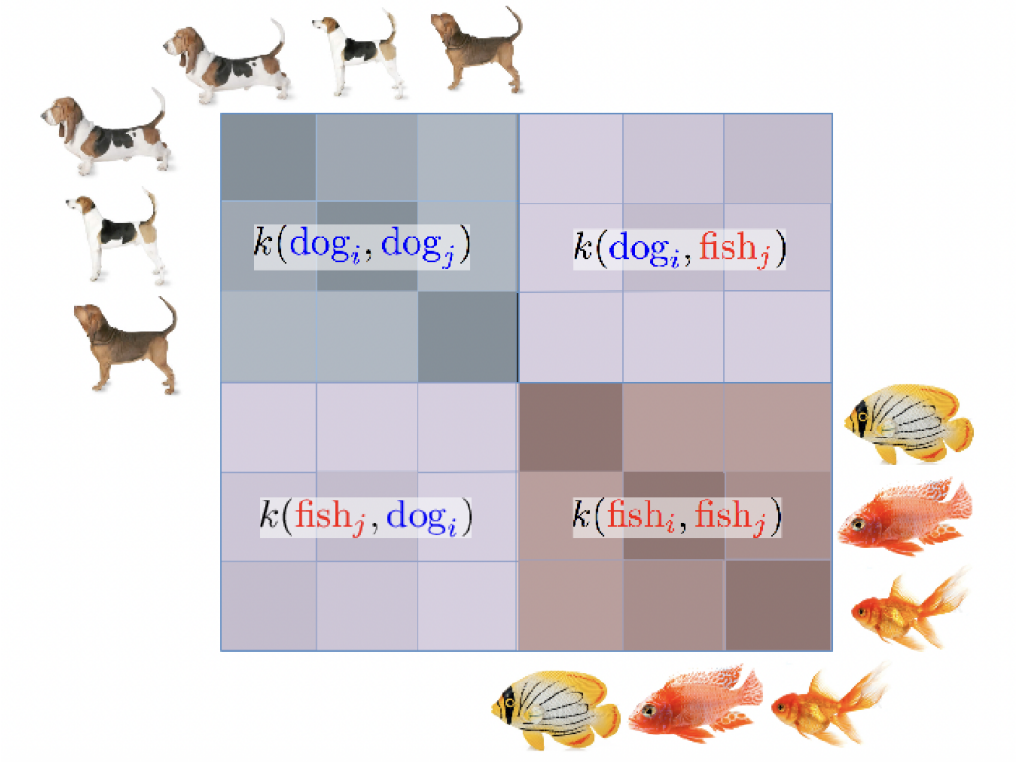
Unbiased esitimation of the MMD

- Data: i.i.d. samples from $\mathbb{P}$ and $\mathbb{Q}$



Unbiased esitimation of the MMD

- Data: i.i.d. samples from $\mathbb{P}$ and $\mathbb{Q}$

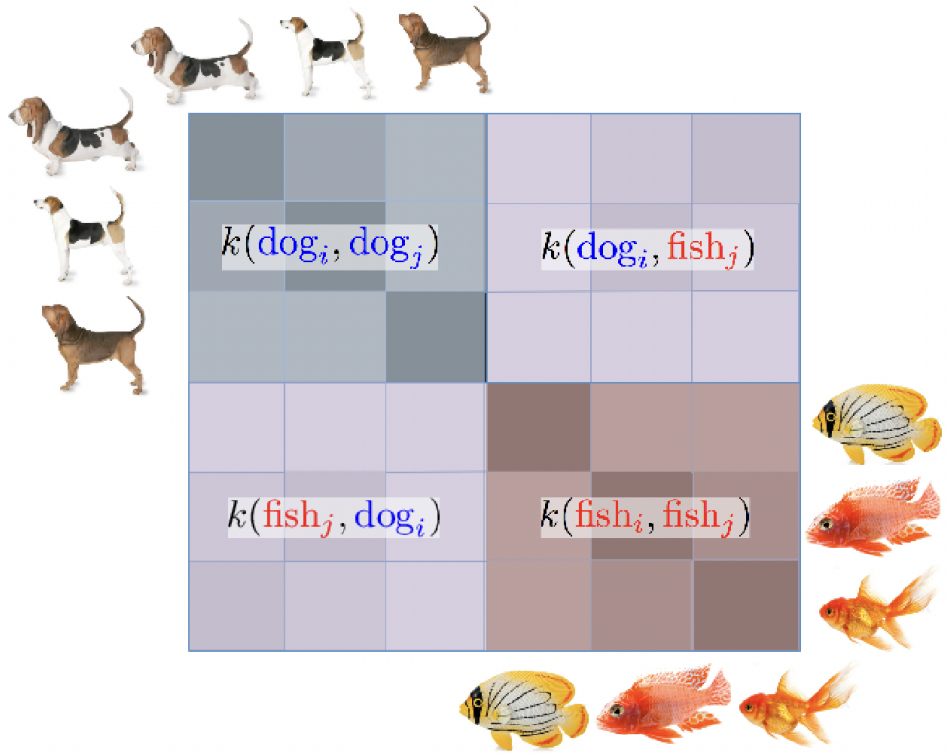


Biased estimate of the MMD^2:

$$\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) = \frac{1}{N^2}\sum_{i,j} K(dog_i, dog_j) + \frac{1}{M^2}\sum_{i,j} K(fish_i, fish_j) - \frac{2}{NM}\sum_{i,j} k(dog_i, fish_j)$$

Unbiased esitimation of the MMD

- Data: i.i.d. samples from $\mathbb{P}$ and $\mathbb{Q}$



Unbiased estimate of the MMD^2:

$$\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) = \frac{1}{N(N-1)} \sum_{i \neq j} K(dog_i, dog_j) + \frac{1}{M(M-1)} \sum_{i \neq j} K(fish_i, fish_j) - \frac{2}{NM} \sum_{i,j} k(dog_i, fish_j)$$

MMD as an Integral Probability Metric

Integral Probability Metric

Let $\mathcal{F}$ be a set of measurable functions. An integral probability metric associated to the class $\mathcal{F}$ is a semi-metric defined as

$$\mathcal{D}_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)].$$

MMD as an Integral Probability Metric

Integral Probability Metric

Let $\mathcal{F}$ be a set of measurable functions. An integral probability metric associated to the class $\mathcal{F}$ is a semi-metric defined as

$$\mathcal{D}_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)].$$

- MMD obtained by choosing $\mathcal{F} = \{f \in \mathcal{H} | \|f\|_{\mathcal{H}} \leq 1\}$:

$$MMD(\mathbb{P}, \mathbb{Q}) := \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$

MMD as an Integral Probability Metric

Integral Probability Metric

Let $\mathcal{F}$ be a set of measurable functions. An integral probability metric associated to the class $\mathcal{F}$ is a semi-metric defined as

$$\mathcal{D}_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)].$$

- MMD obtained by choosing $\mathcal{F} = \{f \in \mathcal{H} | \|f\|_{\mathcal{H}} \leq 1\}$:

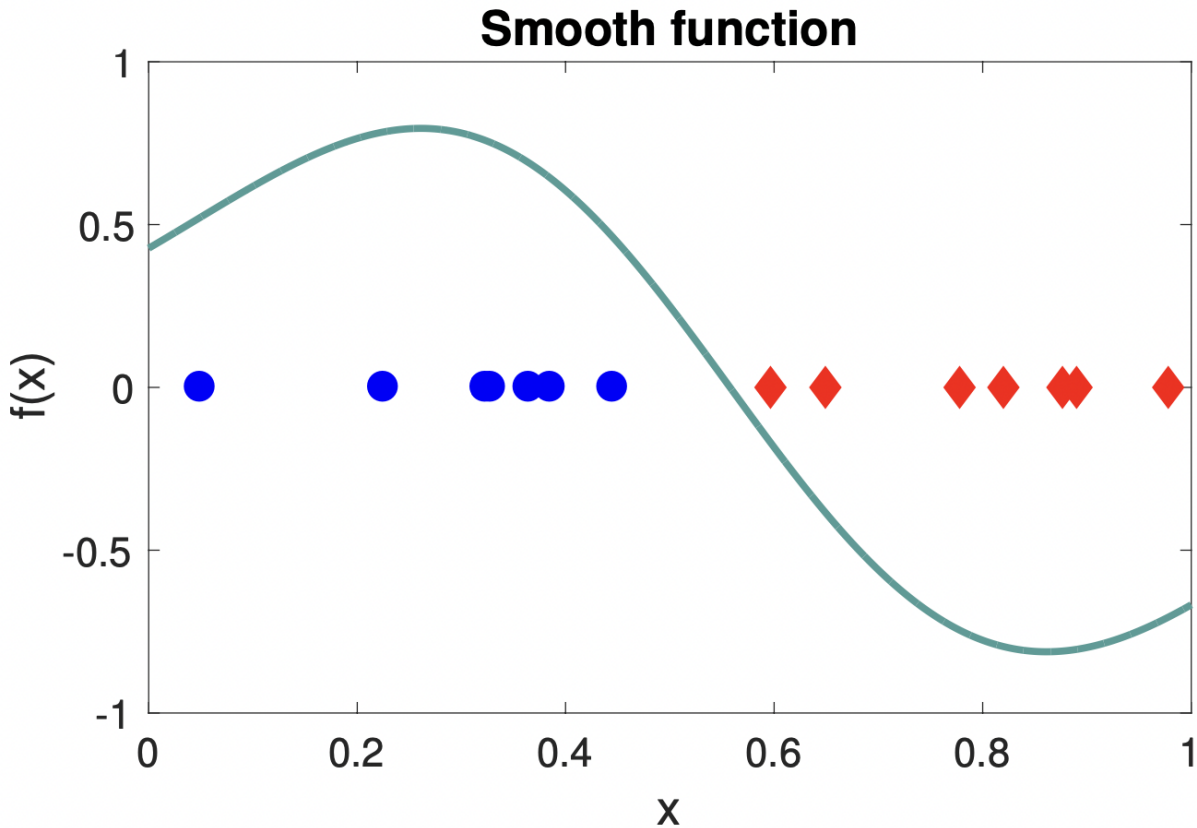
$$MMD(\mathbb{P}, \mathbb{Q}) := \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$

- Other choices for the set $\mathcal{F}$:
  - Bounded continuous $\rightarrow$ Dudley's metric.
  - Bounded variations $\rightarrow$ Kolmogorov metric.
  - Bounded Lipschitz $\rightarrow$ 1-Wasserstein distance.

MMD as an Integral Probability Metric

- MMD obtained by choosing $\mathcal{F} = \{f \in \mathcal{H} | \|f\|_{\mathcal{H}} \leq 1\}$:

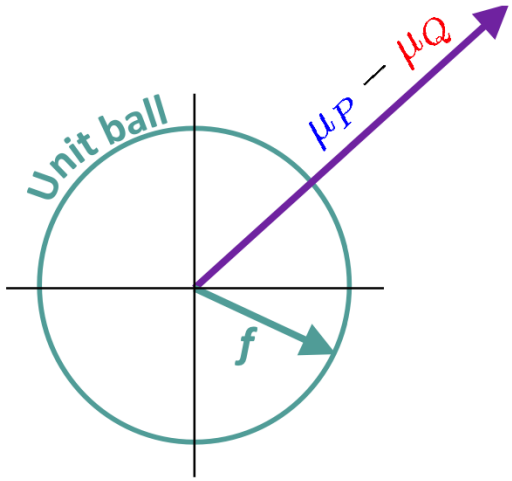
$$MMD(\mathbb{P}, \mathbb{Q}) = \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$



MMD as an Integral Probability Metric

- MMD obtained by choosing $\mathcal{F} = \{f \in \mathcal{H} | \|f\|_{\mathcal{H}} \leq 1\}$:

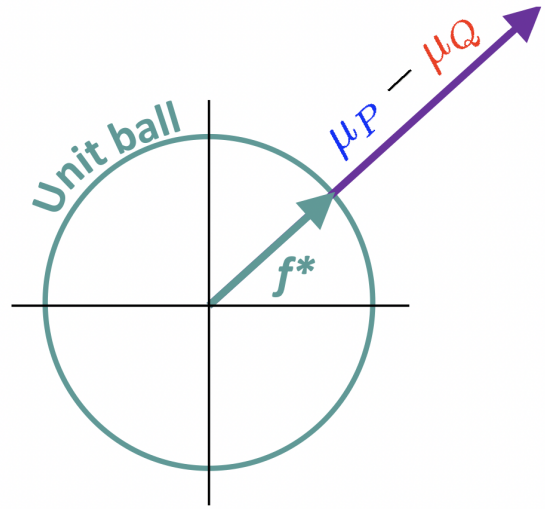
$$
\begin{aligned}
MMD(\mathbb{P}, \mathbb{Q}) &= \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)] \\
&= \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}
\end{aligned}
$$



MMD as an Integral Probability Metric

- MMD obtained by choosing $\mathcal{F} = \{f \in \mathcal{H} | \|f\|_{\mathcal{H}} \leq 1\}$:

$$
\begin{aligned}
MMD(\mathbb{P}, \mathbb{Q}) = & \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)] \\
= & \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
= & \langle f^{\star}, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
= & \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}
\end{aligned}
$$



$$
f^{\star} = \frac{\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}}{\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|}
$$

$f^{\star}$ is called the witness function

Outline

1. Characterizing probabilities with kernels
   - Kernel mean embedding
   - The Maximum Mean Discrepancy
     - Applications (I): Statistical testing using the MMD
     - Applications (II): Learning generative models
   - Characteristic kernels

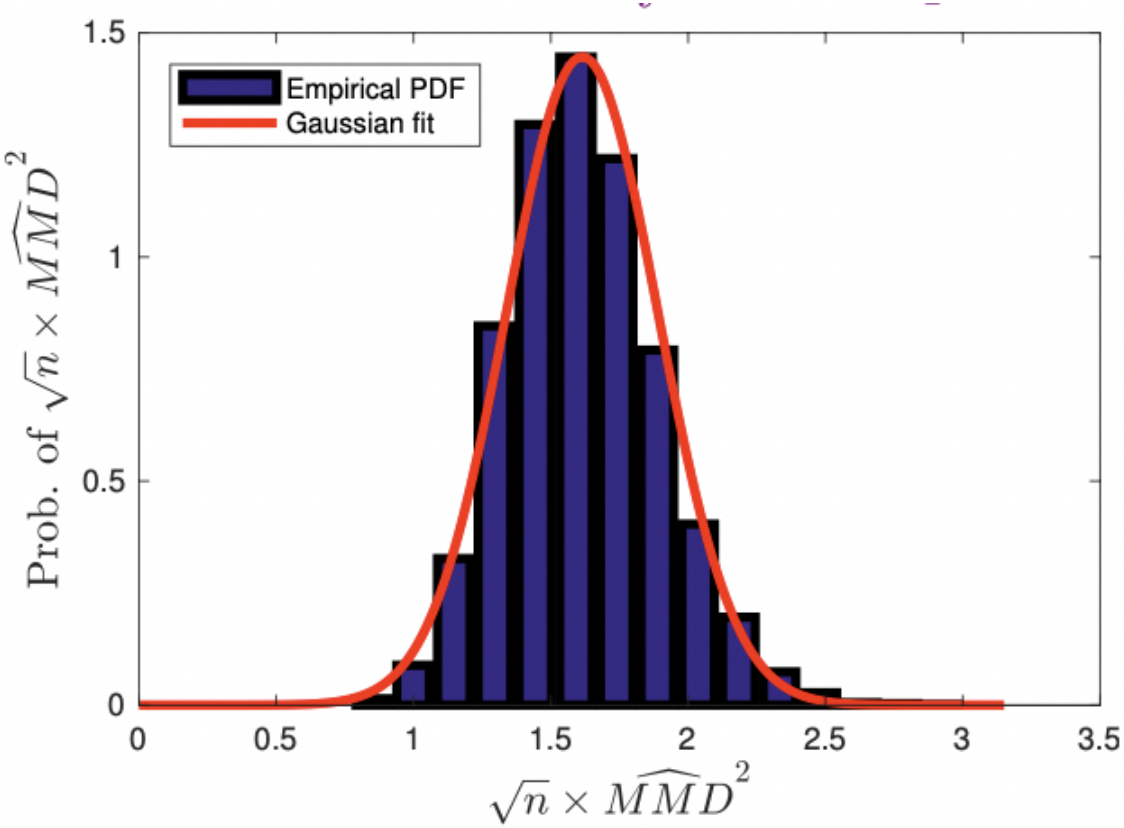
A statistical test using MMD

For simplicity assume same number of samples from $\mathbb{P}$ and $\mathbb{Q}$:

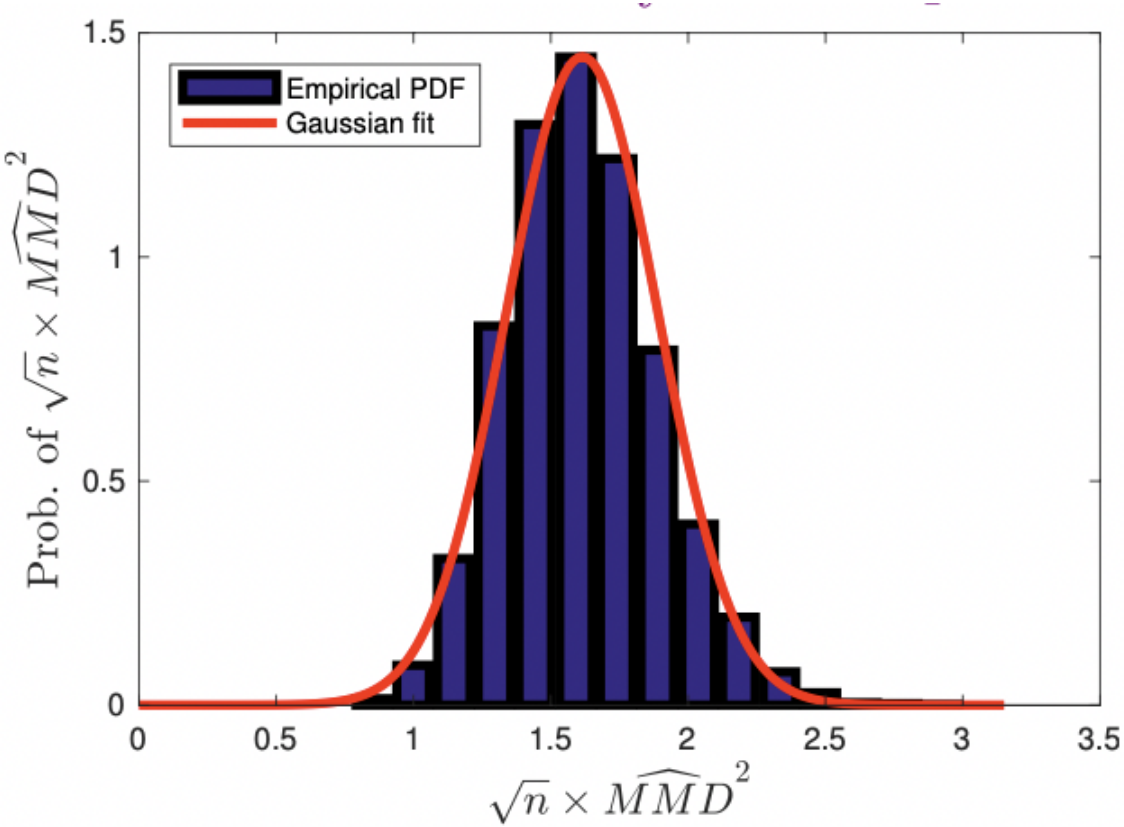
$$\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) = \frac{1}{N(N-1)} \sum_{i \neq j} K(x_i, x_j) + \frac{1}{N(N-1)} \sum_{i \neq j} K(y_i, y_j) - \frac{2}{N^2} \sum_{i,j} K(x_i, y_j)$$

- **Null hypothesis** h_0 when $\mathbb{P} = \mathbb{Q}$.
  $\widehat{MMD^2}(\mathbb{P}, \mathbb{Q})$ should be close to zero.
- **Alternative hypothesis** h_1 when $\mathbb{P} \neq \mathbb{Q}$.
  $\widehat{MMD^2}(\mathbb{P}, \mathbb{Q})$ should be far away from zero.
- What do close or far away mean here?

Behaviour of MMD when $\mathbb{P} \neq \mathbb{Q}$



Behaviour of MMD when $\mathbb{P} \neq \mathbb{Q}$

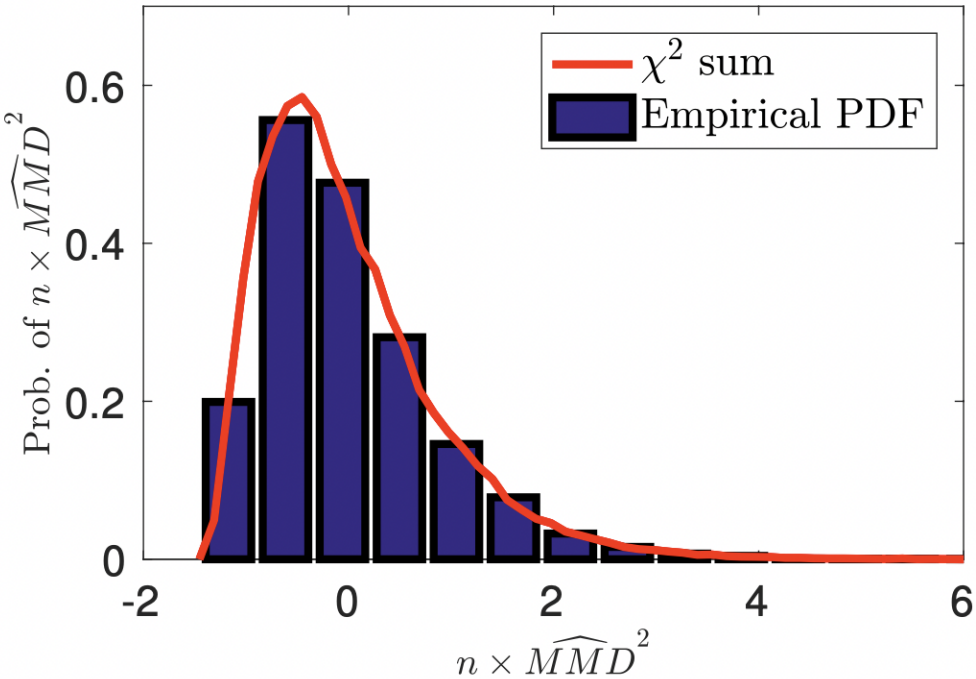


The statistic $\widehat{MMD^2}(\mathbb{P}, \mathbb{Q})$ is asymptotically normal [Gretton, 2006]:

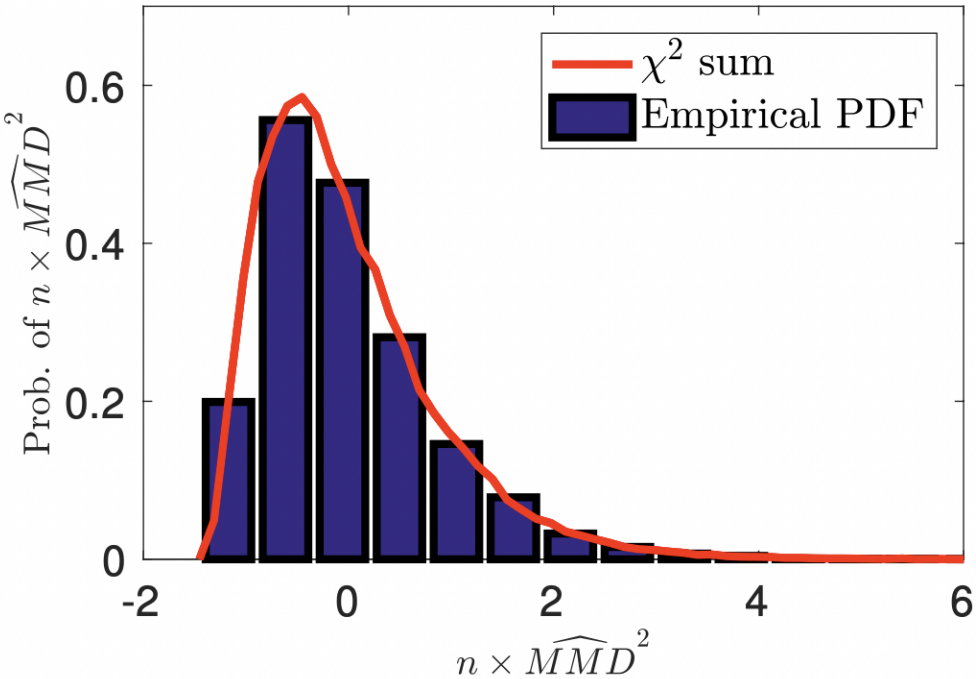
$$\frac{\sqrt{n}(\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) - MMD^2(\mathbb{P}, \mathbb{Q}))}{\sqrt{V(\mathbb{P}, \mathbb{Q})}} \to \mathcal{N}(0, 1).$$

where $V(\mathbb{P}, \mathbb{Q})$ is the asymptotic variance of $\sqrt{n} \times (\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}))$.

Behaviour of MMD when $\mathbb{P} = \mathbb{Q}$



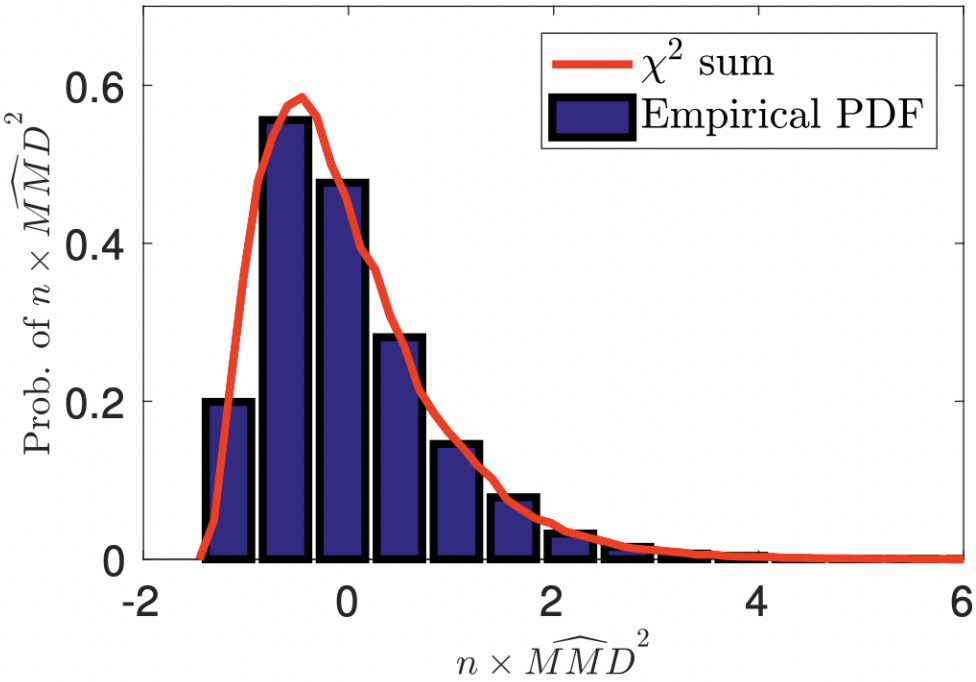
Behaviour of MMD when $\mathbb{P} = \mathbb{Q}$



$n\widehat{MMD^2}(\mathbb{P}, \mathbb{Q})$ has an asymptotic distribution [Gretton, 2006]:

$$n\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) \sim 2 \sum_{i=1}^{\infty} \lambda_i (z_i^2 - 1)$$

Behaviour of MMD when $\mathbb{P} = \mathbb{Q}$



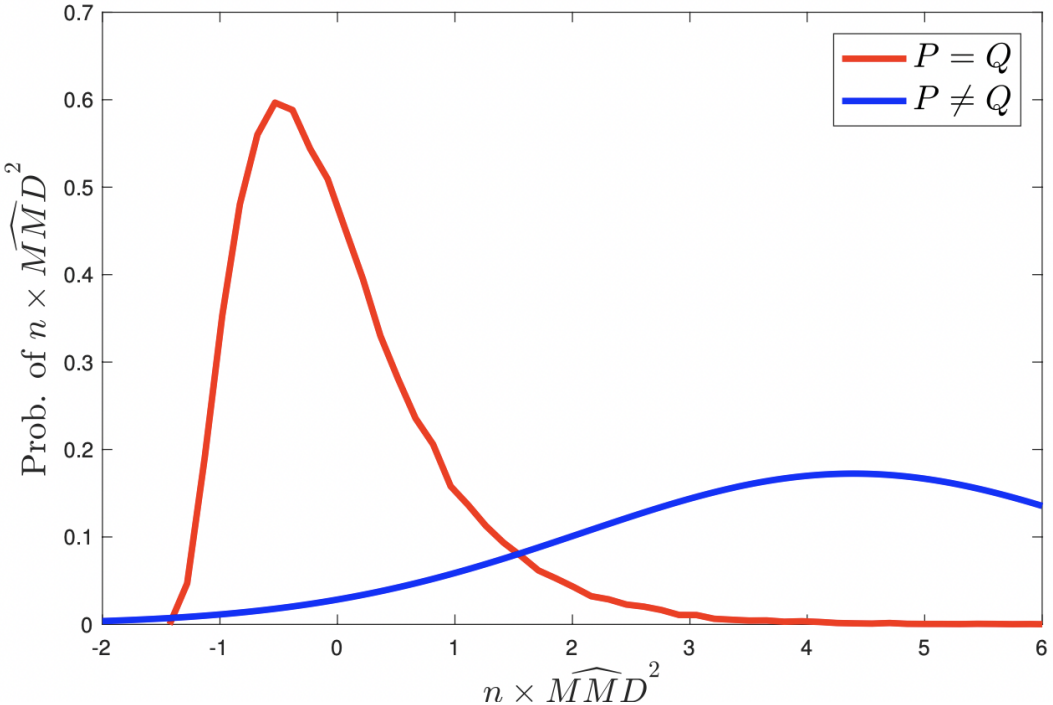
$n\widehat{MMD^2}(\mathbb{P}, \mathbb{Q})$ has an asymptotic distribution [Gretton, 2006]:

$$n\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) \sim 2 \sum_{i=1}^{\infty} \lambda_i (z_i^2 - 1)$$

- z_i are i.i.d. standard gaussians: $z_i \sim \mathcal{N}(0, 1)$
- λ_i are eigenvalues of the operator $f \mapsto \mathbb{E}_{X \sim \mathbb{P}}[\tilde{K}(X, X')f(X)]$
- $\tilde{K}$ the centered kernel:

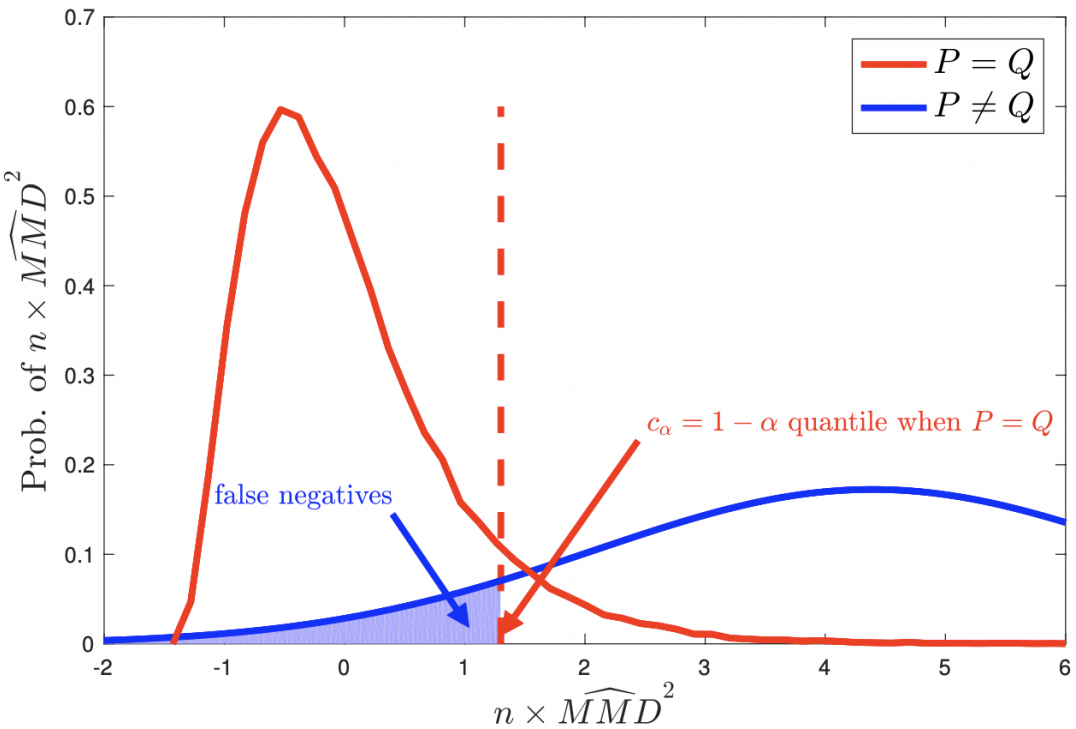
$$\tilde{K}(x, x') = \langle K(x, .) - \mu_{\mathbb{P}}, K(x', .) - \mu_{\mathbb{P}} \rangle_{\mathcal{H}}.$$

A statistical test using MMD



$$T_0 := n\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) \approx \begin{cases} nMMD^2(\mathbb{P}, \mathbb{Q}) + \sqrt{n}\mathcal{N}(0, V(\mathbb{P}, \mathbb{Q})), & \mathbb{P} \neq \mathbb{Q} \\ 2\sum_{i=1}^{\infty} \lambda_i (z_i^2 - 1), & \mathbb{P} = \mathbb{Q}. \end{cases}$$

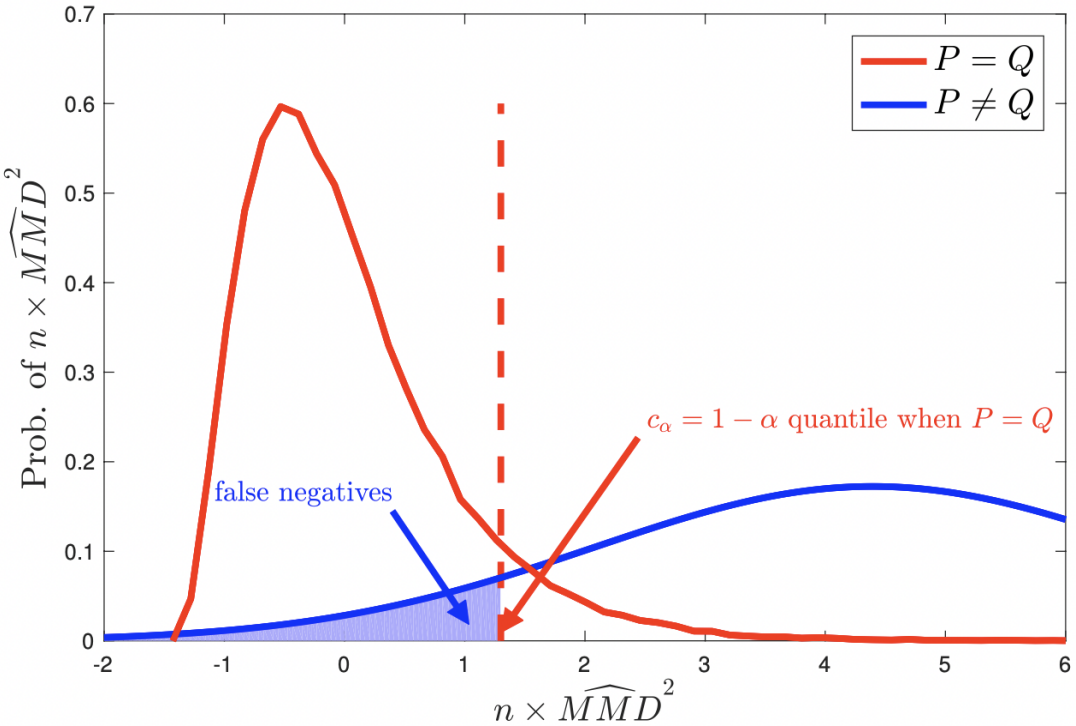
A statistical test using MMD



- Fix a significance level α (usually $\alpha = 0.05$.)
- If $T_0 \geq c_\alpha$, reject the null, i.e. ($\mathbb{P} = \mathbb{Q}$ unlikely)
- Otherwise, cannot reject ($\mathbb{P} = \mathbb{Q}$ is likely).

$$T_0 := n\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) \approx \begin{cases} nMMD^2(\mathbb{P}, \mathbb{Q}) + \sqrt{n}\mathcal{N}(0, V(\mathbb{P}, \mathbb{Q})), & \mathbb{P} \neq \mathbb{Q} \\ 2\sum_{i=1}^{\infty} \lambda_i (z_i^2 - 1), & \mathbb{P} = \mathbb{Q}. \end{cases}$$

A statistical test using MMD



- Fix a significance level α (usually a small value: 0.05.)
- If $T_0 \geq c_\alpha$, reject the null, i.e. ($\mathbb{P} = \mathbb{Q}$ unlikely)
- Otherwise, cannot reject ($\mathbb{P} = \mathbb{Q}$ is likely).

How can we tell if $T_0 := n\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) \geq c_\alpha$?

- Let T be a r.v. under the null distribution: $T \sim 2\sum_{i=1}^{\infty} \lambda_i (z_i^2 - 1)$.
- If the p-value $p := \mathbb{P}_T(T > T_0) \leq \alpha$, then $T_0 \geq c_\alpha$.
- For $T_1, ..., T_J$ samples from the null: $p \approx |\{j | T_j \geq T_0\}|/J$.

Can use a permutation test to construct $T_1, ..., T_J$.

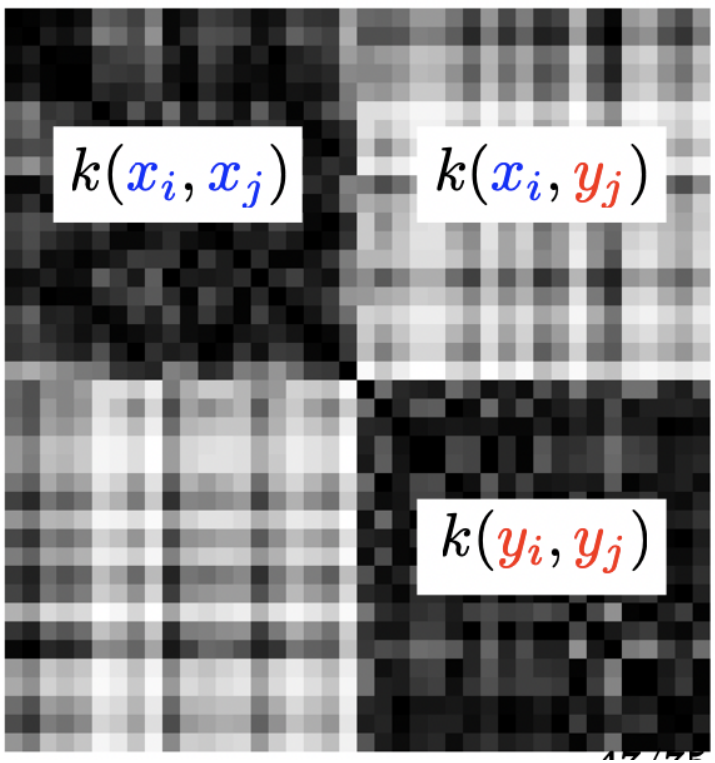
A statistical test using MMD

Original empirical MMD for dogs and fish:

$X = [$  $\dots]$

$Y = [$  $\dots]$

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)$$



For each permutation j set $T_j = nMMD^2(\tilde{\mathbb{P}}, \tilde{\mathbb{Q}})$

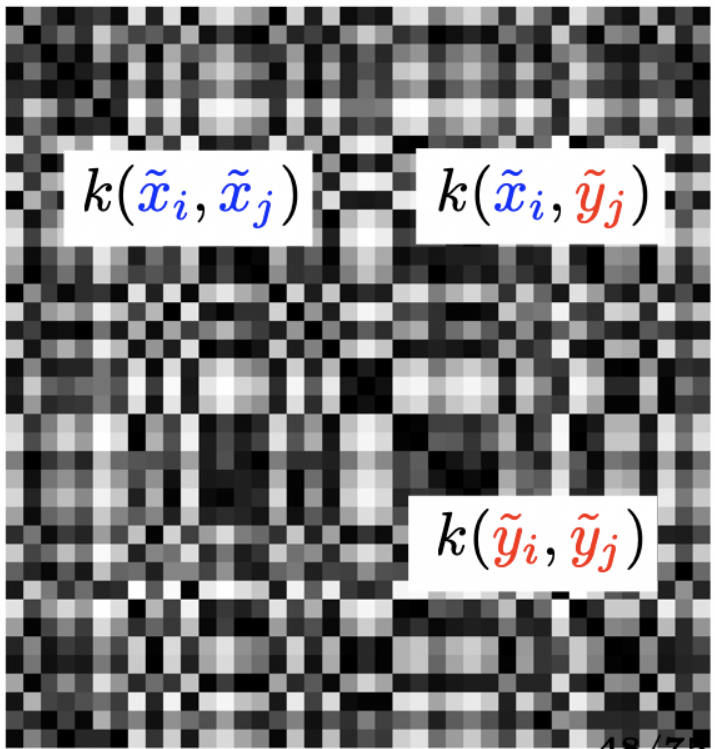
A statistical test using MMD

Permuted dog and fish samples (merdogs):

$\tilde{X} = [$  $\ldots\ ]$

$\tilde{Y} = [$  $\ldots\ ]$

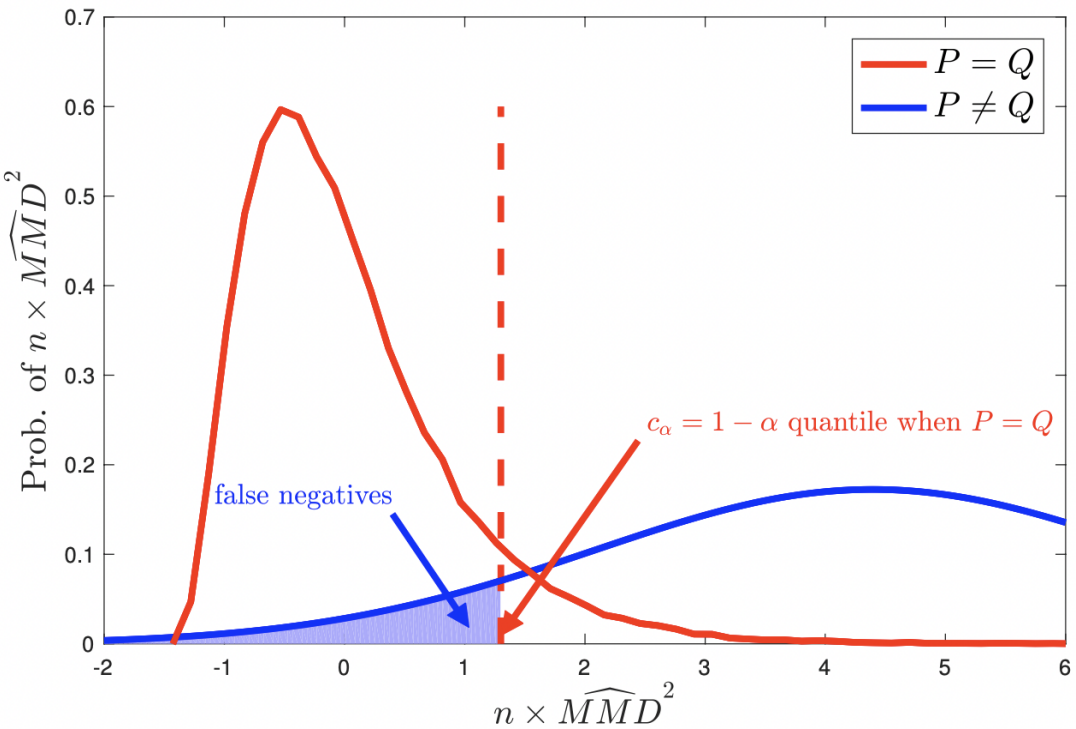
$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(\tilde{x}_i, \tilde{x}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\tilde{y}_i, \tilde{y}_j) - \frac{2}{n^2} \sum_{i,j} k(\tilde{x}_i, \tilde{y}_j)$$



Permutation simulates $P = Q$

For each permutation j set $T_j = nMMD^2(\tilde{\mathbb{P}}, \tilde{\mathbb{Q}})$

A statistical test using MMD



- Fix a significance level α (usually a small value: 0.05.)
- If $T_0 \geq c_\alpha$, reject the null, i.e. ($\mathbb{P} = \mathbb{Q}$ unlikely)
- Otherwise, cannot reject ($\mathbb{P} = \mathbb{Q}$ is likely).

How can we tell if $T_0 := n\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) \geq c_\alpha$?

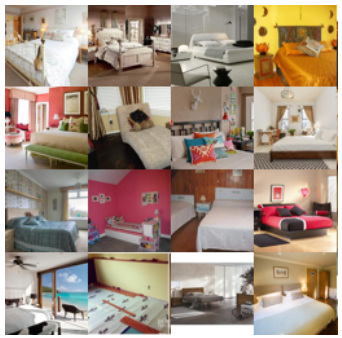
- Let T be a r.v. under the null distribution: $T \sim 2\sum_{i=1}^{\infty} \lambda_i (z_i^2 - 1)$.
- If the p-value $p := \mathbb{P}_T(T > T_0) \leq \alpha$, then $T_0 \geq c_\alpha$.
- For $T_1, ..., T_J$ samples from the null: $p \approx |\{j | T_j \geq T_0\}|/J$.

Can use a permutation test to construct $T_1, ..., T_J$.

Outline

1. Characterizing probabilities with kernels
   - Kernel mean embedding
   - The Maximum Mean Discrepancy
     - Applications (I): Statistical testing using the MMD
     - Applications (II): Learning generative models
   - Characteristic kernels

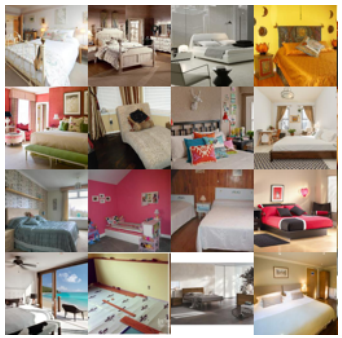
Given samples from a distribution $\mathbb{P}$ over $\mathcal{X}$, want a model that can produce new samples from $\mathbb{Q} \approx \mathbb{P}$



$X \sim \mathbb{P}$

$Y \sim \mathbb{Q}$

Given samples from a distribution $\mathbb{P}$ over $\mathcal{X}$, want a model that can produce new samples from $\mathbb{Q} \approx \mathbb{P}$



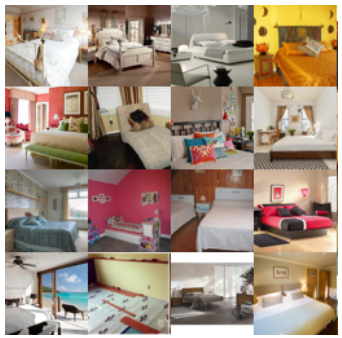
$X \sim \mathbb{P}$

$Y \sim \mathbb{Q}$

EGM: $\mathbb{Q}$ has density $q(Y)$.

- **Support**: the whole space.
- **Training** using maximum likelihood or score matching.
- **Sampling** using MCMC.

Given samples from a distribution $\mathbb{P}$ over $\mathcal{X}$, want a model that can produce new samples from $\mathbb{Q} \approx \mathbb{P}$



$X \sim \mathbb{P}$

$Y \sim \mathbb{Q}$

EGM: $\mathbb{Q}$ has density $q(Y)$.

- **Support**: the whole space.
- **Training** using maximum likelihood or score matching.
- **Sampling** using MCMC.

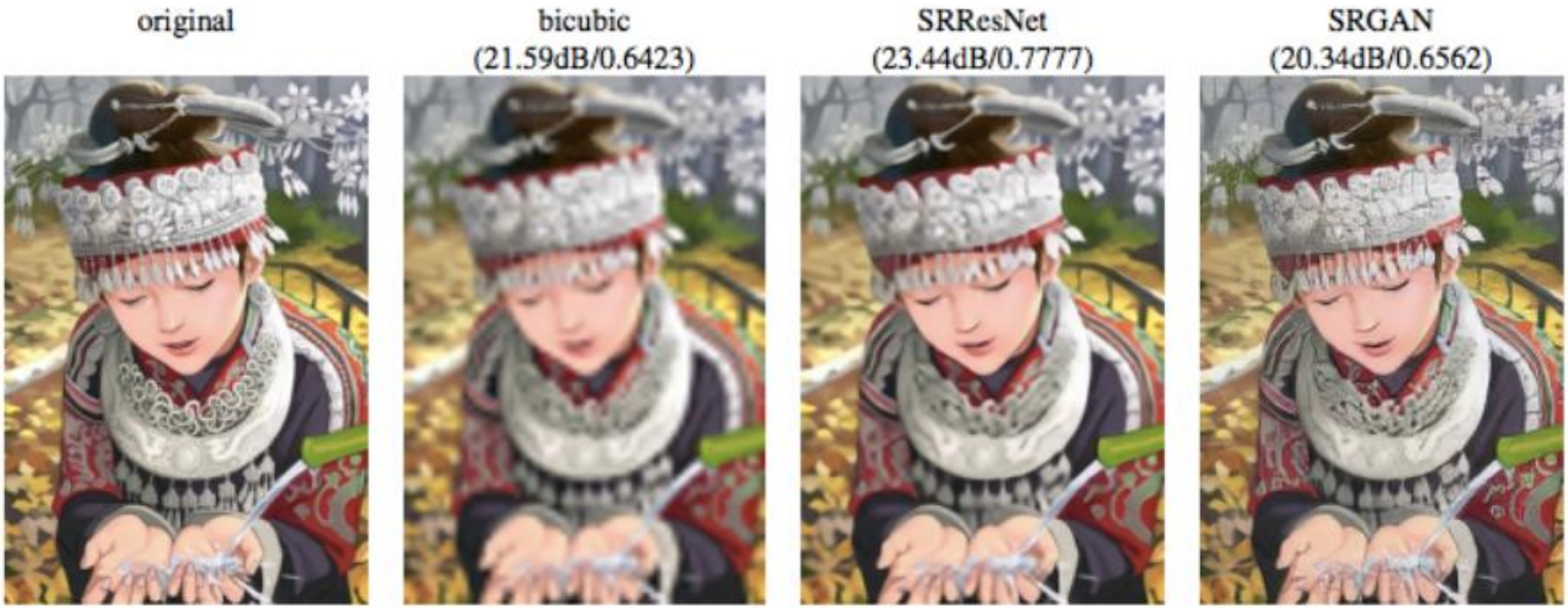
IGM: $Y = G(Z) \sim \mathbb{Q}$ with known $Z \sim \mu$.

- **Support**: low dimensional [Arjovsky 2017.
- **Training** by minimizing some well chosen divergence $D(\mathbb{P}, \mathbb{Q})$.
- **Sampling** by pushing μ forward with G.

Generative Adversarial Networks

Many successful applications:

- Single-image super-resolution

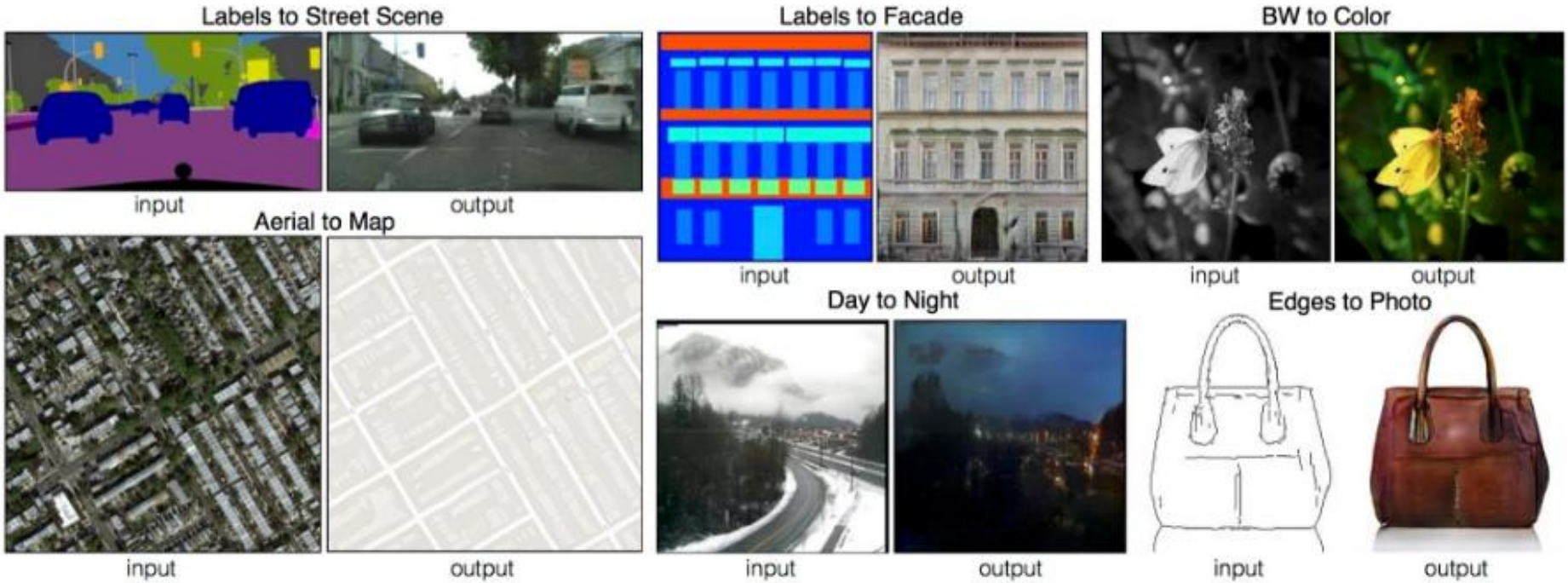


Ledig et al 2015

Generative Adversarial Networks

Many successful applications:

- Image to image translation



Isola et al 2016

Generative Adversarial Networks

Many successful applications:

- Text to image generation

This small blue bird has a short pointy beak and brown on its wings

This bird is completely red with black wings and pointy beak

Zhang et al 2016

Adversarial training [Goodfellow 2014]

Divergence $D(\mathbb{P}, \mathbb{Q})$ defined by maximizing a variational objective $\mathcal{G}$:

$$D(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \mathcal{G}(f, \mathbb{P}, \mathbb{Q})$$

- **Critic**: maximizes $\mathcal{G}(f, \mathbb{P}, \mathbb{Q})$ over $f \in \mathcal{F}$ to find optimal critic $f^\star$.
- **Generator**: minimizes $\mathcal{D}(\mathbb{P}, \mathbb{Q}) = \mathcal{G}(f^\star, \mathbb{P}, \mathbb{Q})$ over $\mathbb{Q}$.
- Recover the MMD when $\mathcal{F}$ is the unit ball in an RKHS $\mathcal{H}$.

Learning generative models using MMD

Goal is to solve the optimization problem:

$$\min_{\theta} MMD^2(\mathbb{P}, \mathbb{Q}_\theta)$$

Learning generative models using MMD

Goal is to solve the optimization problem:

$$\min_{\theta} MMD^2(\mathbb{P}, \mathbb{Q}_\theta)$$

1. Sample a mini-batch of i.i.d samples $X_1, ..., X_B \sim \mathbb{P}$ from data-set.
2. Sample a mini-batch of i.i.d. latent noise $Z_1, ..., Z_B \sim \mu$.
3. Generate IGM samples $Y_b = G_\theta(Z_b) \sim \mathbb{Q}_\theta$ for $1 \leq b \leq B$.
4. Compute empirical loss $\hat{\mathcal{L}}(\theta) := \widehat{MMD^2}(\mathbb{P}, \mathbb{Q}_\theta)$. (Differentiable in θ)
5. Update parameters of the model using SGD:

$$\theta \leftarrow \theta - \gamma \nabla \hat{\mathcal{L}}(\theta).$$

Learning generative models using MMD

IGM trained using an RBF kernel on MNIST dataset.

Need better image features.

- In practice, choice of the kernel is crucial for good performance.
- Hard to design a kernel for high dimensional data like images.
- Why not learning it?

Learning generative models using MMD

Goal is to solve the optimization problem:

$$\min_{\theta} \sup_{k \in \mathcal{K}} MMD_k^2(\mathbb{P}, \mathbb{Q}_\theta)$$

- $\mathcal{K}$ is a family of kernels,
  - ex: parmaterized by a neural network:

$$k(x, y) = h(\varphi(x), \varphi(y))$$

    where φ is a NN and h is a fixed p.d. kernel.
- Adaptively select an *MMD* that best discriminates between $\mathbb{P}$ and current model $\mathbb{Q}$.
- In practice, alternate between gradient steps on k and on θ: (Adversarial training).

Learning generative models using MMD

IGM trained on MNIST dataset.

Samples are better!

Learning generative models using MMD

IGM trained on CelebA dataset.

[A., Sutherland , Binkowski and Gretton, 2018]

Learning generative models using MMD

IGM trained on CelebA dataset.

[A., Sutherland , Binkowski and Gretton, 2018]

- More to the story: regularization, stability in optimization, evaluation, etc

Summary

- It is possible to represent probability distributions using kernels through the concept of mean embeddings.
- The maximum mean discrepancy (MMD), allows to compare probabilities by comparing their mean embeddings.
- MMD can be used for various applications:
  - Two sample tests
  - Learning implicit generative models (like GANs)
- Other applications include
  - Dependence detection
  - Feature selection
  - Bling source separaion (e.g. ICA)
- Often assume good kernels which do not discard information about distributions: characteristic kernels.

Outline

1. Characterizing probabilities with kernels
   - Kernel mean embedding
   - The Maximum Mean Discrepancy
   - Characteristic kernels

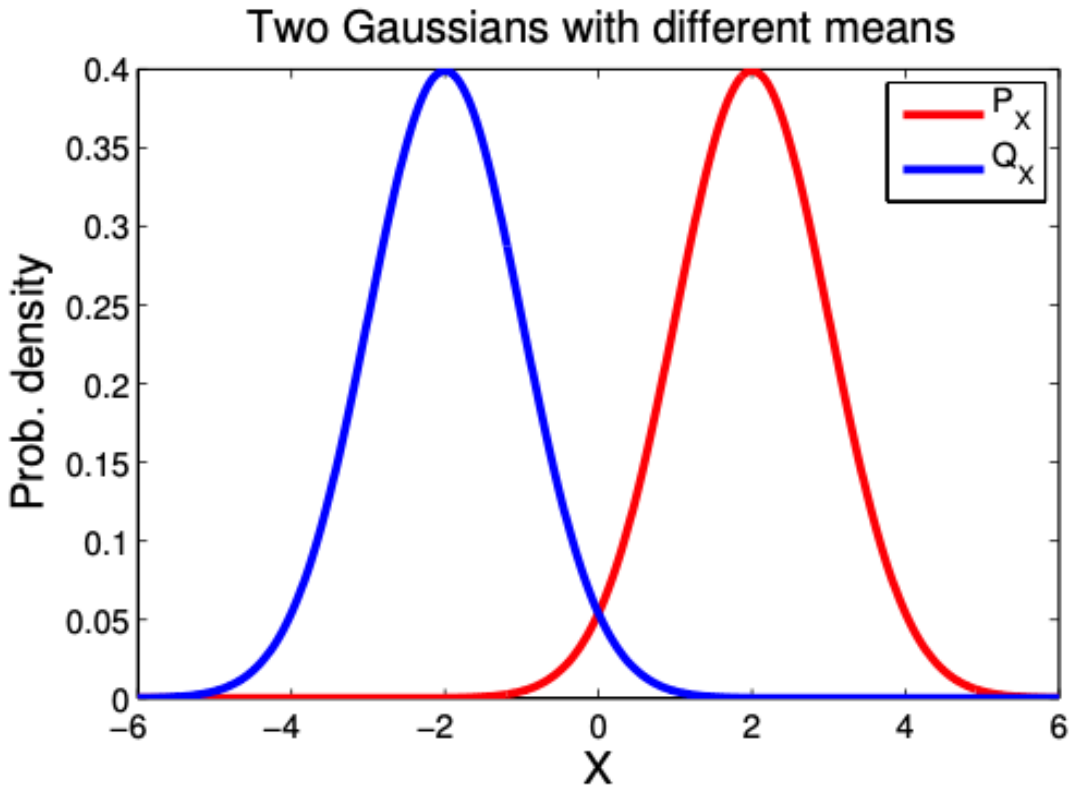
Can mean embeddings characterize probabilities?

Question: Given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ with mean embeddings $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$, can we confidently tell if $\mathbb{P}$ and $\mathbb{Q}$ are different or not based only on the summary given by $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$?

Can mean embeddings characterize probabilities?

Question: Given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ with mean embeddings $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$, can we confidently tell if $\mathbb{P}$ and $\mathbb{Q}$ are different or not based only on the summary given by $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$?

Example 1: Linear kernel $K(x, x') = x^\top x'$.



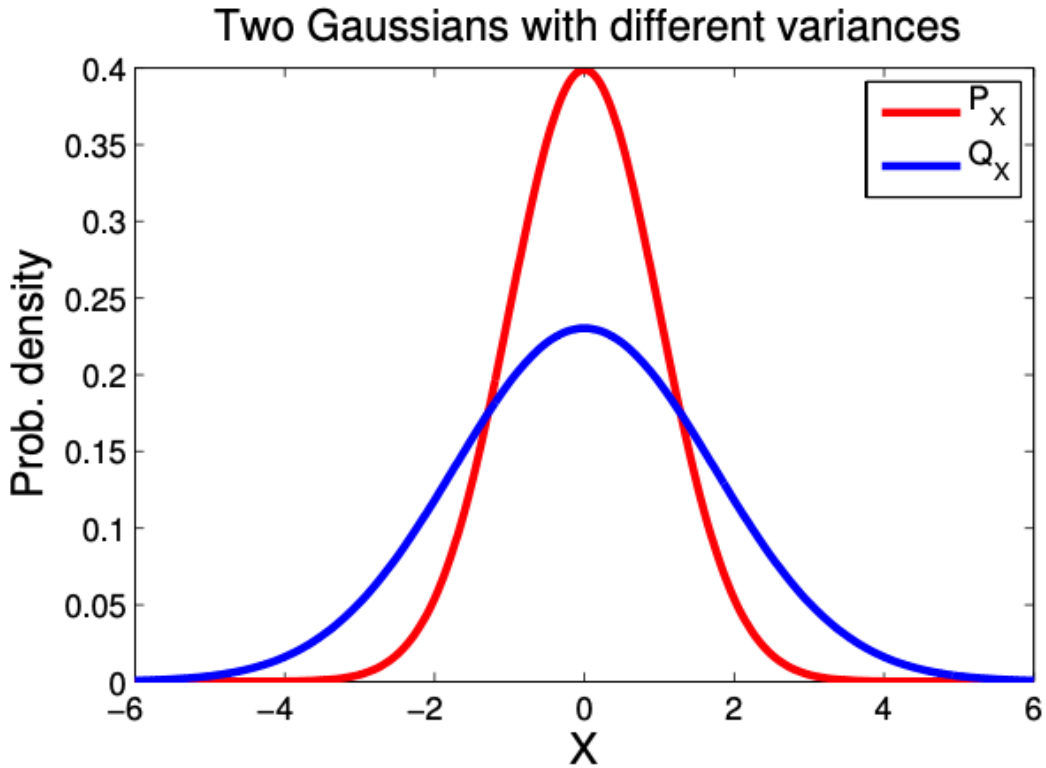
Compare

$$\mu_{\mathbb{P}}(x) = \mathbb{E}_{X \sim \mathbb{P}}[X]^\top x$$
$$\neq$$
$$\mu_{\mathbb{Q}}(x) = \mathbb{E}_{X \sim \mathbb{Q}}[X]^\top x$$

Can mean embeddings characterize probabilities?

Question: Given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ with mean embeddings $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$, can we confidently tell if $\mathbb{P}$ and $\mathbb{Q}$ are different or not based only on the summary given by $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$?

Example 1: Linear kernel $K(x, x') = x^\top x'$.

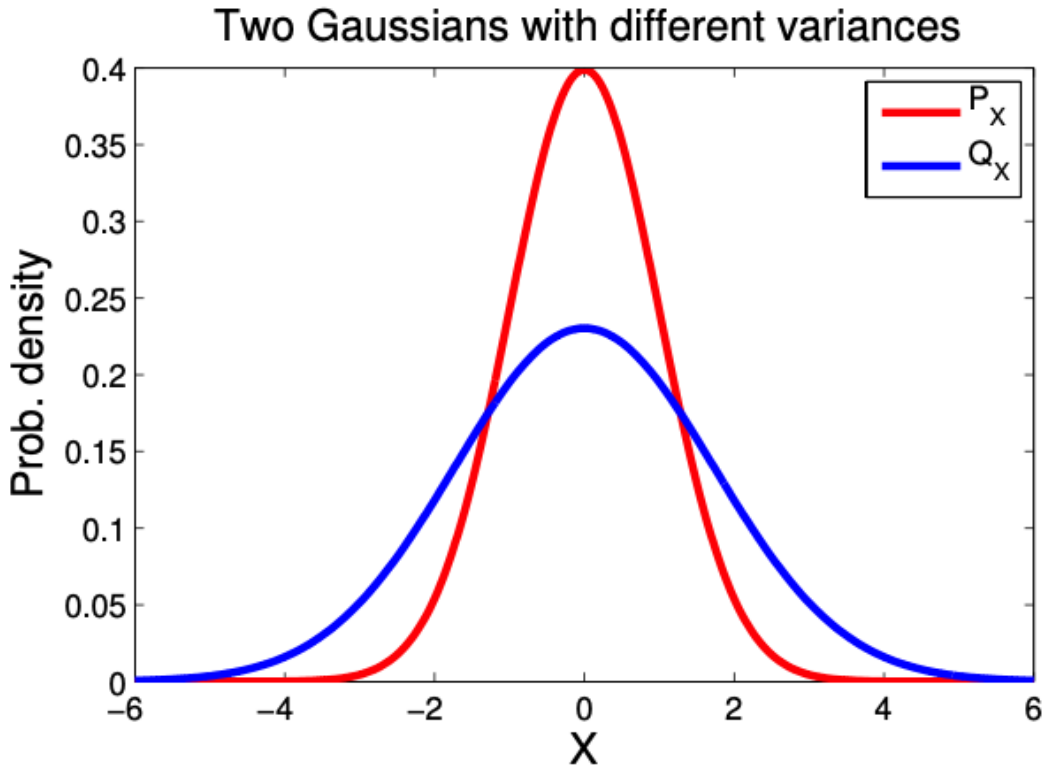


$$\mu_{\mathbb{P}}(x) = \mathbb{E}_{X\sim\mathbb{P}}[X]^\top x$$
$$=$$
$$\mu_{\mathbb{Q}}(x) = \mathbb{E}_{X\sim\mathbb{Q}}[X]^\top x$$

Can mean embeddings characterize probabilities?

Question: Given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ with mean embeddings $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$, can we confidently tell if $\mathbb{P}$ and $\mathbb{Q}$ are different or not based only on the summary given by $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$?

Example 2: Polynomial kernel $K(x, x') = (x^\top x')^2$.

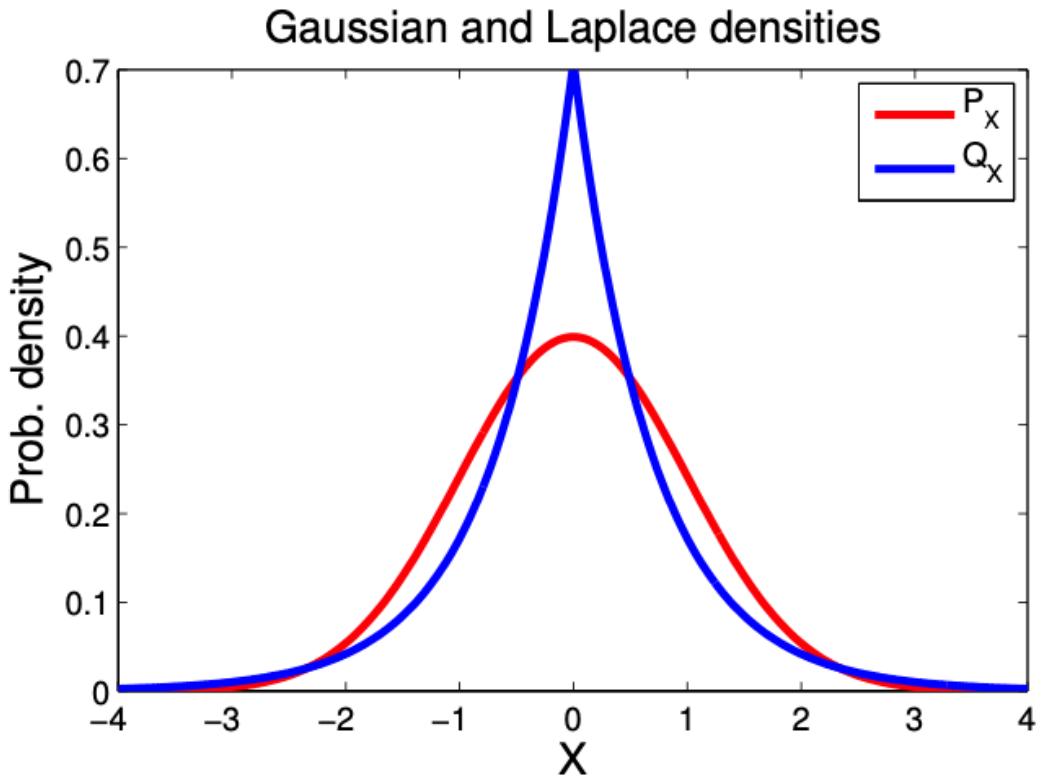


$$\mu_{\mathbb{P}}(x) = Tr(\mathbb{E}_{X\sim\mathbb{P}}[XX^\top]xx^\top)$$
$$\neq$$
$$\mu_{\mathbb{Q}}(x) = Tr(\mathbb{E}_{X\sim\mathbb{Q}}[XX^\top]xx^\top)$$

Can mean embeddings characterize probabilities?

Question: Given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ with mean embeddings $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$, can we confidently tell if $\mathbb{P}$ and $\mathbb{Q}$ are different or not based only on the summary given by $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$?

Example 2: Polynomial kernel of order 2: $K(x, x') = (x^\top x')^2$.



$$\mu_{\mathbb{P}}(x) = Tr(\mathbb{E}_{X\sim\mathbb{P}}[XX^\top]xx^\top)$$
$$=$$
$$\mu_{\mathbb{Q}}(x) = Tr(\mathbb{E}_{X\sim\mathbb{Q}}[XX^\top]xx^\top)$$

Can mean embeddings characterize probabilities?

Question: Are there kernels for which two mean embeddings $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ are equal iff $\mathbb{P} = \mathbb{Q}$?

Example 3: Exponential kernel $K(x, y) = \exp(x^\top y)$.

$$\mu_{\mathbb{P}}(y) = \mathbb{E}_{X \sim \mathbb{P}}[\exp(X^\top y)]$$

Can mean embeddings characterize probabilities?

Question: Are there kernels for which two mean embeddings $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ are equal iff $\mathbb{P} = \mathbb{Q}$?

Example 3: Exponential kernel $K(x, y) = \exp(x^\top y)$.

$$\mu_{\mathbb{P}}(y) = \underbrace{\mathbb{E}_{X \sim \mathbb{P}}[\exp(X^\top y)]}_{\text{Moment generating function}}$$

Can mean embeddings characterize probabilities?

Question: Are there kernels for which two mean embeddings $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ are equal iff $\mathbb{P} = \mathbb{Q}$?

Example 3: Exponential kernel $K(x, y) = \exp(x^\top y)$.

$$\mu_{\mathbb{P}}(y) = \underbrace{\mathbb{E}_{X\sim\mathbb{P}}[\exp(X^\top y)]}_{\text{Moment generating function}}$$

Classical result: If two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ have the same moment generating functions, then $\mathbb{P} = \mathbb{Q}$, meaning that:

$$\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)], \qquad \forall f \in \mathcal{C}_b(\mathcal{X}).$$

Can mean embeddings characterize probabilities?

Question: Are there kernels for which two mean embeddings $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ are equal iff $\mathbb{P} = \mathbb{Q}$?

Example 3: Exponential kernel $K(x, y) = \exp(x^\top y)$.

$$\mu_{\mathbb{P}}(y) = \underbrace{\mathbb{E}_{X\sim\mathbb{P}}[\exp(X^\top y)]}_{\text{Moment generating function}}$$

Classical result: If two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ have the same moment generating functions, then $\mathbb{P} = \mathbb{Q}$, meaning that:

$$\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)], \qquad \forall f \in \mathcal{C}_b(\mathcal{X}).$$

Intuitively: The RKHS and, in particular, the set of functions $\{K_y : x \mapsto exp(x^\top y)\}_{y\in\mathcal{X}}$ is rich enough so that $\mathbb{E}_{\mathbb{P}}[K_y(X)] = \mathbb{E}_{\mathbb{Q}}[K_y(X)]$ for all $y \in \mathcal{X}$ guarantees that $\mathbb{P} = \mathbb{Q}$.

Characteristic kernels

Definition

Let $\mathcal{X}$ be a topological set and $\mathcal{P}$ the set of Borel probability measures on $\mathcal{X}$. Consider a bounded measurable p.d. kernel K defined on $\mathcal{X}$ and let $\mathcal{H}$ be its RKHS. The kernel K is said to be characteristic if the map $\mathcal{P} \ni \mathbb{P} \mapsto \mu_{\mathbb{P}} = \mathbb{E}_{X\sim\mathbb{P}}[K_X] \in \mathcal{H}$ is injective, i.e.:

$$\forall \mathbb{P}, \mathbb{Q} \in \mathcal{P} : \mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \implies \mathbb{P} = \mathbb{Q}.$$

- Equality of mean embeddings $\iff$ equality of expectations of functions in $\mathcal{H}$, i.e.:

$$\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \iff \mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)], \qquad \forall f \in \mathcal{H}$$

Characteristic kernels

Definition

Let $\mathcal{X}$ be a topological set and $\mathcal{P}$ the set of Borel probability measures on $\mathcal{X}$. Consider a bounded measurable p.d. kernel K defined on $\mathcal{X}$ and let $\mathcal{H}$ be its RKHS. The kernel K is said to be characteristic if the map $\mathcal{P} \ni \mathbb{P} \mapsto \mu_{\mathbb{P}} = \mathbb{E}_{X\sim\mathbb{P}}[K_X] \in \mathcal{H}$ is injective, i.e.:

$$\forall \mathbb{P}, \mathbb{Q} \in \mathcal{P} : \mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \implies \mathbb{P} = \mathbb{Q}.$$

- Equality of mean embeddings $\iff$ equality of expectations of functions in $\mathcal{H}$, i.e.:

$$\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \iff \mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)], \qquad \forall f \in \mathcal{H}$$

- Equality of probability distributions $\iff$ Equality of expectations of continuous and bounded functions on $\mathcal{X}$, i.e.:

$$\mathbb{P} = \mathbb{Q} \iff \mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)], \qquad \forall f \in \mathcal{C}_b(\mathcal{X}).$$

Characteristic kernels

Definition

Let $\mathcal{X}$ be a topological set and $\mathcal{P}$ the set of Borel probability measures on $\mathcal{X}$. Consider a bounded measurable p.d. kernel K defined on $\mathcal{X}$ and let $\mathcal{H}$ be its RKHS. The kernel K is said to be characteristic if the map $\mathcal{P} \ni \mathbb{P} \mapsto \mu_{\mathbb{P}} = \mathbb{E}_{X\sim\mathbb{P}}[K_X] \in \mathcal{H}$ is injective, i.e.:

$$\forall \mathbb{P}, \mathbb{Q} \in \mathcal{P} : \mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \implies \mathbb{P} = \mathbb{Q}.$$

- Equality of mean embeddings $\iff$ equality of expectations of functions in $\mathcal{H}$, i.e.:

$$\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \iff \mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)], \qquad \forall f \in \mathcal{H}$$

- Equality of probability distributions $\iff$ Equality of expectations of continuous and bounded functions on $\mathcal{X}$, i.e.:

$$\mathbb{P} = \mathbb{Q} \iff \mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)], \qquad \forall f \in \mathcal{C}_b(\mathcal{X}).$$

- A kernel K is characteristic if RKHS $\mathcal{H}$ is rich enough!

Characteristic kernels via Universality

Definition

Let K be a p.d. kernel with RKHS $\mathcal{H}$ on a compact set $\mathcal{X}$. K is universal if $y \mapsto K(x, y)$ is continuous for all $x \in \mathcal{X}$ and $\mathcal{H}$ is dense in $\mathcal{C}(\mathcal{X})$ in the maximum norm $\|.\|_\infty$.

Proposition

Assume $\mathcal{X}$ is compact. If K is universal, then K is characteristic.

proof: Let $\mathbb{P}$ and $\mathbb{Q}$ such that $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}}$. We need to show that

$$\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)], \forall f \in \mathcal{C}(\mathcal{X}).$$

Fix $f \in \mathcal{C}(\mathcal{X})$. By universality of K, $\mathcal{H}$ is dense in $\mathcal{C}(\mathcal{X})$ in the sup norm. Hence, for any $\epsilon > 0$, there exists $g \in \mathcal{H}$ such that $\|f - g\|_\infty \leq \epsilon$.

Characteristic kernels via Universality

Proof Next we make the expansion

$$\begin{aligned}|\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)]| \leq & |\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{X\sim\mathbb{P}}[g(X)]| \\ & + |\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)] - \mathbb{E}_{Y\sim\mathbb{Q}}[g(Y)]| \\ & + |\mathbb{E}_{X\sim\mathbb{P}}[g(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[g(Y)]|.\end{aligned}$$

The first two terms are upper-bounded by ϵ by definition of g.The last term is equal to 0 since $\mathbb{E}_{X\sim\mathbb{P}}[g(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[g(Y)] = \langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\rangle_{\mathcal{H}}$ and $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}}$ by assumption.
Hence, we have shown that for any $\epsilon > 0$:

$$|\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)]| \leq 2\epsilon$$

directly implying that $|\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)]| = 0$.
The above holds for any $f \in \mathcal{C}(\mathcal{X})$, meaning that $\mathbb{P} = \mathbb{Q}$.

Criteria for Universality

Proposition

Let $0 < r \leq \infty$ and $f : (-r, r) \to \mathbb{R}$ be a C^∞ function that admits an expansion as a Taylor series in 0: $f(x) = \sum_{i=0}^{\infty} a_i x^i$. Let $\mathcal{X}$ be a compact set in the open centered ball in $\mathbb{R}^d$ of radius $\sqrt{r}$. If $a_i > 0$ for all $i \geq 0$, then $k(x, y) = f(\langle x, y \rangle)$ defines a universal kernel on $\mathcal{X}$.

Example 1: Exp kernel: $K(x, y) = \exp \langle x, y \rangle$ on any compact $\mathcal{X}$.

$$f(x) = exp(x) = \sum_{i=0}^{\infty} \frac{1}{i!} x^i, \qquad K(x, y) = f(\langle x, y \rangle).$$

Example 2: Gaussian kernel on the Unit Sphere
$K(x, y) = \exp\left(-\frac{1}{2}\|x - y\|^2\right)$.

$$f(x) = e^{-1} \exp(x) = e^{-1} \sum_{i=0}^{\infty} \frac{1}{i!} x^i, \qquad K(x, y) = f(\langle x, y \rangle).$$

Criteria for Universality

Proposition (Steinwart 2001)

Let $f : [0, 2\pi] \to \mathbb{R}$ be a continuous function that can be expanded in a pointwise absolutely convergent Fourier series: $f(t) = \sum_{n=0}^{\infty} a_n cos(nt)$. If $a_n > 0$ for all $n \geq 0$, then the Kernel $K(x, y) := \prod_{i=1}^{d} f(|x_i - y_i|)$ defines a universal kernel on every compact subset of $[0, 2\pi)^d$.

Example 1: The stronger regularized Fourier kernel (Vapnik 1998, p.470)

$$k(x, y) = (1 - q^2)/(2 - 4qcos(x - y) + 2q^2)$$

for any $0 < q < 1$.

Just in case ...

Definition

Let A be a vector space and $\times : A \times A \to A$ be a binary operation on A. Then A is an algebra if $\times$ is bilinear, i.e. for all $x, y, z \in A$ and $a, b \in \mathbb{R}$:

$$z \times (x + y) = z \times x + z \times y$$
$$(x + y) \times z = x \times z + y \times z$$
$$(ax) \times (by) = (ab)(x \times y).$$

Theorem: Stone-Weierstrass

Let $(\mathcal{X}, d)$ be a compact metric space and A a linear subspace of $\mathcal{C}(\mathcal{X})$. Then A is dense in $\mathcal{C}(\mathcal{X})$ if

- A is an algebra for the product of functions.
- A does not vanish: For all $x \in \mathcal{X}$, there exists $f \in A$ s.t. $f(x) \neq 0$.
- A separates points: For all $x, y \in \mathcal{X}$ with $x \neq y$, there exists $f \in A$, s.t. $f(x) \neq f(y)$.

General criterion for Universality

Theorem: General criterion for universality (Steinwart, 2001)

Let $\mathcal{X}$ be a compact metric space and k be a continuous kernel on $\mathcal{X}$ with $k(x,x) > 0$. Suppose there is an injective map $\Phi(x) = \{\varphi_i(x)\}_{i\geq 0}$ such that $k(x,y) = \sum_{i=0}^{\infty} \varphi_i(x)\varphi_i(y)$. If the set $A := span\{\varphi_i | i \geq 0\}$ is an algebra, then k is universal.

Proof:

- A is a subset of $\mathcal{C}(\mathcal{X})$. Follows by continuity of the map $x \mapsto \Phi(x)$. Indeed, $\|\Phi(x)-\Phi(y)\|^2 = K(x,x)+K(y,y)-2K(x,y) \leq \epsilon$ for any $\epsilon > 0$ provided that y is close enough to x since K is continuous.
- A does not vanish. Otherwise, we can find x such that $\varphi_i(x) = 0$ for all $i \geq 0$, meaning that $K(x,x) = 0$: contradicts $K(x,x) > 0$.
- A separates points. Otherwise, there exists x, y with $x \neq y$ and $\varphi_i(x) = \varphi_i(y)$ for all $i \geq 0$, hence $\Phi(x) = \Phi(y)$: contradicts Φ injective.

Hence A is dense in $\mathcal{C}(\mathcal{X})$ by Stone-Weierstrass theorem.

General criterion for Universality

Theorem: General criterion for universality (Steinwart, 2001)

Let $\mathcal{X}$ be a compact metric space and k be a continuous kernel on $\mathcal{X}$ with $k(x,x) > 0$. Suppose there is an injective map $\Phi(x) = \{\varphi_i(x)\}_{i\geq 0}$ such that $k(x,y) = \sum_{i=0}^{\infty} \varphi_i(x)\varphi_i(y)$. If the set $A := span\{\varphi_i | i \geq 0\}$ is an algebra, then k is universal.

Proof Continued: Let $f \in \mathcal{C}(\mathcal{X})$ and $\epsilon > 0$.

- Since A is dense in $\mathcal{C}(\mathcal{X})$, there exists $g \in A$ s.t. $\|f - g\|_\infty < \epsilon$.
- By definition of A, the function g is of the form $g(x) = \langle w, \Phi(x)\rangle_{l_2}$ with $w = (w_i)_{i\geq 0}$ s.t. $w_i = 0$ for any $i > N$ for some $N < \infty$.
- Hence, g belongs to the unique RKHS $\mathcal{H}$ of K. This shows that $\mathcal{H}$ is dense in $\mathcal{C}(\mathcal{X})$, hence K is universal.

Criteria for Universality

Proposition

Let $0 < r \leq \infty$ and $f : (-r, r) \to \mathbb{R}$ be a C^∞ function that admits an expansion as a Taylor series in 0: $f(x) = \sum_{i=0}^{\infty} a_i x^i$. Let $\mathcal{X}$ be a compact set in the open centered ball in $\mathbb{R}^d$ of radius $\sqrt{r}$. If $a_i > 0$ for all $i \geq 0$, then $k(x, y) = f(\langle x, y \rangle)$ defines a universal kernel on $\mathcal{X}$.

Proof: For simplicity, take $d = 1$.

- K is continuous and of the form:

$$K(x, y) := \sum_{i=0}^{\infty} a_i x^i y^i = \langle \Phi(x), \Phi(y) \rangle_{l_2}$$

  with $\Phi(x) = (\sqrt{a_i} x^i)_{i \geq 0}$ which is injective.
- $K(x, x) = \sum_{i=0}^{\infty} a_i x^{2i} > 0$ since $a_i > 0$ for all $i \geq 0$.
- $A := span(\{\varphi_n | n \geq 0\})$ is the algebra of polynomials.
- Hence K universal by the general criterion for universality.

Criteria for Universality

Proposition (Steinwart 2001)

Let $f : [0, 2\pi] \to \mathbb{R}$ be a continuous function that can be expanded in a pointwise absolutely convergent Fourier series: $f(t) = \sum_{n=0}^{\infty} a_n cos(nt)$. If $a_n > 0$ for all $n \geq 0$, then the Kernel $K(x, y) := \prod_{i=1}^{d} f(|x_i - y_i|)$ defines a universal kernel on every compact subset of $[0, 2\pi)^d$.

Proof: For simplicity, take $d=1$.

- K is continuous and of the form:

$$K(x, y) = a_0 + \sum_{n=0}^{\infty} a_n (sin(nx)sin(ny) + cos(nx)cos(ny)) = \langle \Phi(x), \Phi(y) \rangle_{l_2}$$

  where $\Phi(x) = (\varphi_n(x))_{n \geq 0}$ defined by $\varphi_0(x) = a_0$, $\varphi_{2n-1} = \sqrt{a_n} sin(nx)$ and $\varphi_{2n}(x) = \sqrt{a_n} cos(nx)$ for $n \geq 1$ is injective.
- $K(x, x) = \sum_{n=0}^{\infty} a_n > 0$ since $a_n > 0$ for all $n \geq 0$.
- $A := span(\{\varphi_n | n \geq 0\})$ is an algebra (by trigonometric identities).
- Hence K universal by the general criterion for universality.

Summary: Characteristic kernels via Universality

- On a compact metric set $\mathcal{X}$, a universal kernel is a continuous kernel whose RKHS (H) is dense in $\mathcal{C}(\mathcal{X})$ in the maximum norm.
- Any universal kernel on $\mathcal{X}$ is characteristic, i.e. the mean embedding map $\mathbb{P} \mapsto \mu_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}[K_X] \in \mathcal{H}$ defined on the set $\mathcal{P}$ of probability distributions on $\mathcal{X}$ is injective:

$$\forall \mathbb{P}, \mathbb{Q} \in \mathcal{P} : \mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \implies \mathbb{P} = \mathbb{Q}.$$

- Can construct a large class of universal kernels using Taylor series or Fourier series with positive coefficients.
- Both constructions follow from the General criterion for universality, itself a consequence of Stone-Weierstrass theorem for compact metric sets.
- Question: What if $\mathcal{X}$ is not compact?

Characteristic kernels via Fourier transform

- Consider a translation invariant kernel K on $\mathbb{R}^d$ of the form $K(x, y)=\kappa(x - y)$ with $\kappa : \mathbb{R}^d \to \mathbb{R}$.
- Bochner's theorem implies the existence of a finite non-negative Borel measure Λ on $\mathbb{R}^d$ such that $\kappa(z) = \int e^{-iz^\top w} d\Lambda(w)$.
- Can express K as a Hermitian product in $L_2(\Lambda)$ of Fourier features:

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{L_2(\Lambda)}, \qquad w \mapsto \Phi(x)(w) = e^{-ix^\top w}$$

- Can express the mean embedding $\mu_{\mathbb{P}}$ in terms of $\mathcal{F}(\mathbb{P})=\mathbb{E}_{X\sim\mathbb{P}}[\Phi(X)]$ the of Fourier transform of $\mathbb{P}$:

$$\mu_{\mathbb{P}}(y) = \mathbb{E}_{X\sim\mathbb{P}}[\langle \Phi(X), \Phi(y) \rangle_{L_2(\Lambda)}] = \langle \mathcal{F}(\mathbb{P}), \Phi(y) \rangle_{L_2(\Lambda)}$$

Fourier inversion theorem (Dudley 2002, Theorem 9.5.4)

If $\mathbb{P}$ and $\mathbb{Q}$ are two probability distributions on $\mathbb{R}^d$ with the same Fourier transform: $\mathcal{F}(\mathbb{P})=\mathcal{F}(\mathbb{Q})$, then $\mathbb{P}=\mathbb{Q}$.

The measure Λ must "preserve information contained" in the Fourier transform $\mathcal{F}(\mathbb{P})$.

Characteristic kernels via Fourier transform

Translation invariant characteristic kernels: (Sriperumbudur 2008)

Let K be a translation invariant kernel on $\mathbb{R}^d$ of the form $K(x,y)=\kappa(x-y)$ with $\kappa(z) = \int e^{-iz^\top w} d\Lambda(w)$ for some finite non-negative Borel measure Λ on $\mathbb{R}^d$. The kernel K is characteristic if and only if $\operatorname{supp}(\Lambda) = \mathbb{R}^d$.

Example 1: Gaussian kernel $K(x,y) = e^{-\frac{\sigma^2}{2}\|x-y\|^2}$. The measure Λ is a gaussian on $\mathbb{R}^d$ with density $w \mapsto \sqrt{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\|w\|^2}$. Since $supp(\Lambda) = \mathbb{R}^d$, K is characteristic.

Example 2: Let $\kappa(z) = \|z\|^{-\frac{d}{2}} J_{d/2}(\|z\|)$, where J_p is Bessel's function of the first kind. Then $K(x,y) = \kappa(x-y)$ is not characteristic: Λ is the uniform distribution on the unit ball.

Characteristic kernels via Fourier transform

- Bochner's theorem: A translation invariant kernel on $\mathbb{R}^d$ is characterized by unique finite non-negative measure Λ
- Main result: K is characteristic if and only if $supp(\Lambda) = \mathbb{R}^d$.
- Similar reasoning can be applied to any space where Bochner's theorem holds:
  - Locally compact Abelian groups
  - Compact, non-Abelian groups (orthogonal matrices)
  - The semigroup $\mathbb{R}^d_+$.

Characteristic kernels: Summary

Definition

Let $\mathcal{X}$ be a topological set and $\mathcal{P}$ the set of Borel probability measures on $\mathcal{X}$. Consider a bounded measurable p.d. kernel K defined on $\mathcal{X}$ and let $\mathcal{H}$ be its RKHS. The kernel K is said to be characteristic if the map $\mathcal{P} \ni \mathbb{P} \mapsto \mu_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}[K_X] \in \mathcal{H}$ is injective, i.e.:

$$\forall \mathbb{P}, \mathbb{Q} \in \mathcal{P} : \mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \implies \mathbb{P} = \mathbb{Q}.$$

Criteria for characteristic kernels

- On a compact set $\mathcal{X}$, can use criteria for universality: A kernel is universal if it continuous and its RKHS is dense in $\mathcal{C}(\mathcal{X})$.
  - If K admits a Taylor expansion with positive coefficients.
  - If K admits a Fourier expansion with positive coefficients.
- If $\mathcal{X} = \mathbb{R}^d$ and K is translation invariant with associated non-negative measure Λ: K characteristic $\iff \operatorname{supp}(\Lambda) = \mathbb{R}^d$

References I