RKHS in ML: Comparing Two Samples

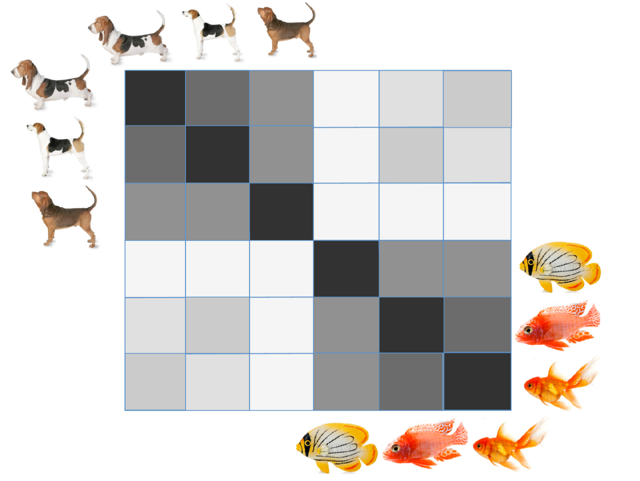
Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

November 2, 2024

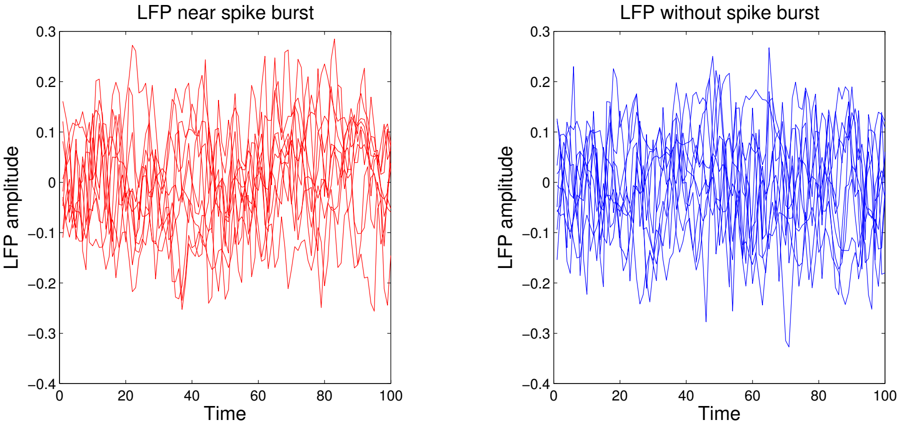
Comparing two samples

- Given: Samples from unknown distributions P and Q.
- Goal: do P and Q differ?



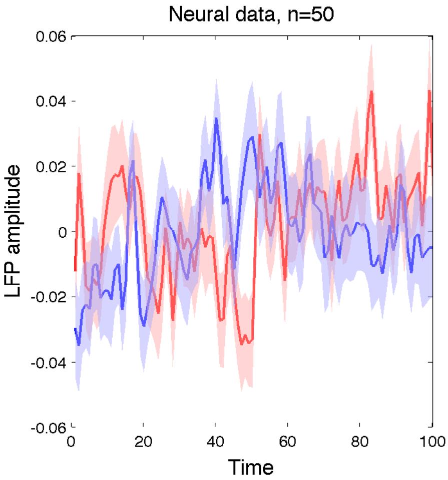
A real-life example: two-sample tests

- The problem:Do local field potential (LFP) signals change when measured near a spike burst?



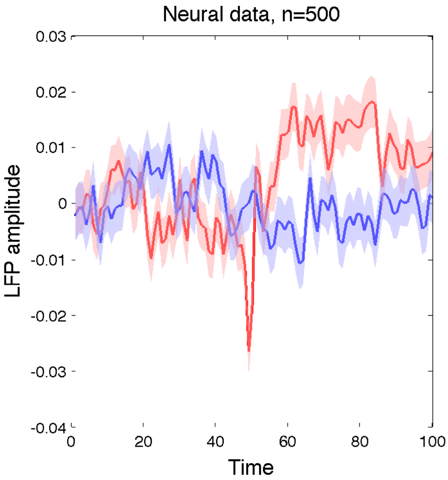
A real-life example: two-sample tests

- The problem:Do local field potential (LFP) signals change when measured near a spike burst?



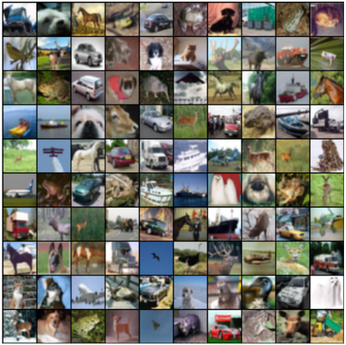
A real-life example: two-sample tests

- The problem:Do local field potential (LFP) signals change when measured near a spike burst?

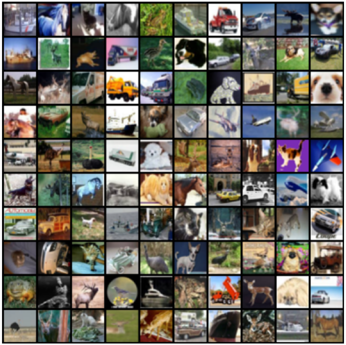


A real-life example: two-sample tests

- Goal: do P and Q differ?



CIFAR 10 samples



Cifar 10.1 samples

Significant difference?

Feng, Xu, Lu, Zhang, G., Sutherland, Learning Deep Kernels for Non-Parametric Two-Sample Tests, ICML 2020

Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017.

A real-life example: discrete domains

How do you compare distributions in a discrete domain?

X_1: Now disturbing reports out of Newfoundland show that the fragile snow crab industry is in serious decline. First the west coast salmon, the east coast salmon and the cod, and now the snow crabs off Newfoundland.

X_2: To my pleasant surprise he responded that he had personally visited those wharves and that he had already announced money to fix them. What wharves did the minister visit in my riding and how much additional funding is he going to provide for Delaps Cove, Hampton, Port Lorne,

...

$$P_X \stackrel{?}{=} Q_Y$$

Y_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.

Y_2: On the grain transportation system we have had the Estey report and the Kroeger report. We could go on and on. Recently programs have been announced over and over by the government such as money for the disaster in agriculture on the prairies and across Canada.

...

Are the gray extracts from the same distribution as the pink ones?

Outline

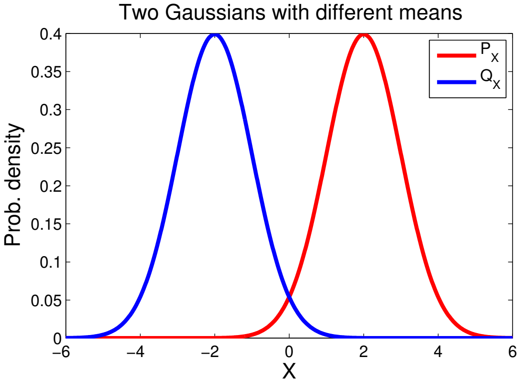
Two sample testing

- Test statistic: Maximum Mean Discrepancy (MMD)...
  - ...as a difference in feature means
  - ...as an integral probability metric (not just a technicality!)
- Statistical testing with the MMD
- "How to choose the best kernel"
  - when are feature means unique?
  - what kernel gives the most powerful test?

Maximum Mean Discrepancy

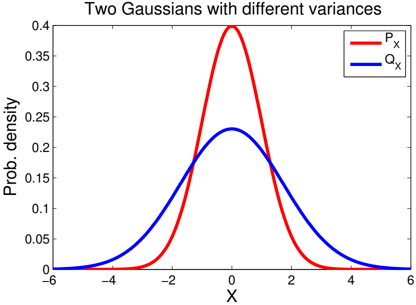
Feature mean difference

- Simple example: 2 Gaussians with different means
- Answer: t-test



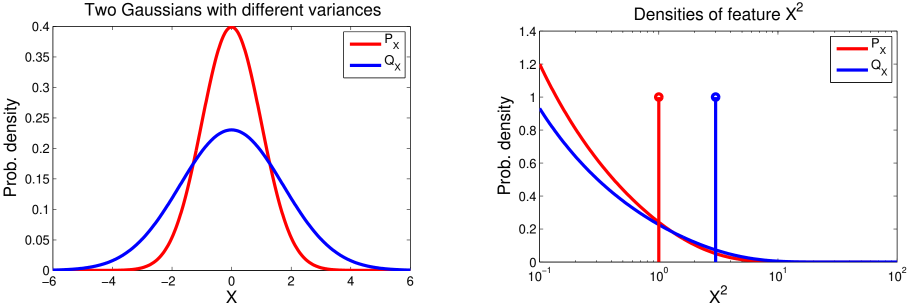
Feature mean difference

- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$



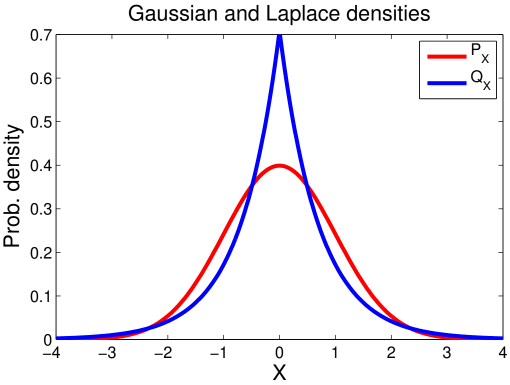
Feature mean difference

- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$



Feature mean difference

- Gaussian and Laplace distributions
- Same mean *and* same variance
- Difference in means using higher order features...RKHS



Infinitely many features using kernels

Kernels: dot products of features

Feature map $\varphi(x) \in \mathcal{F}$,

$$\varphi(x) = [\dots \varphi_i(x) \dots] \in \ell_2$$

For positive definite k,

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Infinitely many features $\varphi(x)$, dot product in closed form!

Infinitely many features using kernels

Kernels: dot products of features

Feature map $\varphi(x) \in \mathcal{F}$,

$$\varphi(x) = [\ldots \varphi_i(x) \ldots] \in \ell_2$$

For positive definite k,

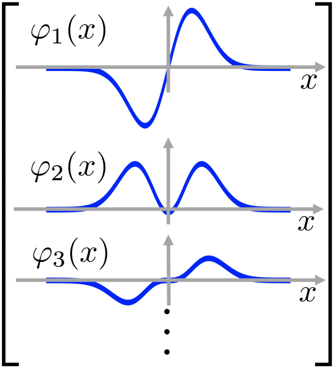
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Infinitely many features $\varphi(x)$, dot product in closed form!

Exponentiated quadratic kernel

$$k(x, x') = \exp\left(-\gamma \left\| x - x' \right\|^2\right)$$

$$\varphi(x) =$$



Features: Gaussian Processes for Machine learning, Rasmussen and Williams, Ch. 4.

Infinitely many features of distributions

Given P a Borel probability measure on $\mathcal{X}$, define feature map of probability P,

$$\mu_P = [\ldots \mathrm{E}_P [\varphi_i(X)] \ldots]$$

For positive definite $k(x, x')$,

$$\langle \mu_P, \mu_Q \rangle_{\mathcal{F}} = \mathrm{E}_{P,Q} k(x, y)$$

for $x \sim P$ and $y \sim Q$.

Infinitely many features of distributions

Given P a Borel probability measure on $\mathcal{X}$, define feature map of probability P,

$$\mu_P = [\ldots \mathrm{E}_P [\varphi_i(X)] \ldots]$$

For positive definite $k(x, x')$,

$$\langle \mu_P, \mu_Q \rangle_{\mathcal{F}} = \mathrm{E}_{P,Q} k(x, y)$$

for $x \sim P$ and $y \sim Q$.

Expectations of RKHS functions

Function evaluation in an RKHS:

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}}$$

Expectation evaulation in an RKHS:

$$\mathrm{E}_P(f(X)) = \langle f, \mu_P \rangle_{\mathcal{F}}$$

$$\boxed{\mu_P \text{ gives you expectations of all RKHS functions}}$$

Empirical mean embedding:

$$\widehat{\mu}_P = m^{-1} \sum_{i=1}^{m} \varphi_{x_i} \qquad x_i \overset{\text{i.i.d.}}{\sim} P$$

... does this reasoning work in infinite dimensions?

Expectations of RKHS functions

Function evaluation in an RKHS:

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}}$$

Expectation evaulation in an RKHS:

$$\mathrm{E}_P(f(X)) = \langle f, \mu_P \rangle_{\mathcal{F}}$$

μ_P gives you expectations of all RKHS functions

Empirical mean embedding:

$$\widehat{\mu}_P = m^{-1} \sum_{i=1}^{m} \varphi_{x_i} \qquad x_i \overset{\text{i.i.d.}}{\sim} P$$

... does this reasoning work in infinite dimensions?

Does the feature space mean exist?

Does there exist an element $\mu_P \in \mathcal{F}$ such that

$$\mathrm{E}_P f(x) = \langle f, \mu_P \rangle_{\mathcal{F}} \qquad \forall f \in \mathcal{F}$$

We recall the concept of a bounded operator: a linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when

$$|Af| \leq \lambda_A \|f\|_{\mathcal{F}} \quad \forall f \in \mathcal{F}.$$

Riesz representation theorem: In a Hilbert space $\mathcal{F}$, all bounded linear operators A can be written $\langle \cdot, g_A \rangle_{\mathcal{F}}$, for some $g_A \in \mathcal{F}$,

$$Af = \langle f(\cdot), g_A(\cdot) \rangle_{\mathcal{F}}$$

Does the feature space mean exist?

Does there exist an element $\mu_P \in \mathcal{F}$ such that

$$\mathrm{E}_P f(x) = \langle f, \mu_P \rangle_{\mathcal{F}} \qquad \forall f \in \mathcal{F}$$

We recall the concept of a bounded operator: a linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when

$$|Af| \leq \lambda_A \|f\|_{\mathcal{F}} \quad \forall f \in \mathcal{F}.$$

Riesz representation theorem: In a Hilbert space $\mathcal{F}$, all bounded linear operators A can be written $\langle \cdot, g_A \rangle_{\mathcal{F}}$, for some $g_A \in \mathcal{F}$,

$$Af = \langle f(\cdot), g_A(\cdot) \rangle_{\mathcal{F}}$$

Does the feature space mean exist?

Does there exist an element $\mu_P \in \mathcal{F}$ such that

$$\mathrm{E}_P f(x) = \langle f, \mu_P \rangle_{\mathcal{F}} \qquad \forall f \in \mathcal{F}$$

We recall the concept of a bounded operator: a linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when

$$|Af| \leq \lambda_A \|f\|_{\mathcal{F}} \quad \forall f \in \mathcal{F}.$$

Riesz representation theorem: In a Hilbert space $\mathcal{F}$, all bounded linear operators A can be written $\langle \cdot, g_A \rangle_{\mathcal{F}}$, for some $g_A \in \mathcal{F}$,

$$Af = \langle f(\cdot), g_A(\cdot) \rangle_{\mathcal{F}}$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathrm{E}_P\sqrt{k(x,x)} = \mathrm{E}_P\|\varphi(x)\|_{\mathcal{F}} < \infty$ then $\exists \mu_P \in \mathcal{F}$.

Proof:

The linear operator $T_P f := \mathrm{E}_P f(x)$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$\begin{aligned} |T_P f| &= |\mathrm{E}_P f(x)| \\ &\leq \mathrm{E}_P |f(x)| \\ &= \mathrm{E}_P |\langle f, \varphi(x)\rangle_{\mathcal{F}}| \\ &\leq \mathrm{E}_P\left(\sqrt{k(x,x)}\,\|f\|_{\mathcal{F}}\right) \end{aligned}$$

Hence by Riesz (with $\lambda_{T_P} = \mathrm{E}_P\sqrt{k(x,x)}$), $\exists \mu_P \in \mathcal{F}$ such that

$$T_P f = \langle f, \mu_P\rangle_{\mathcal{F}}.$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathrm{E}_P\sqrt{k(x,x)} = \mathrm{E}_P \|\varphi(x)\|_{\mathcal{F}} < \infty$ then $\exists \mu_P \in \mathcal{F}$.

Proof:

The linear operator $T_P f := \mathrm{E}_P f(x)$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$
\begin{aligned}
|T_P f| &= |\mathrm{E}_P f(x)|\,. \\
&\leq \mathrm{E}_P |f(x)| \\
&= \mathrm{E}_P |\langle f, \varphi(x)\rangle_{\mathcal{F}}| \\
&\leq \mathrm{E}_P \left(\sqrt{k(x,x)}\,\|f\|_{\mathcal{F}}\right)
\end{aligned}
$$

Hence by Riesz (with $\lambda_{T_P} = \mathrm{E}_P\sqrt{k(x,x)}$), $\exists \mu_P \in \mathcal{F}$ such that

$$T_P f = \langle f, \mu_P\rangle_{\mathcal{F}}\,.$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathrm{E}_P\sqrt{k(x,x)} = \mathrm{E}_P\,\|\varphi(x)\|_{\mathcal{F}} < \infty$ then $\exists \mu_P \in \mathcal{F}$.

Proof:

The linear operator $T_P f := \mathrm{E}_P f(x)$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$\begin{aligned}
|T_P f| &= |\mathrm{E}_P f(x)|\,. \\
&\leq \mathrm{E}_P\,|f(x)| \\
&= \mathrm{E}_P\,|\langle f, \varphi(x)\rangle_{\mathcal{F}}| \\
&\leq \mathrm{E}_P\left(\sqrt{k(x,x)}\,\|f\|_{\mathcal{F}}\right)
\end{aligned}$$

Hence by Riesz (with $\lambda_{T_P} = \mathrm{E}_P\sqrt{k(x,x)}$), $\exists \mu_P \in \mathcal{F}$ such that

$$T_P f = \langle f, \mu_P\rangle_{\mathcal{F}}\,.$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathrm{E}_P\sqrt{k(x,x)} = \mathrm{E}_P \|\varphi(x)\|_{\mathcal{F}} < \infty$ then $\exists \mu_P \in \mathcal{F}$.

Proof:

The linear operator $T_P f := \mathrm{E}_P f(x)$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$\begin{aligned} |T_P f| &= |\mathrm{E}_P f(x)| \,. \\ &\leq \mathrm{E}_P |f(x)| \\ &= \mathrm{E}_P |\langle f, \varphi(x)\rangle_{\mathcal{F}}| \\ &\leq \mathrm{E}_P \left(\sqrt{k(x,x)} \|f\|_{\mathcal{F}} \right) \end{aligned}$$

Hence by Riesz (with $\lambda_{T_P} = \mathrm{E}_P\sqrt{k(x,x)}$), $\exists \mu_P \in \mathcal{F}$ such that

$$T_P f = \langle f, \mu_P \rangle_{\mathcal{F}} \,.$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathrm{E}_P\sqrt{k(x,x)} = \mathrm{E}_P \|\varphi(x)\|_{\mathcal{F}} < \infty$ then $\exists \mu_P \in \mathcal{F}$.

Proof:

The linear operator $T_P f := \mathrm{E}_P f(x)$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$
\begin{aligned}
|T_P f| &= |\mathrm{E}_P f(x)| . \\
&\leq \mathrm{E}_P |f(x)| \\
&= \mathrm{E}_P |\langle f, \varphi(x)\rangle_{\mathcal{F}}| \\
&\leq \mathrm{E}_P \left(\sqrt{k(x,x)} \|f\|_{\mathcal{F}} \right)
\end{aligned}
$$

Hence by Riesz (with $\lambda_{T_P} = \mathrm{E}_P\sqrt{k(x,x)}$), $\exists \mu_P \in \mathcal{F}$ such that

$$T_P f = \langle f, \mu_P \rangle_{\mathcal{F}} .$$

μ_P as a function in the RKHS

Embedding of P to feature space

- Mean embedding $\mu_P \in \mathcal{F}$,

$$\langle \mu_P, f \rangle_{\mathcal{F}} = E_P f(x).$$

- What does prob. feature map look like?

$$\begin{aligned}\mu_P(t) &= \langle \mu_P, \varphi(t) \rangle_{\mathcal{F}} \\ &= \langle \mu_P, k(\cdot, t) \rangle_{\mathcal{F}} \\ &= \mathrm{E}_{x \sim P} k(x, t)\end{aligned}$$

Expectation of kernel!

μ_P as a function in the RKHS

Embedding of P to feature space

- Mean embedding $\mu_P \in \mathcal{F}$,

$$\langle \mu_P, f \rangle_{\mathcal{F}} = E_P f(x).$$

- What does prob. feature map look like?

$$\begin{aligned} \mu_P(t) &= \langle \mu_P, \varphi(t) \rangle_{\mathcal{F}} \\ &= \langle \mu_P, k(\cdot, t) \rangle_{\mathcal{F}} \\ &= \mathrm{E}_{x\sim P} k(x, t) \end{aligned}$$

Expectation of kernel!

μ_P as a function in the RKHS

Embedding of P to feature space

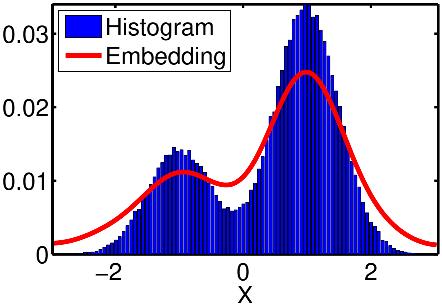
- Mean embedding $\mu_P \in \mathcal{F}$,

$$\langle \mu_P, f \rangle_{\mathcal{F}} = E_P f(x).$$

- What does prob. feature map look like?

$$\begin{aligned} \mu_P(t) &= \langle \mu_P, \varphi(t) \rangle_{\mathcal{F}} \\ &= \langle \mu_P, k(\cdot, t) \rangle_{\mathcal{F}} \\ &= \mathrm{E}_{x \sim P} k(x, t) \end{aligned}$$

Expectation of kernel!



The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\
&= \underbrace{\mathrm{E}_P k(x, x')}_{\text{(a)}} + \underbrace{\mathrm{E}_Q k(y, y')}_{\text{(a)}} - \underbrace{2\mathrm{E}_{P,Q} k(x, y)}_{\text{(b)}}
\end{aligned}
$$

(a)= within distrib. similarity, (b)= cross-distrib. similarity.

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\
&= \underbrace{\mathrm{E}_P k(x, x')}_{\text{(a)}} + \underbrace{\mathrm{E}_Q k(y, y')}_{\text{(a)}} - 2\underbrace{\mathrm{E}_{P,Q} k(x, y)}_{\text{(b)}}
\end{aligned}
$$

Proof:

$$
\begin{aligned}
\|\mu_P - \mu_Q\|^2_{\mathcal{F}} &= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\
&= \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2 \langle \mu_P, \mu_Q \rangle
\end{aligned}
$$

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\
&= \underbrace{\mathrm{E}_P k(x, x')}_{\text{(a)}} + \underbrace{\mathrm{E}_Q k(y, y')}_{\text{(a)}} - 2\underbrace{\mathrm{E}_{P,Q} k(x, y)}_{\text{(b)}}
\end{aligned}
$$

Proof:

$$
\begin{aligned}
\|\mu_P - \mu_Q\|^2_{\mathcal{F}} &= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\
&= \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2 \langle \mu_P, \mu_Q \rangle
\end{aligned}
$$

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\
&= \underbrace{\mathrm{E}_P k(x, x')}_{\text{(a)}} + \underbrace{\mathrm{E}_Q k(y, y')}_{\text{(a)}} - \underbrace{2\mathrm{E}_{P,Q} k(x, y)}_{\text{(b)}}
\end{aligned}
$$

Proof:

$$
\begin{aligned}
\|\mu_P - \mu_Q\|^2_{\mathcal{F}} &= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\
&= \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2 \langle \mu_P, \mu_Q \rangle \\
&= \mathrm{E}_P[\mu_P(x)] + \ldots
\end{aligned}
$$

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\
&= \underbrace{\mathrm{E}_P k(x, x')}_{\text{(a)}} + \underbrace{\mathrm{E}_Q k(y, y')}_{\text{(a)}} - 2\underbrace{\mathrm{E}_{P,Q} k(x, y)}_{\text{(b)}}
\end{aligned}
$$

Proof :

$$
\begin{aligned}
\|\mu_P - \mu_Q\|^2_{\mathcal{F}} &= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\
&= \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2 \langle \mu_P, \mu_Q \rangle \\
&= \mathrm{E}_P[\mu_P(x)] + \ldots \\
&= \mathrm{E}_P \langle \mu_P, k(x, \cdot) \rangle + \ldots
\end{aligned}
$$

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\
&= \underbrace{\mathrm{E}_P k(x, x')}_{\text{(a)}} + \underbrace{\mathrm{E}_Q k(y, y')}_{\text{(a)}} - 2\underbrace{\mathrm{E}_{P,Q} k(x, y)}_{\text{(b)}}
\end{aligned}
$$

Proof :

$$
\begin{aligned}
\|\mu_P - \mu_Q\|^2_{\mathcal{F}} &= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\
&= \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2 \langle \mu_P, \mu_Q \rangle \\
&= \mathrm{E}_P[\mu_P(x)] + \ldots \\
&= \mathrm{E}_P \langle \mu_P, k(x, \cdot) \rangle + \ldots \\
&= \mathrm{E}_P k(x, x') + \mathrm{E}_Q k(y, y') - 2\mathrm{E}_{P,Q} k(x, y)
\end{aligned}
$$

Illustration of MMD

- Dogs ($= P$) and fish ($= Q$) example revisited
- Each entry is one of $k(\text{dog}_i, \text{dog}_j)$, $k(\text{dog}_i, \text{fish}_j)$, or $k(\text{fish}_i, \text{fish}_j)$

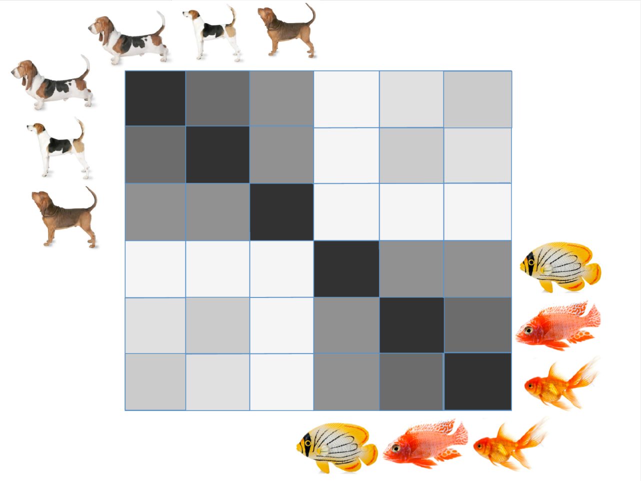
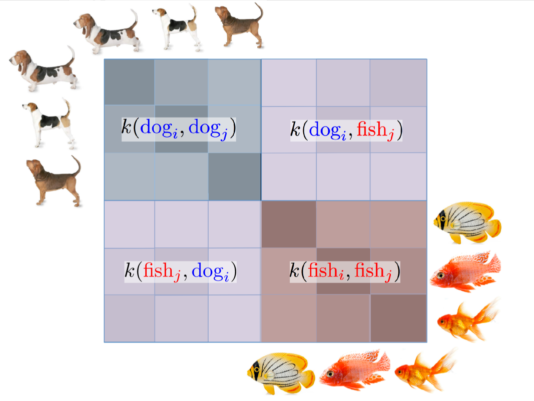


Illustration of MMD

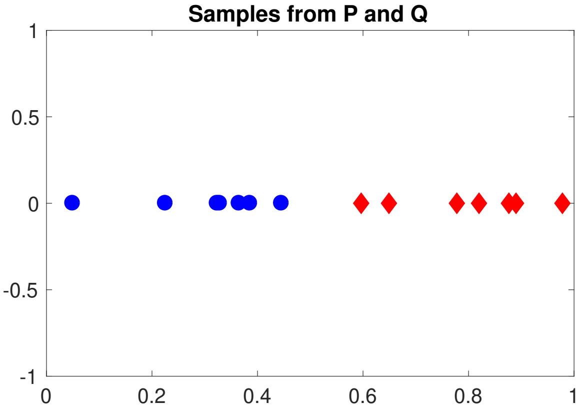
The maximum mean discrepancy:

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(\text{dog}_i, \text{dog}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\text{fish}_i, \text{fish}_j) - \frac{2}{n^2} \sum k(\text{dog}_i, \text{fish}_j)$$



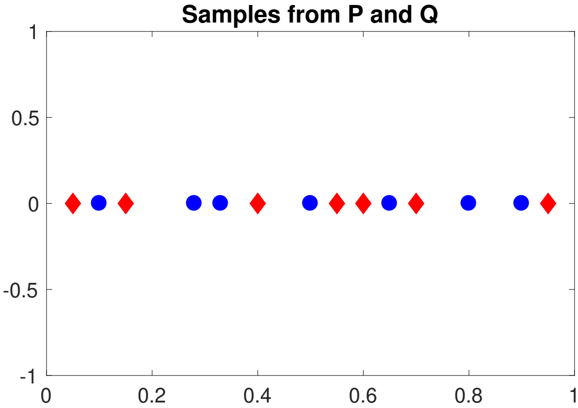
MMD as an integral probability metric

Are P and Q different?



MMD as an integral probability metric

Are P and Q different?

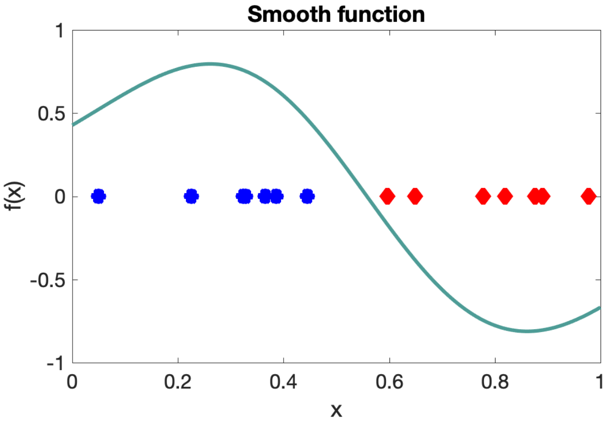


MMD as an integral probability metric

Integral probability metric:

Find a "well behaved function" $f(x)$ to maximize

$$\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y)$$

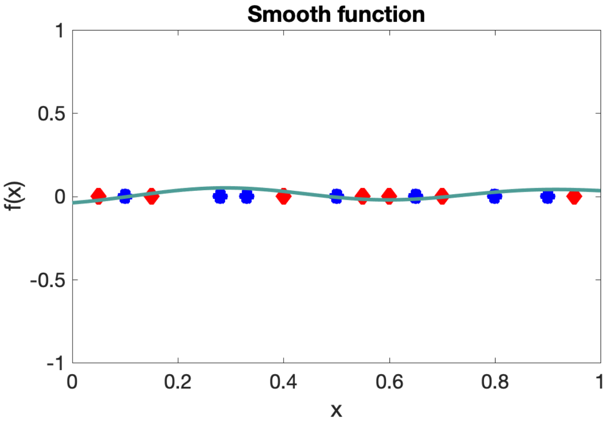


MMD as an integral probability metric

Integral probability metric:

Find a "well behaved function" $f(x)$ to maximize

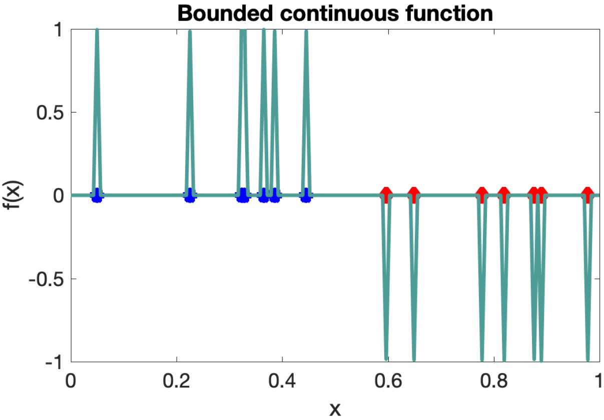
$$\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y)$$



MMD as an integral probability metric

What if the function is not smooth?

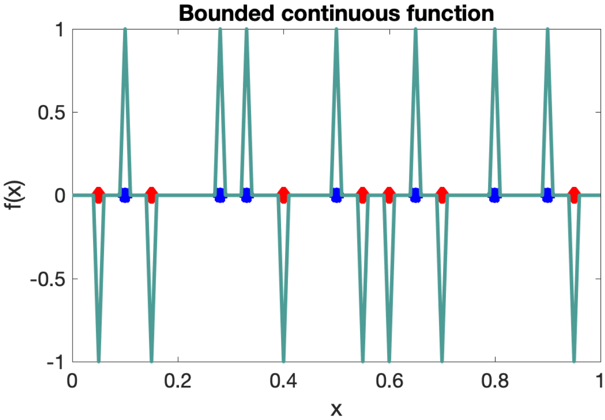
$$\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y)$$



MMD as an integral probability metric

What if the function is not smooth?

$$\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y)$$

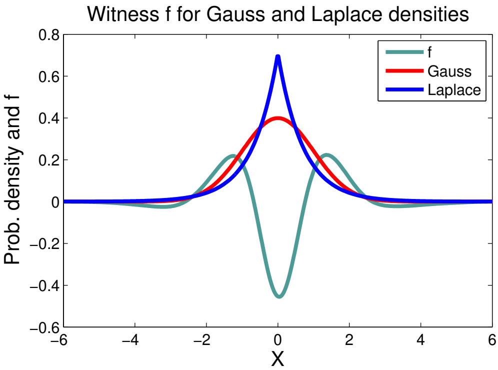


MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

$$MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right]$$

$$(F = \text{unit ball in RKHS } \mathcal{F})$$



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

$$MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right]$$

$$(F = \text{unit ball in RKHS } \mathcal{F})$$

For characteristic RKHS $\mathcal{F}$, $MMD(P, Q; F) = 0$ iff $P = Q$

Other choices for witness function class:

- Bounded continuous [Dudley, 2002]
- Bounded varation 1 (Kolmogorov metric) [Müller, 1997]
- Bounded Lipschitz (Wasserstein distances) [Dudley, 2002]

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

$$MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right]$$

$$(F = \text{unit ball in RKHS } \mathcal{F})$$

A reminder for the proof on the next slide:

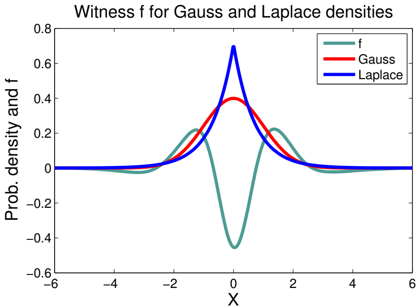
$$\mathrm{E}_P(f(X)) = \langle f, \mathrm{E}_P \varphi(X) \rangle_{\mathcal{F}} = \langle f, \mu_P \rangle_{\mathcal{F}}$$

(always true if kernel is bounded)

Integral prob. metric vs feature difference

The MMD:

$$MMD(P, Q; F) = \sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)]$$



Integral prob. metric vs feature difference

The MMD:

$$
\begin{aligned}
& MMD(P, Q; F) \\
&= \sup_{\|f\|_{\mathcal{F}} \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right] \\
&= \sup_{\|f\|_{\mathcal{F}} \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$

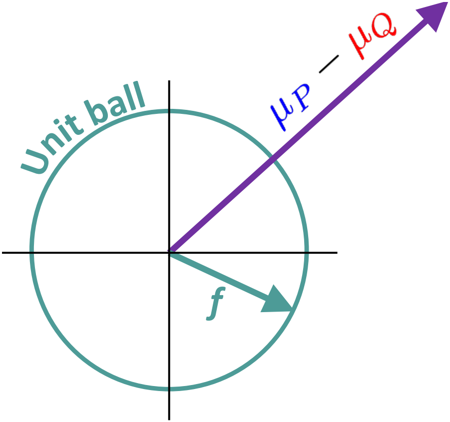
use

$$\mathrm{E}_P f(X) = \langle \mu_P, f \rangle_{\mathcal{F}}$$

Integral prob. metric vs feature difference

The MMD:

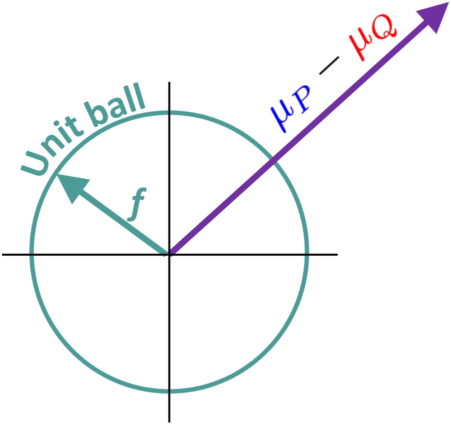
$$MMD(P, Q; F)$$
$$= \sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y)]$$
$$= \sup_{\|f\|_{\mathcal{F}} \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature difference

The MMD:

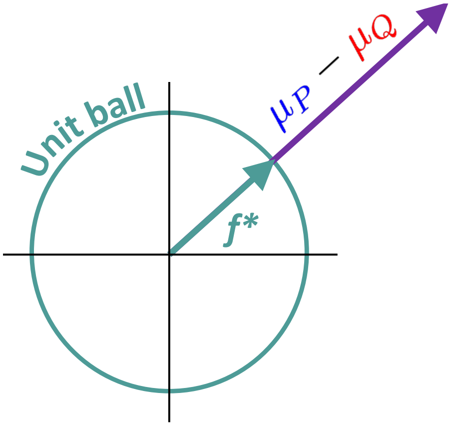
$$MMD(P, Q; F)$$
$$= \sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y)]$$
$$= \sup_{\|f\|_{\mathcal{F}} \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature difference

The MMD:

$$
\begin{aligned}
&MMD(P, Q; F) \\
&= \sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y)] \\
&= \sup_{\|f\|_{\mathcal{F}} \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



$$f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|}$$

Integral prob. metric vs feature difference

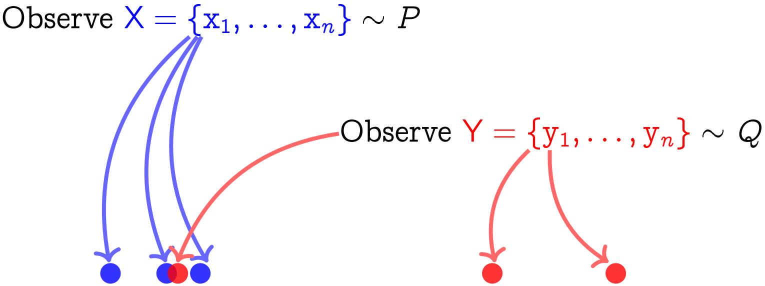
The MMD:

$$
\begin{aligned}
&MMD(P, Q; F) \\
&= \sup_{\|f\|_{\mathcal{F}} \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right] \\
&= \sup_{\|f\|_{\mathcal{F}} \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\
&= \|\mu_P - \mu_Q\|
\end{aligned}
$$

$\boxed{\text{Function view and feature view equivalent}}$

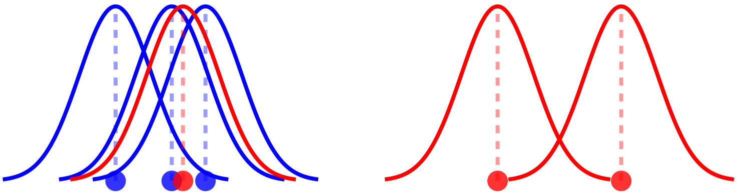
Construction of MMD witness

Construction of empirical witness function (proof: next slide!)



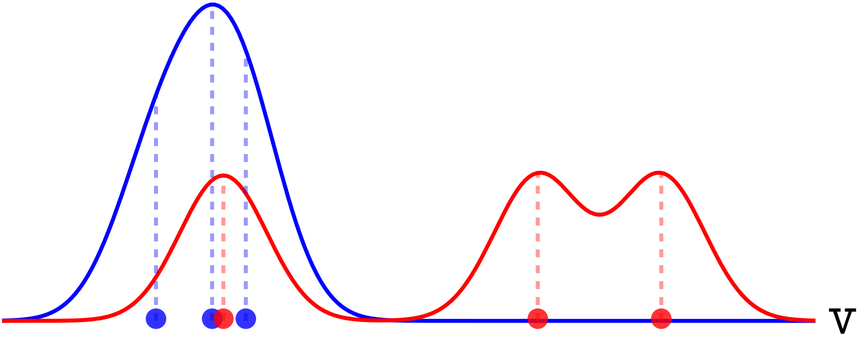
Construction of MMD witness

Construction of empirical witness function (proof: next slide!)



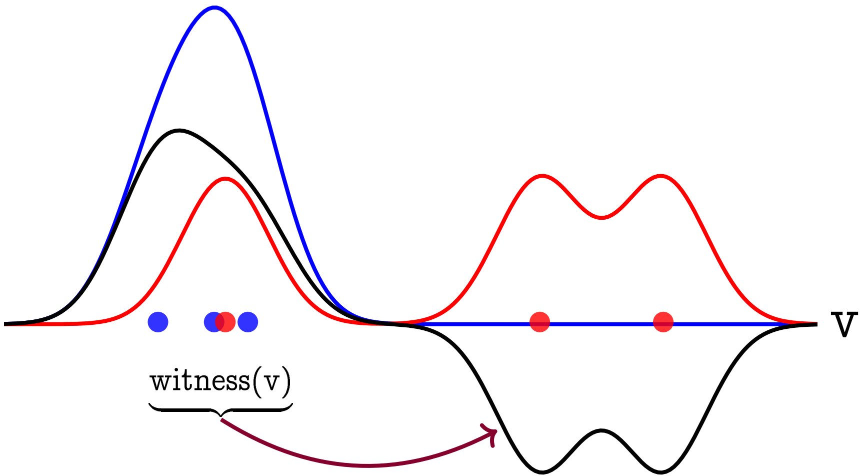
Construction of MMD witness

Construction of empirical witness function (proof: next slide!)



Construction of MMD witness

Construction of empirical witness function (proof: next slide!)



Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

$$\widehat{\mu}_P := \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i)$$

Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

$$\widehat{\mu}_P := \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i)$$

The empirical witness function at v

$$f^*(v) = \langle f^*, \varphi(v) \rangle_{\mathcal{F}}$$

Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

$$\widehat{\mu}_P := \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i)$$

The empirical witness function at v

$$\begin{aligned} f^*(v) &= \langle f^*, \varphi(v) \rangle_{\mathcal{F}} \\ &\propto \langle \widehat{\mu}_P - \widehat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \end{aligned}$$

Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

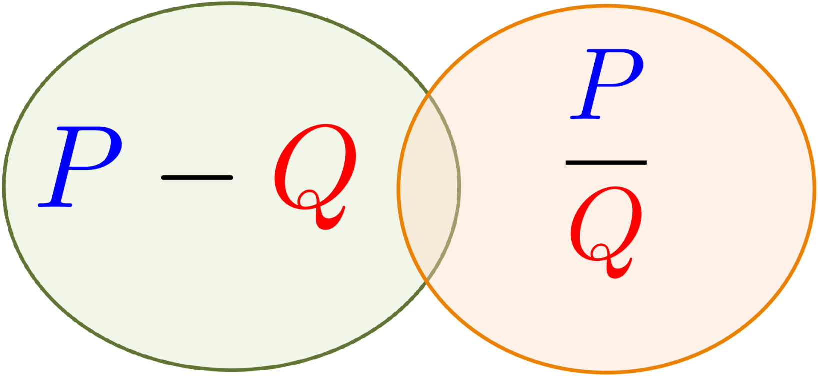
$$\widehat{\mu}_P := \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i)$$

The empirical witness function at v

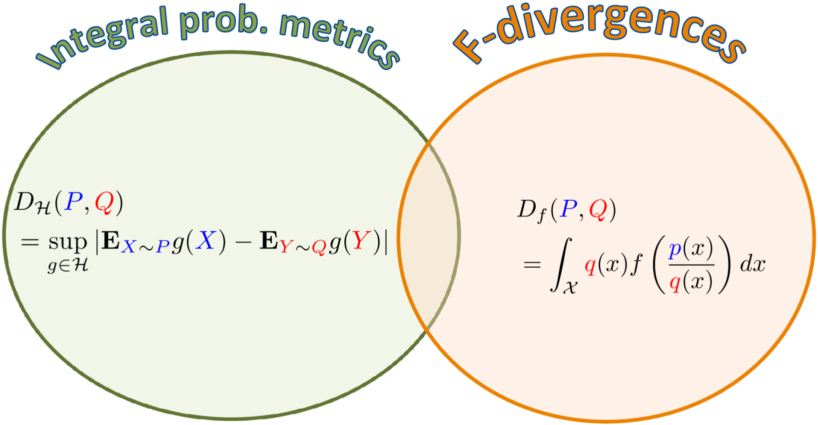
$$\begin{aligned} f^*(v) &= \langle f^*, \varphi(v) \rangle_{\mathcal{F}} \\ &\propto \langle \widehat{\mu}_P - \widehat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \\ &= \frac{1}{n}\sum_{i=1}^{n} k(x_i, v) - \frac{1}{n}\sum_{i=1}^{n} k(\mathrm{y}_i, v) \end{aligned}$$

Don't need explicit feature coefficients $f^* := \left[\begin{array}{ccc} f_1^* & f_2^* & \ldots \end{array} \right]$

Interlude: divergence measures


$P - Q$
$\frac{P}{Q}$

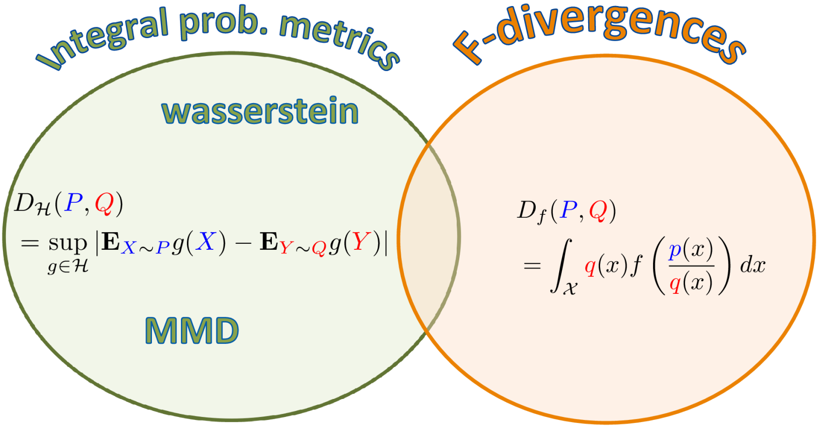
Divergences



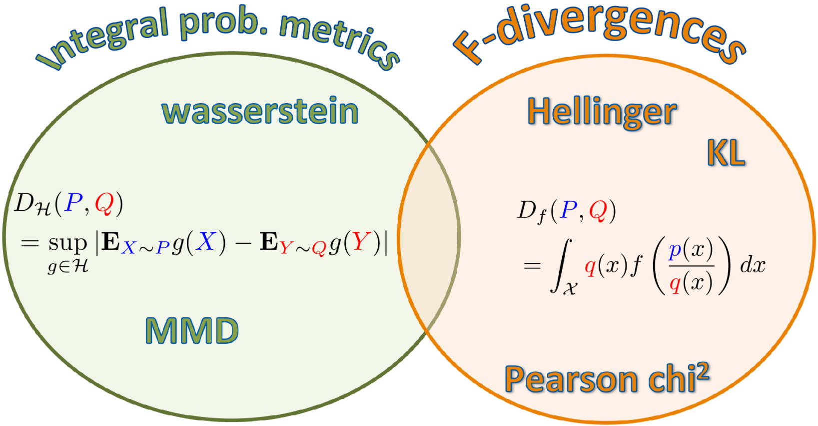
$$D_{\mathcal{H}}(P,Q) = \sup_{g\in\mathcal{H}} \left|\mathbf{E}_{X\sim P}\, g(X) - \mathbf{E}_{Y\sim Q}\, g(Y)\right|$$

$$D_f(P,Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

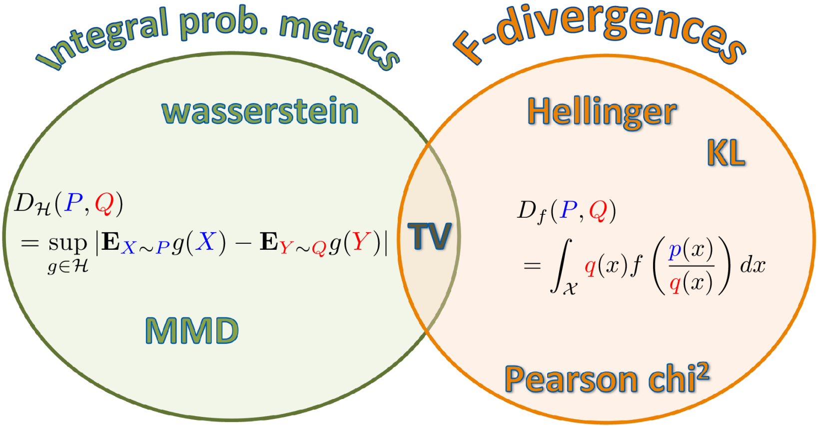
Divergences



Divergences



Divergences



Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet (EJS, 2012, Theorem A.1)

Two-Sample Testing with MMD

A statistical test using MMD

The empirical MMD:

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

How does this help decide whether $P = Q$?

A statistical test using MMD

The empirical MMD:

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

Perspective from statistical hypothesis testing:

- Null hypothesis $\mathcal{H}_0$ when $P = Q$
  - should see $\widehat{MMD}^2$ "close to zero".
- Alternative hypothesis $\mathcal{H}_1$ when $P \neq Q$
  - should see $\widehat{MMD}^2$ "far from zero"

A statistical test using MMD

The empirical MMD:

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

Perspective from statistical hypothesis testing:

- Null hypothesis $\mathcal{H}_0$ when $P = Q$
  - should see $\widehat{MMD}^2$ "close to zero".
- Alternative hypothesis $\mathcal{H}_1$ when $P \neq Q$
  - should see $\widehat{MMD}^2$ "far from zero"

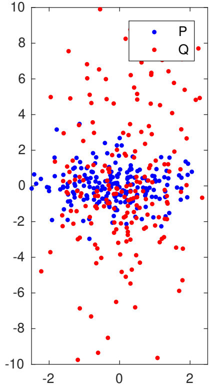
Want Threshold c_α for $\widehat{MMD}^2$ to get false positive rate α

Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Draw $n = 200$ i.i.d samples from P and Q

- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.2$

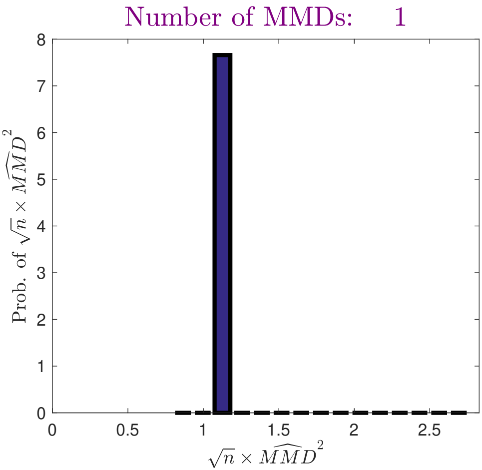
$\sqrt{n} \times \widehat{MMD}^2 = 1.2$

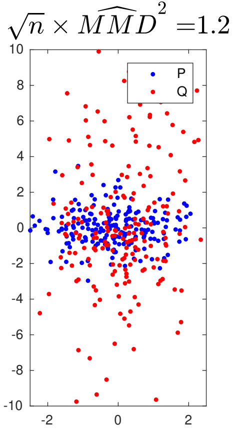


Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Draw $n = 200$ i.i.d samples from P and Q

- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.2$

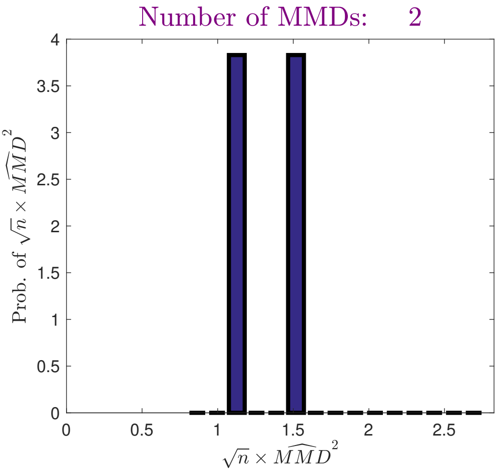


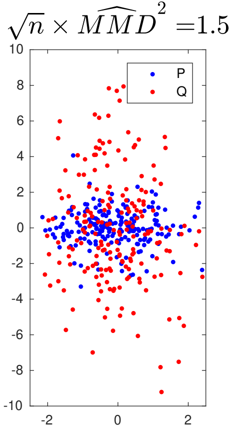


Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Draw $n = 200$ new samples from P and Q

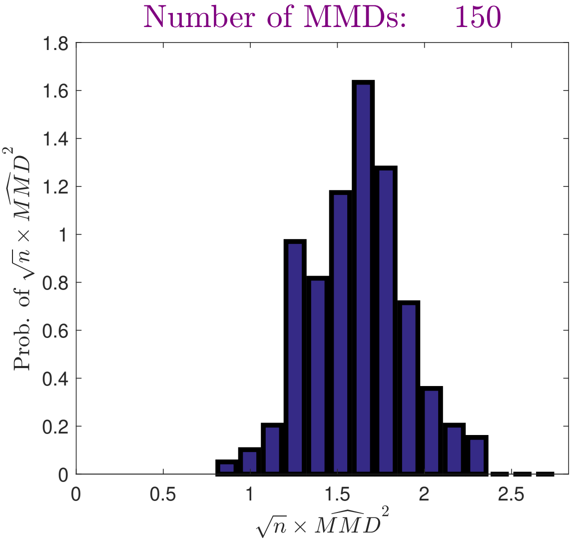
- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.5$





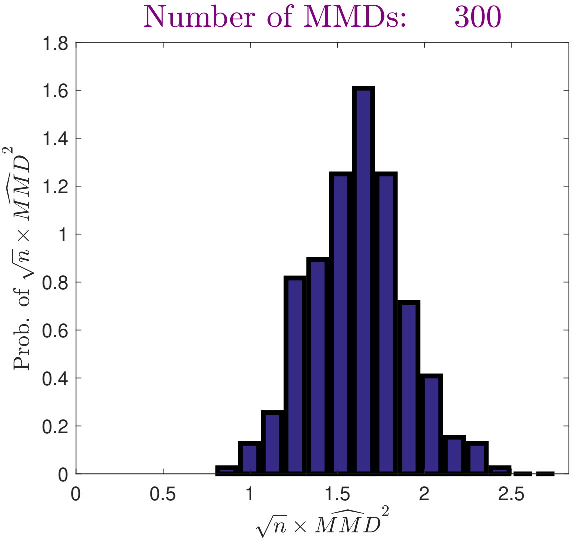
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 150 times . . .



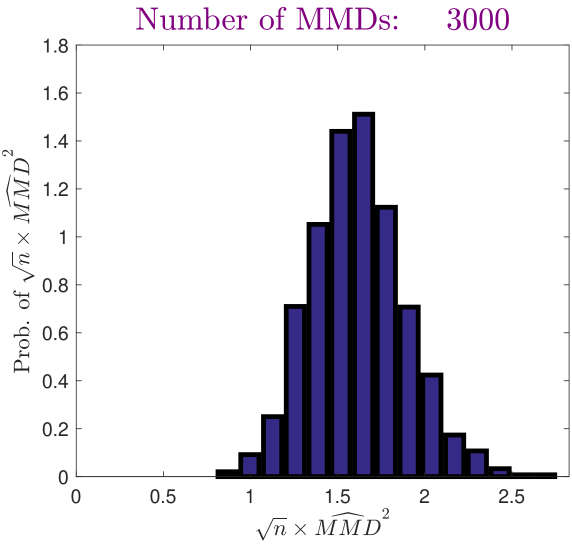
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 300 times . . .



Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 3000 times ...

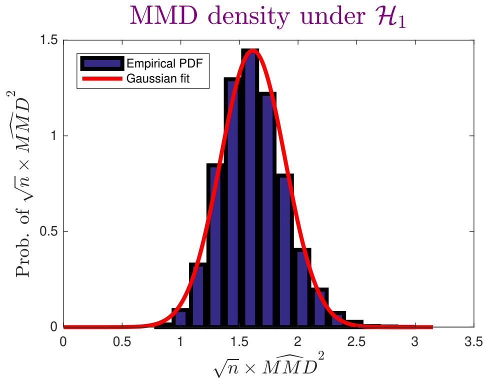


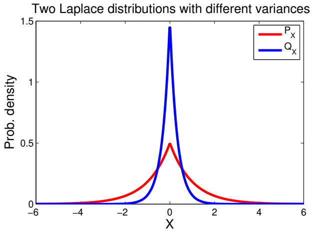
Asymptotics of $\widehat{MMD}^2$ when $P \neq Q$

When $P \neq Q$, statistic is asymptotically normal,

$$\frac{\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2(P, Q)}{\sqrt{V_n(P, Q)}} \xrightarrow{D} \mathcal{N}(0, 1),$$

where variance $V_n(P, Q) = O\left(n^{-1}\right)$.



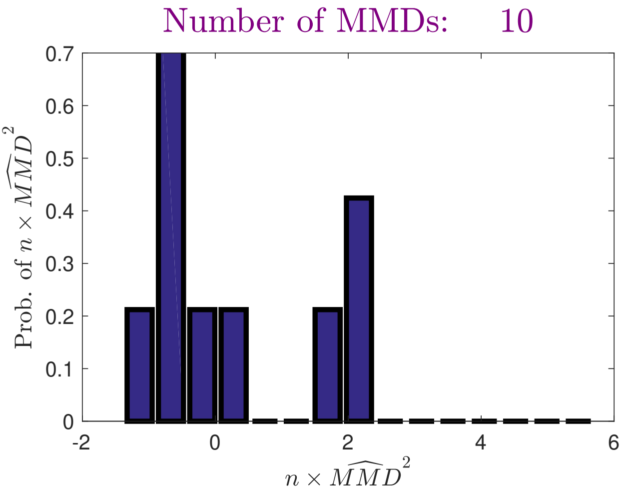


Behaviour of $\widehat{MMD}^2$ when $P = Q$

What happens when P and Q are the same?

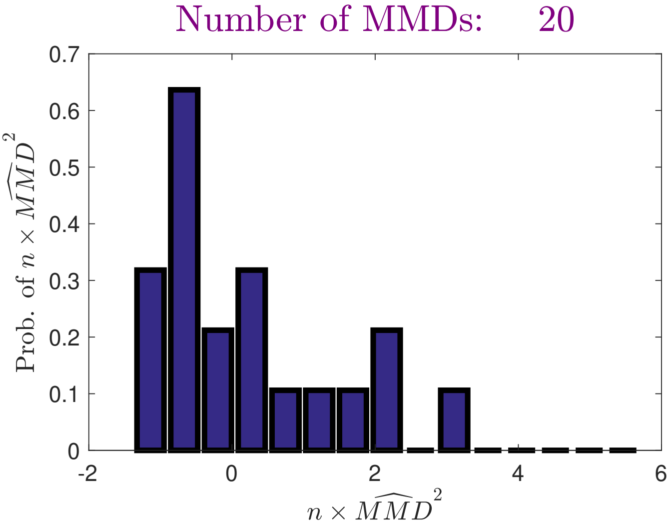
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0,1)$



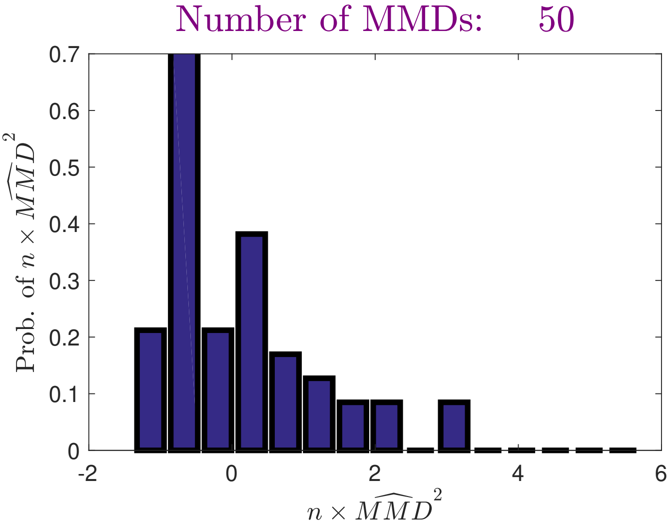
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0, 1)$



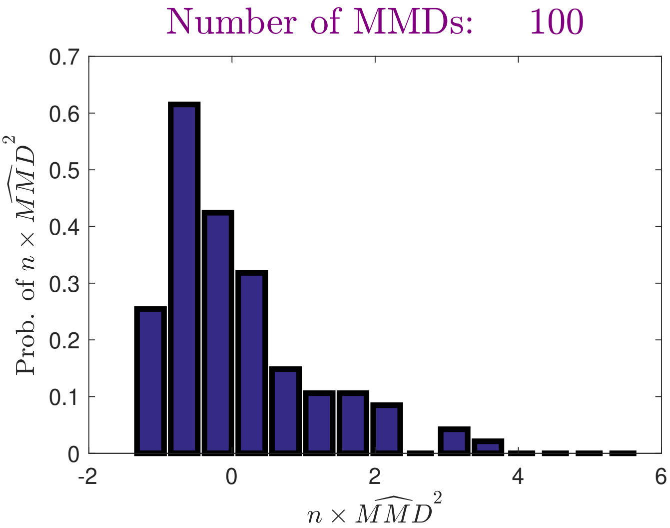
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0,1)$



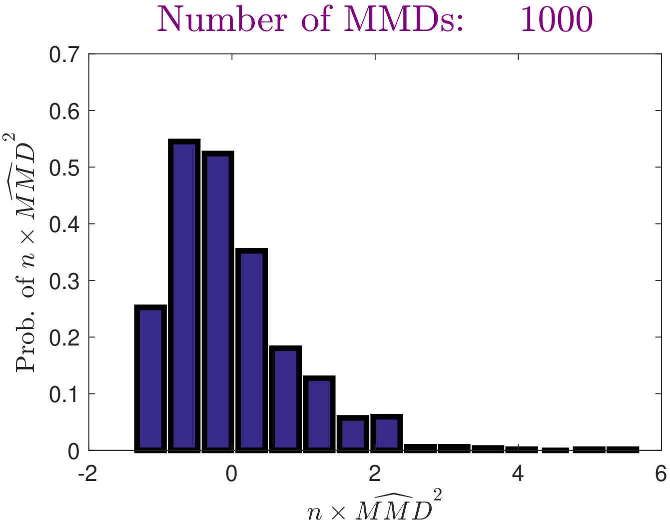
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0, 1)$



Behaviour of $\widehat{MMD}^2$ when $P = Q$

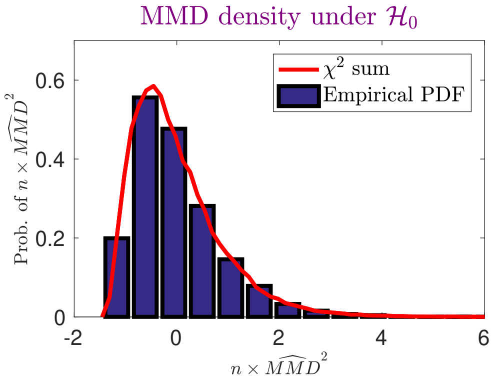
- Case of $P = Q = \mathcal{N}(0,1)$



Asymptotics of $\widehat{MMD}^2$ when $P = Q$

Where $P = Q$, statistic has asymptotic distribution

$$n\widehat{\mathrm{MMD}}^2 \sim \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 2 \right]$$



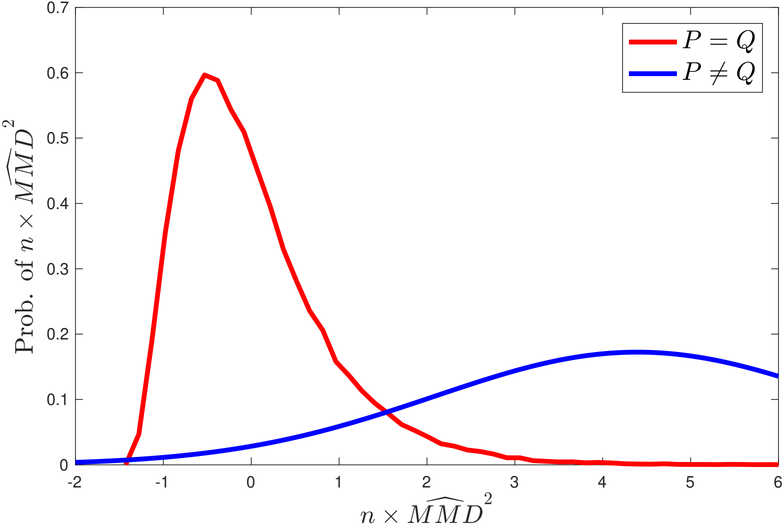
where

$$\lambda_i \psi_i(x') = \int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) dP(x)$$

$$z_l \sim \mathcal{N}(0, 2) \quad \text{i.i.d.}$$

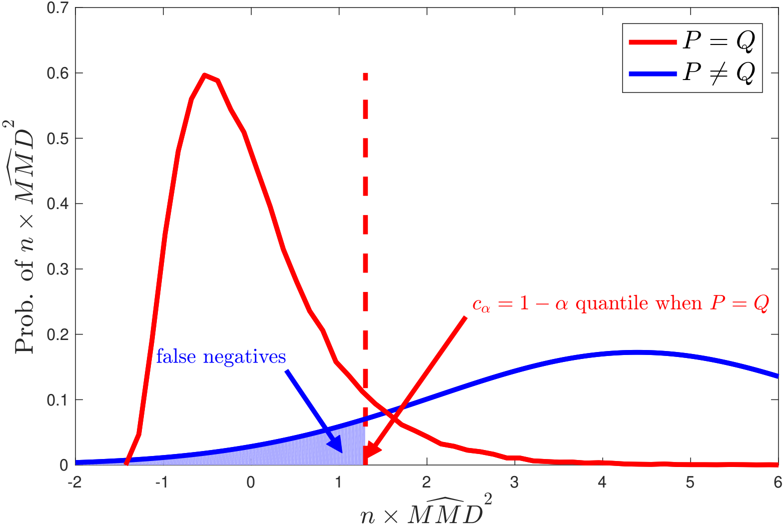
A statistical test

A summary of the asymptotics:



A statistical test

Test construction: (G., Borgwardt, Rasch, Schoelkopf, and Smola, JMLR 2012)

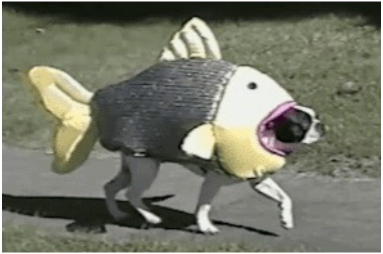


How do we get test threshold c_α?

Permuted dog and fish samples (merdogs):

$\widetilde{X} = [$  $\dots]$

$\widetilde{Y} = [$  $\dots]$



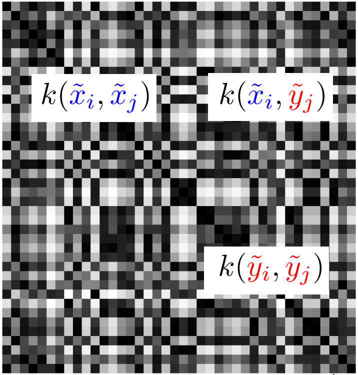
How do we get test threshold c_α?

Permuted dog and fish samples (merdogs):

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

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

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



Permutation simulates

$P = Q$

How do we get test threshold c_α?

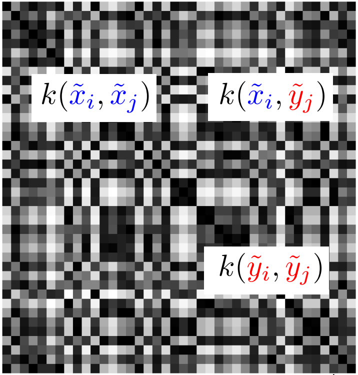
Permuted dog and fish samples (merdogs):

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

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

Exact level α (upper bound on false positive rate) at finite n and number of permutations (when unpermuted statistic included in pool)

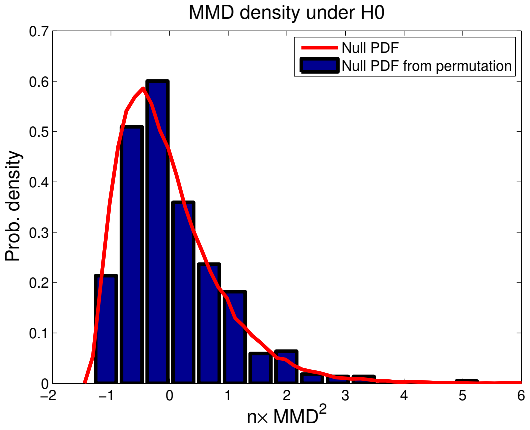
Proposition 1, Schrab, Kim, Albert, Laurent, Guedj, Gretton (2021), MMD Aggregated Two-Sample Test, arXiv:2110.15073



Approx. null distribution of $\widehat{MMD}^2$ via permutation

Null distribution estimated from 500 permutations

Example: $P = Q = \mathcal{N}(0, 1)$



Consistent test w/o bootstrap (not examinable)

Maximum mean discrepancy (MMD):

$$MMD^2(P, Q; \mathcal{F}) = \|\mu_P - \mu_Q\|^2_{\mathcal{F}}$$

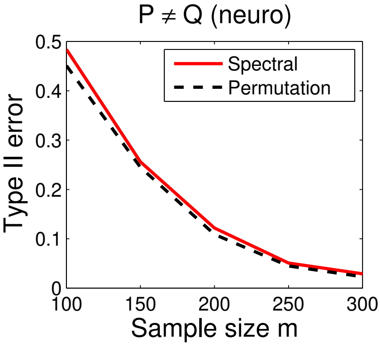
Is $\widehat{\mathrm{MMD}}^2$ significantly > 0?

$P = Q$, null distrib. of $\widehat{\mathrm{MMD}}$:

$$n\widehat{\mathrm{MMD}} \underset{D}{\rightarrow} \sum_{l=1}^{\infty} \lambda_l (z_l^2 - 2),$$

λ_l is lth eigenvalue of centered kernel $\tilde{k}(x_i, x_j)$

Use Gram matrix spectrum for $\hat{\lambda}_l$: consistent test without permutation



Gretton, Fukumizu, Harchaoui, Sriperumbudur, A fast, consistent kernel two-sample test, NeurIPS (2009)

How to choose the best kernel (1)
characteristic kernels

Characteristic kernels

Characteristic: MMD a metric $MMD = 0$ iff $P = Q$)

[NeurIPS07b, JMLR10]

In the next slides:

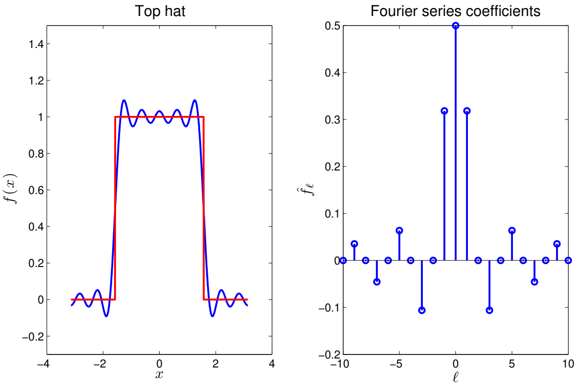
- Characteristic property on $[-\pi, \pi]$ with periodic boundary
- Characteristic property on $\mathbb{R}^d$
- Characteristic property via Universality

Characteristic kernels on $[-\pi, \pi]$

Reminder: Fourier series

Function on $[-\pi, \pi]$ with periodic boundary.

$$f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell x) = \sum_{l=-\infty}^{\infty} \hat{f}_\ell \left(\cos(\ell x) + \imath \sin(\ell x)\right).$$

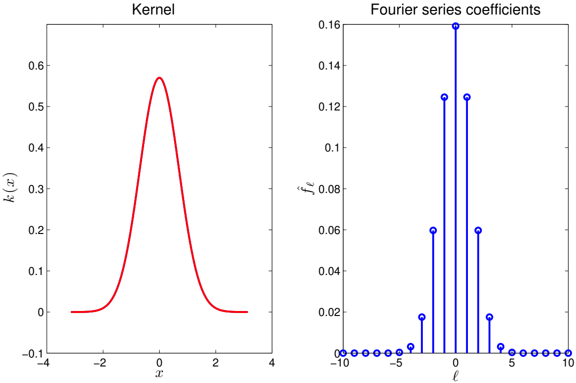


Characteristic kernels on $[-\pi, \pi]$

Jacobi theta kernel (close to exponentiated quadratic):

$$k(x-y) = \frac{1}{2\pi}\vartheta\left(\frac{x-y}{2\pi}, \frac{\imath\sigma^2}{2\pi}\right), \qquad \hat{k}_\ell = \frac{1}{2\pi}\exp\left(\frac{-\sigma^2\ell^2}{2}\right).$$

ϑ is the Jacobi theta function, close to Gaussian when σ^2 small



The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

- Fourier series for P is characteristic function $\varphi_{P,\ell}$
- Fourier series for mean embedding is product of fourier series! (convolution theorem)

$$\begin{aligned}\mu_P(x) &= \langle \mu_P, k(\cdot, x)\rangle_{\mathcal{F}} \\ &= E_{t\sim P} k(t-x) \\ &= \int_{-\pi}^{\pi} k(t-x)\, dP(t) \qquad \hat{\mu}_{P,\ell} = \hat{k}_\ell \times \bar{\varphi}_{P,\ell}\end{aligned}$$

The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

- Fourier series for P is characteristic function $\varphi_{P,\ell}$
- Fourier series for mean embedding is product of fourier series! (convolution theorem)

$$\begin{aligned}\mu_P(x) &= \langle \mu_P, k(\cdot, x)\rangle_{\mathcal{F}} \\ &= E_{t\sim P} k(t-x) \\ &= \int_{-\pi}^{\pi} k(t-x)\, dP(t) \qquad \hat{\mu}_{P,\ell} = \hat{k}_\ell \times \bar{\varphi}_{P,\ell}\end{aligned}$$

The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

- Fourier series for P is characteristic function $\varphi_{P,\ell}$
- Fourier series for mean embedding is product of fourier series! (convolution theorem)

$$\begin{aligned}\mu_P(x) &= \langle \mu_P, k(\cdot, x)\rangle_{\mathcal{F}} \\ &= E_{t\sim P} k(t-x) \\ &= \int_{-\pi}^{\pi} k(t-x)\, dP(t) \qquad \hat{\mu}_{P,\ell} = \hat{k}_\ell \times \bar{\varphi}_{P,\ell}\end{aligned}$$

The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

- Fourier series for P is characteristic function $\varphi_{P,\ell}$
- Fourier series for mean embedding is product of fourier series! (convolution theorem)

$$\begin{aligned}
\mu_P(x) &= \langle \mu_P, k(\cdot, x)\rangle_{\mathcal{F}} \\
&= E_{t\sim P} k(t-x) \\
&= \int_{-\pi}^{\pi} k(t-x)\, dP(t) \qquad \hat{\mu}_{P,\ell} = \hat{k}_\ell \times \bar{\varphi}_{P,\ell}
\end{aligned}$$

The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

- Fourier series for P is characteristic function $\varphi_{P,\ell}$
- Fourier series for mean embedding is product of fourier series! (convolution theorem)

$$\mu_P(x) = \langle \mu_P, k(\cdot, x) \rangle_{\mathcal{F}}$$
$$= E_{t \sim P} k(t - x)$$
$$= \int_{-\pi}^{\pi} k(t - x) \, dP(t) \qquad \hat{\mu}_{P,\ell} = \hat{k}_\ell \times \bar{\varphi}_{P,\ell}$$

MMD can be written in terms of Fourier series:

$$MMD(P, Q; \mathcal{F}) = \|\mu_P - \mu_Q\|_{\mathcal{F}}$$
$$= \left\| \sum_{\ell=-\infty}^{\infty} \left[\left(\bar{\varphi}_{P,\ell} - \bar{\varphi}_{Q,\ell} \right) \hat{k}_\ell \right] \exp(\imath \ell x) \right\|_{\mathcal{F}}$$

A simpler Fourier representation for MMD

From previous slide,

$$MMD(P, Q; \mathcal{F}) = \left\| \sum_{\ell=-\infty}^{\infty} \left[\left(\bar{\varphi}_{P,\ell} - \bar{\varphi}_{Q,\ell} \right) \hat{k}_\ell \right] \exp(\imath \ell x) \right\|_{\mathcal{F}}$$

Reminder: the squared norm of a function f in $\mathcal{F}$ is:

$$\|f\|_{\mathcal{F}}^2 = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_\ell|^2}{\hat{k}_\ell}.$$

Simple, interpretable expression for squared MMD:

$$MMD^2(P, Q; \mathcal{F}) = \sum_{\ell=-\infty}^{\infty} \frac{|\varphi_{P,\ell} - \varphi_{Q,\ell}|^2 \hat{k}_\ell^2}{\hat{k}_\ell} = \sum_{\ell=-\infty}^{\infty} |\varphi_{P,\ell} - \varphi_{Q,\ell}|^2 \hat{k}_\ell$$

A simpler Fourier representation for MMD

From previous slide,

$$MMD(P, Q; \mathcal{F}) = \left\| \sum_{\ell=-\infty}^{\infty} \left[\left(\bar{\varphi}_{P,\ell} - \bar{\varphi}_{Q,\ell} \right) \hat{k}_\ell \right] \exp(\imath \ell x) \right\|_{\mathcal{F}}$$

Reminder: the squared norm of a function f in $\mathcal{F}$ is:

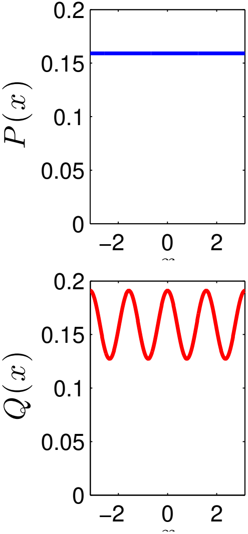
$$\|f\|_{\mathcal{F}}^2 = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_\ell|^2}{\hat{k}_\ell}.$$

Simple, interpretable expression for squared MMD:

$$MMD^2(P, Q; \mathcal{F}) = \sum_{\ell=-\infty}^{\infty} \frac{|\varphi_{P,\ell} - \varphi_{Q,\ell}|^2 \hat{k}_\ell^2}{\hat{k}_\ell} = \sum_{\ell=-\infty}^{\infty} |\varphi_{P,\ell} - \varphi_{Q,\ell}|^2 \hat{k}_\ell$$

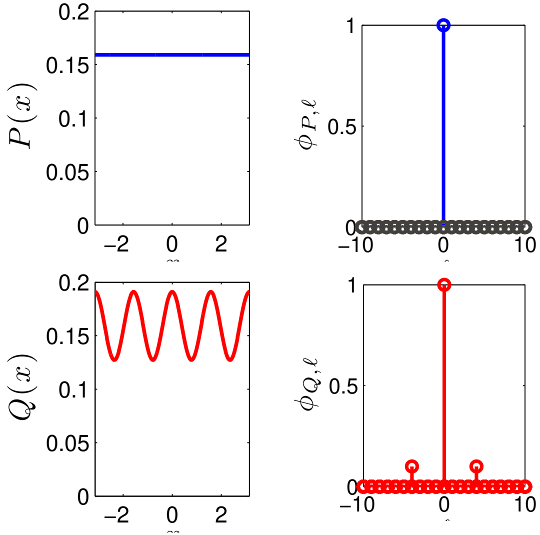
Characteristic kernels on $[-\pi, \pi]$

Example: P differs from Q at one frequency:



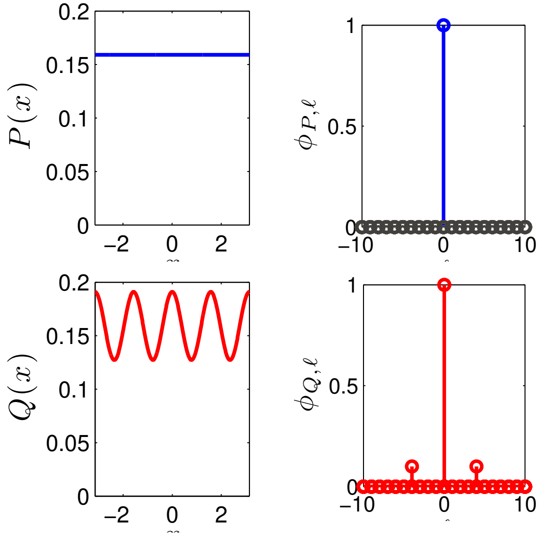
Characteristic kernels on $[-\pi, \pi]$

Example: P differs from Q at one frequency:

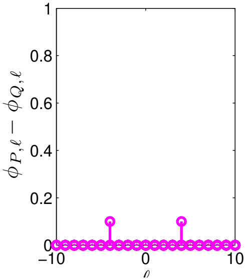


Characteristic kernels on $[-\pi, \pi]$

Example: P differs from Q at one frequency:

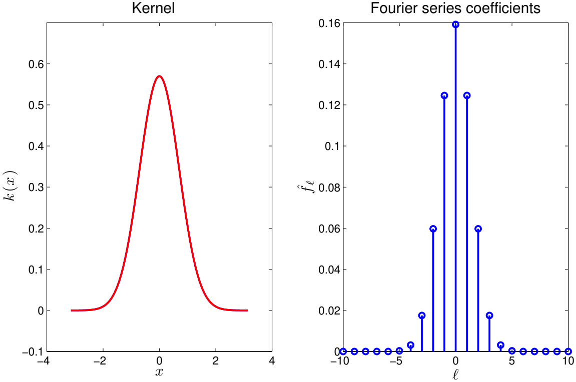


Characteristic function difference



Characteristic kernels on $[-\pi, \pi]$

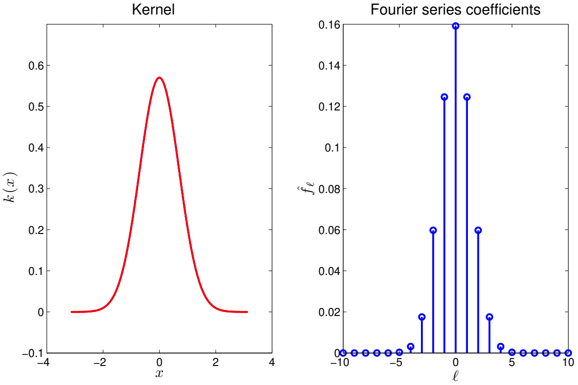
Is the Gaussian spectrum kernel characteristic?



$$MMD^2(P, Q; F) = \sum_{\ell=-\infty}^{\infty} \left|\varphi_{P,\ell} - \varphi_{Q,\ell}\right|^2 \hat{k}_\ell$$

Characteristic kernels on $[-\pi, \pi]$

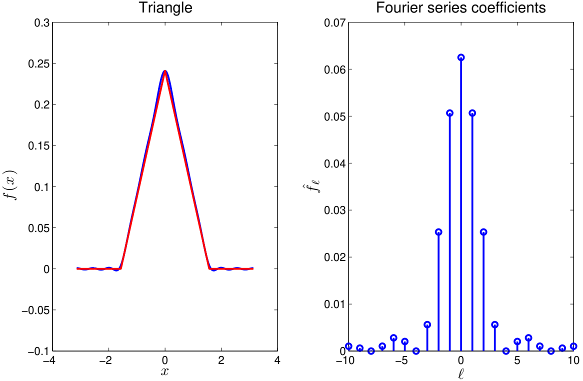
Is the Gaussian spectrum kernel characteristic? YES



$$MMD^2(P, Q; F) = \sum_{\ell=-\infty}^{\infty} \left|\varphi_{P,\ell} - \varphi_{Q,\ell}\right|^2 \hat{k}_\ell$$

Characteristic kernels on $[-\pi, \pi]$

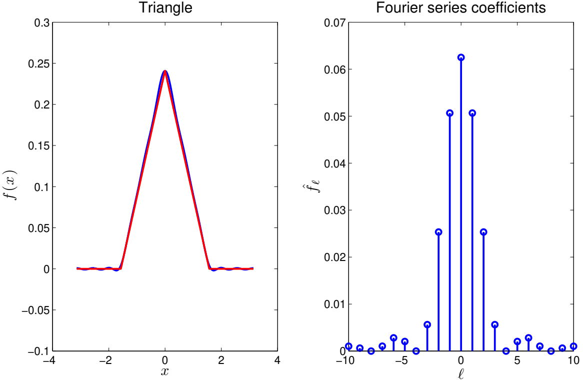
Is the triangle kernel characteristic?



$$MMD^2(P, Q; F) = \sum_{l=-\infty}^{\infty} \left|\varphi_{P,\ell} - \varphi_{Q,\ell}\right|^2 \hat{k}_\ell$$

Characteristic kernels on $[-\pi, \pi]$

Is the triangle kernel characteristic? NO



$$MMD^2(P, Q; F) = \sum_{l=-\infty}^{\infty} \left|\varphi_{P,\ell} - \varphi_{Q,\ell}\right|^2 \hat{k}_\ell$$

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

Can we prove characteristic on $\mathbb{R}^d$?

Characteristic function of P via Fourier transform

$$\varphi_P(\omega) = \int_{\mathbb{R}^d} e^{ix^\top \omega} dP(x)$$

For translation invariant kernels: $k(x, y) = k(x - y)$, Bochner's theorem:

$$k(x - y) = \int_{\mathbb{R}^d} e^{-i(x-y)^\top \omega} d\Lambda(\omega)$$

$\Lambda(\omega)$ finite non-negative Borel measure.

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

Can we prove characteristic on $\mathbb{R}^d$?

Characteristic function of P via Fourier transform

$$\varphi_P(\omega) = \int_{\mathbb{R}^d} e^{ix^\top \omega} dP(x)$$

For translation invariant kernels: $k(x, y) = k(x - y)$, Bochner's theorem:

$$k(x - y) = \int_{\mathbb{R}^d} e^{-i(x-y)^\top \omega} d\Lambda(\omega)$$

$\Lambda(\omega)$ finite non-negative Borel measure.

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

Fourier representation of MMD on $\mathbb{R}^d$:

$$MMD^2(P, Q; F) = \int |\varphi_P(\omega) - \varphi_Q(\omega)|^2 \, d\Lambda(\omega)$$

Proof:

$$\begin{aligned}
& MMD^2(P, Q; F) \\
&:= E_P k(x - x') + E_Q k(y - y') - 2E_{P,Q} k(x - y) \\
&= \int \int \left[k(s - t) \, d(P - Q)(s) \right] d(P - Q)(t) \\
&\overset{(a)}{=} \int \int \int_{\mathbb{R}^d} e^{-i(s-t)^T \omega} \, d\Lambda(\omega) \, d(P - Q)(s) \, d(P - Q)(t) \\
&\overset{(b)}{=} \int \int_{\mathbb{R}^d} e^{-is^T \omega} \, d(P - Q)(s) \int_{\mathbb{R}^d} e^{it^T \omega} \, d(P - Q)(t) \, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\phi_P(\omega) - \phi_Q(\omega)|^2 \, d\Lambda(\omega)
\end{aligned}$$

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

Fourier representation of MMD on $\mathbb{R}^d$:

$$MMD^2(P, Q; F) = \int |\varphi_P(\omega) - \varphi_Q(\omega)|^2 \, d\Lambda(\omega)$$

Proof:

$$\begin{aligned}
& MMD^2(P, Q; F) \\
& := E_P k(x - x') + E_Q k(y - y') - 2E_{P,Q} k(x - y) \\
& = \int\int \left[k(s - t) \, d(P - Q)(s) \right] d(P - Q)(t) \\
& \stackrel{(a)}{=} \int\int\int_{\mathbb{R}^d} e^{-i(s-t)^T\omega} \, d\Lambda(\omega) \, d(P - Q)(s) \, d(P - Q)(t) \\
& \stackrel{(b)}{=} \int\int_{\mathbb{R}^d} e^{-is^T\omega} \, d(P - Q)(s) \int_{\mathbb{R}^d} e^{it^T\omega} \, d(P - Q)(t) \, d\Lambda(\omega) \\
& = \int_{\mathbb{R}^d} |\phi_P(\omega) - \phi_Q(\omega)|^2 \, d\Lambda(\omega)
\end{aligned}$$

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

Fourier representation of MMD on $\mathbb{R}^d$:

$$MMD^2(P, Q; F) = \int |\varphi_P(\omega) - \varphi_Q(\omega)|^2 \, d\Lambda(\omega)$$

Proof:

$$\begin{aligned}
& MMD^2(P, Q; F) \\
&:= E_P k(x - x') + E_Q k(y - y') - 2E_{P,Q} k(x - y) \\
&= \int \int \left[k(s - t) \, d(P - Q)(s) \right] d(P - Q)(t) \\
&\overset{(a)}{=} \int \int \int_{\mathbb{R}^d} e^{-i(s-t)^T \omega} \, d\Lambda(\omega) \, d(P - Q)(s) \, d(P - Q)(t) \\
&\overset{(b)}{=} \int \int_{\mathbb{R}^d} e^{-is^T \omega} \, d(P - Q)(s) \int_{\mathbb{R}^d} e^{it^T \omega} \, d(P - Q)(t) \, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\phi_P(\omega) - \phi_Q(\omega)|^2 \, d\Lambda(\omega)
\end{aligned}$$

(a) Using Bochner's theorem...

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

Fourier representation of MMD on $\mathbb{R}^d$:

$$MMD^2(P, Q; F) = \int |\varphi_P(\omega) - \varphi_Q(\omega)|^2 \, d\Lambda(\omega)$$

Proof:

$$\begin{aligned}
& MMD^2(P, Q; F) \\
& := E_P k(x - x') + E_Q k(y - y') - 2E_{P,Q} k(x - y) \\
& = \int \int \left[k(s - t) \, d(P - Q)(s) \right] d(P - Q)(t) \\
& \overset{(a)}{=} \int \int \int_{\mathbb{R}^d} e^{-i(s-t)^T \omega} \, d\Lambda(\omega) \, d(P - Q)(s) \, d(P - Q)(t) \\
& \overset{(b)}{=} \int \int_{\mathbb{R}^d} e^{-is^T \omega} \, d(P - Q)(s) \int_{\mathbb{R}^d} e^{it^T \omega} \, d(P - Q)(t) \, d\Lambda(\omega) \\
& = \int_{\mathbb{R}^d} |\phi_P(\omega) - \phi_Q(\omega)|^2 \, d\Lambda(\omega)
\end{aligned}$$

(a) Using Bochner's theorem......(b) and using Fubini's theorem.

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

Fourier representation of MMD on $\mathbb{R}^d$:

$$MMD^2(P, Q; F) = \int \left|\varphi_P(\omega) - \varphi_Q(\omega)\right|^2 d\Lambda(\omega)$$

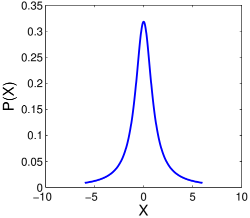
Proof:

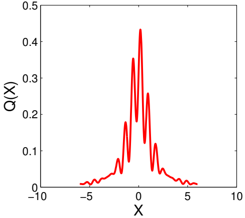
$$\begin{aligned}
& MMD^2(P, Q; F) \\
& := E_P k(x - x') + E_Q k(y - y') - 2E_{P,Q} k(x - y) \\
& = \int \int \left[k(s - t)\, d(P - Q)(s) \right] d(P - Q)(t) \\
& \stackrel{(a)}{=} \int \int \int_{\mathbb{R}^d} e^{-i(s-t)^T \omega}\, d\Lambda(\omega)\, d(P - Q)(s)\, d(P - Q)(t) \\
& \stackrel{(b)}{=} \int \int_{\mathbb{R}^d} e^{-is^T \omega}\, d(P - Q)(s) \int_{\mathbb{R}^d} e^{it^T \omega}\, d(P - Q)(t)\, d\Lambda(\omega) \\
& = \int_{\mathbb{R}^d} \left|\phi_P(\omega) - \phi_Q(\omega)\right|^2\, d\Lambda(\omega)
\end{aligned}$$

(a) Using Bochner's theorem......(b) and using Fubini's theorem.

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

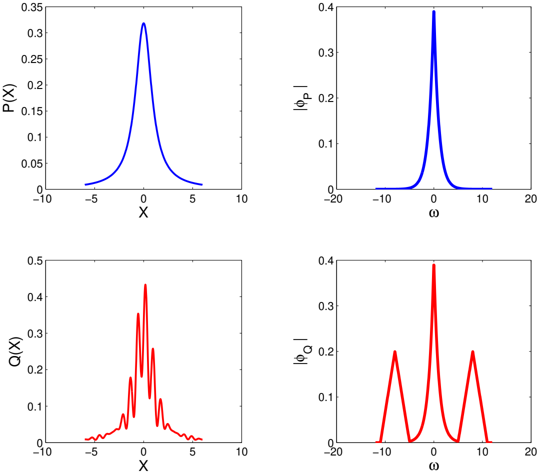
Example: P differs from Q at roughly one frequency:





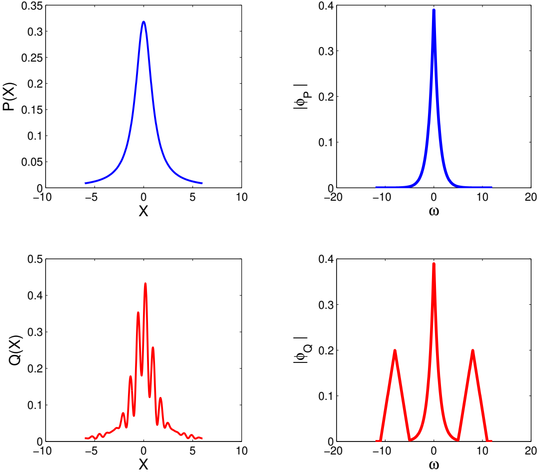
Characteristic kernels on $\mathbb{R}^d$

Example: P differs from Q at roughly one frequency:

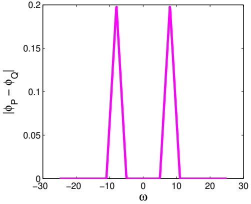


Characteristic kernels on $\mathbb{R}^d$

Example: P differs from Q at roughly one frequency:



Characteristic function difference

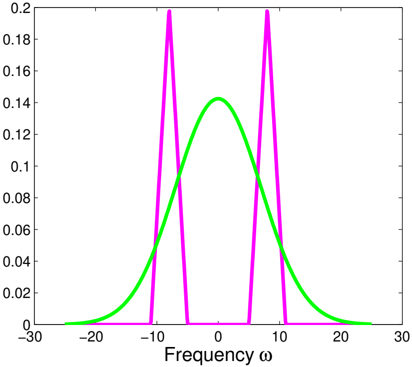


Characteristic kernels on $\mathbb{R}^d$

Example: P differs from Q at (roughly) one frequency:

Exponentiated quadraric kernel spectrum $\Lambda(\omega)$

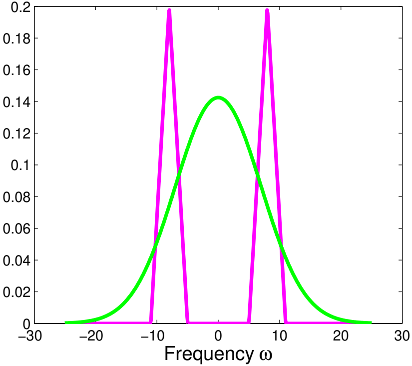
Difference $|\varphi_P - \varphi_Q|$



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

Example: P differs from Q at (roughly) one frequency:

Characteristic

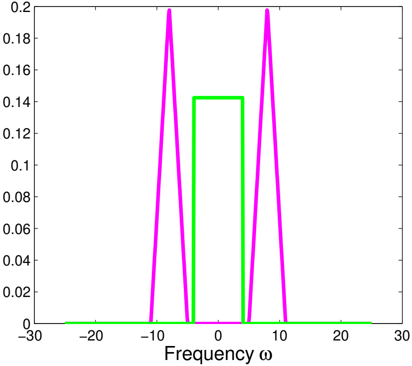


Characteristic kernels on $\mathbb{R}^d$

Example: P differs from Q at (roughly) one frequency:

Sinc kernel spectrum $\Lambda(\omega)$

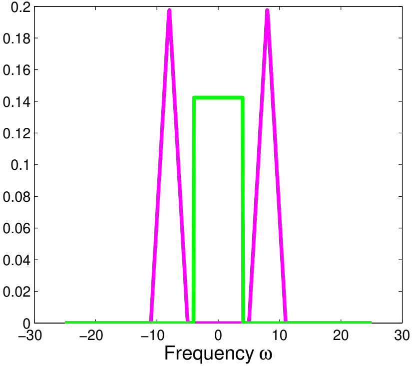
Difference $|\varphi_P - \varphi_Q|$



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

Example: P differs from Q at (roughly) one frequency:

Not characteristic

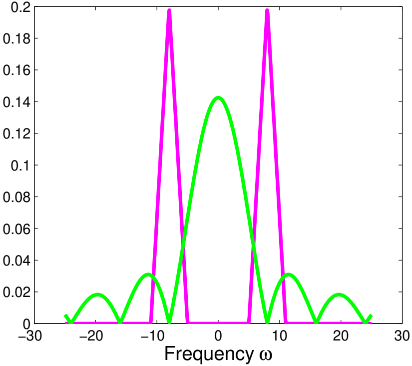


Characteristic kernels on $\mathbb{R}^d$

Example: P differs from Q at (roughly) one frequency:

Triangle (B-spline) kernel spectrum $\Lambda(\omega)$

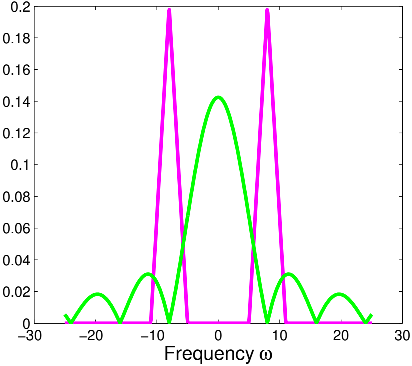
Difference $|\phi_P - \phi_Q|$



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

Example: P differs from Q at (roughly) one frequency:

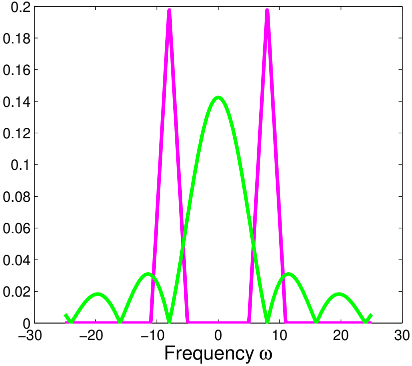
???



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

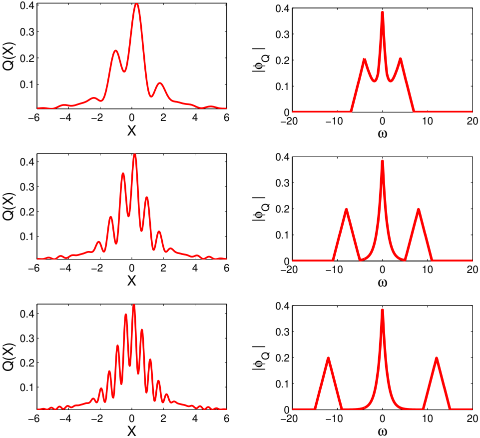
Example: P differs from Q at (roughly) one frequency:

Characteristic

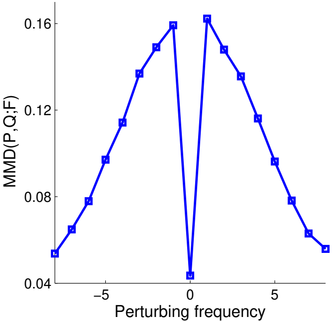


Choosing the best kernel (Fourier)

Exponentiated quadratic kernel:

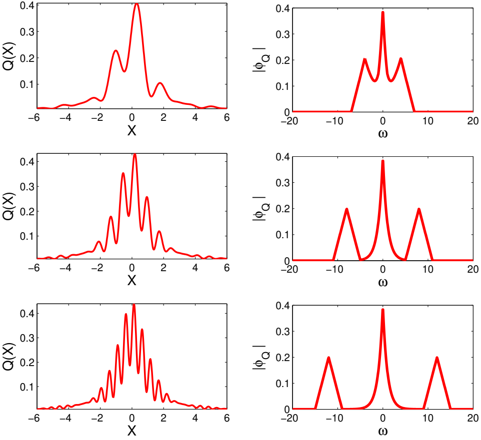


MMD vs frequency of perturbation to P

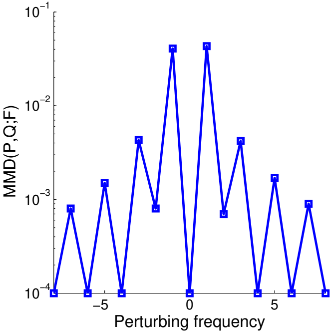


Choosing the best kernel (Fourier)

B-Spline kernel:



MMD vs frequency of perturbation to P



MMD decay with increasing perturbation freq.

Recall simple MMD, Fourier series on $[-\pi, \pi]$:

$$MMD^2(P, Q; \mathcal{F}) = \sum_{\ell=-\infty}^{\infty} \left|\varphi_{P,\ell} - \varphi_{Q,\ell}\right|^2 \hat{k}_\ell$$

where $\hat{k}_\ell$ decays as ℓ grows.

Fourier series representation for more general case on $\mathbb{R}^d$:

$$MMD^2(P, Q; \mathcal{F}) = \int_{\mathbb{R}^d} \left|\phi_P(\omega) - \phi_Q(\omega)\right|^2 \, d\Lambda(\omega)$$

has similar behaviour.

Summary: characteristic kernels on $\mathbb{R}^d$

Characteristic kernel: $MMD = 0$ iff $P = Q$ Fukumizu et al. [NIPS07b], Sriperumbudur et al.[COLT08]

Main theorem: A translation invariant k is characteristic for prob. measures on $\mathbb{R}^d$ if and only if

$$\operatorname{supp}(\Lambda) = \mathbb{R}^d$$

(i.e. support zero on at most a countable set) Sriperumbudur et al. [COLT08, JMLR10]

Corollary: any continuous, compactly supported k characteristic (since Fourier spectrum $\Lambda(\omega)$ cannot be zero on an interval).

1-D proof sketch from [Mallat, 99, Theorem 2.6], proof on $\mathbb{R}^d$ via distribution theory in Sriperumbudur et al. [JMLR10, Corollary 10 p. 1535]

Summary: characteristic kernels on $\mathbb{R}^d$

Characteristic kernel: $MMD = 0$ iff $P = Q$ Fukumizu et al. [NIPS07b], Sriperumbudur et al.[COLT08]

Main theorem: A translation invariant k is characteristic for prob. measures on $\mathbb{R}^d$ if and only if

$$\mathrm{supp}(\Lambda) = \mathbb{R}^d$$

(i.e. support zero on at most a countable set) Sriperumbudur et al. [COLT08, JMLR10]

Corollary: any continuous, compactly supported k characteristic (since Fourier spectrum $\Lambda(\omega)$ cannot be zero on an interval).

1-D proof sketch from [Mallat, 99, Theorem 2.6], proof on $\mathbb{R}^d$ via distribution theory in Sriperumbudur et al. [JMLR10, Corollary 10 p. 1535]

Summary: characteristic kernels on $\mathbb{R}^d$

Characteristic kernel: $MMD = 0$ iff $P = Q$ Fukumizu et al. [NIPS07b], Sriperumbudur et al.[COLT08]

Main theorem: A translation invariant k is characteristic for prob. measures on $\mathbb{R}^d$ if and only if

$$\mathrm{supp}(\Lambda) = \mathbb{R}^d$$

(i.e. support zero on at most a countable set) Sriperumbudur et al. [COLT08, JMLR10]

Corollary: any continuous, compactly supported k characteristic (since Fourier spectrum $\Lambda(\omega)$ cannot be zero on an interval).

1-D proof sketch from [Mallat, 99, Theorem 2.6], proof on $\mathbb{R}^d$ via distribution theory in Sriperumbudur et al. [JMLR10, Corollary 10 p. 1535]

Characteristic kernels (via Universality)

Characteristic kernels: $MMD = 0$ iff $P = Q$

Classical result:
$P = Q$ if and only if $E_P(f(x)) = E_Q(f(y))$ for all $f \in C(\mathcal{X})$, the space of bounded continuous functions on $\mathcal{X}$ Dudley (2002)

Universal RKHS:
$k(x, x')$ continuous, $\mathcal{X}$ compact, and $\mathcal{F}$ dense in $C(\mathcal{X})$ with respect to L_∞ Steinwart (2001)

If $\mathcal{F}$ universal, then $MMD(P, Q; \mathcal{F}) = 0$ iff $P = Q$

Characteristic kernels (via Universality)

Characteristic kernels: $MMD = 0$ iff $P = Q$

Classical result:
$P = Q$ if and only if $E_P(f(x)) = E_Q(f(y))$ for all $f \in C(\mathcal{X})$, the space of bounded continuous functions on $\mathcal{X}$ Dudley (2002)

Universal RKHS:
$k(x, x')$ continuous, $\mathcal{X}$ compact, and $\mathcal{F}$ dense in $C(\mathcal{X})$ with respect to L_∞ Steinwart (2001)

If $\mathcal{F}$ universal, then $MMD(P, Q; \mathcal{F}) = 0$ iff $P = Q$

Characteristic kernels (via Universality)

Characteristic kernels: $MMD = 0$ iff $P = Q$

Classical result:
$P = Q$ if and only if $E_P(f(x)) = E_Q(f(y))$ for all $f \in C(\mathcal{X})$, the space of bounded continuous functions on $\mathcal{X}$ Dudley (2002)

Universal RKHS:
$k(x, x')$ continuous, $\mathcal{X}$ compact, and $\mathcal{F}$ dense in $C(\mathcal{X})$ with respect to L_∞ Steinwart (2001)

If $\mathcal{F}$ universal, then $MMD(P, Q; \mathcal{F}) = 0$ iff $P = Q$

Characteristic kernels (via Universality)

Characteristic kernels: $MMD = 0$ iff $P = Q$

Classical result:
$P = Q$ if and only if $E_P(f(x)) = E_Q(f(y))$ for all $f \in C(\mathcal{X})$, the space of bounded continuous functions on $\mathcal{X}$ Dudley (2002)

Universal RKHS:
$k(x, x')$ continuous, $\mathcal{X}$ compact, and $\mathcal{F}$ dense in $C(\mathcal{X})$ with respect to L_∞ Steinwart (2001)

If $\mathcal{F}$ universal, then $MMD(P, Q; \mathcal{F}) = 0$ iff $P = Q$

Characteristic kernels (via Universality)

Proof:

First, it is clear that $P = Q$ implies $MMD(P, Q; \mathcal{F})$ is zero.

Converse: by the universality of $\mathcal{F}$, for any given $\epsilon > 0$ and $f \in C(\mathcal{X})$, $\exists g \in \mathcal{F}$

$$\|f - g\|_\infty \leq \epsilon.$$

We next make the expansion

$$\begin{aligned} &|\mathrm{E}_P f(x) - \mathrm{E}_Q f(y)| \\ &\leq |\mathrm{E}_P f(x) - \mathrm{E}_P g(x)| + |\mathrm{E}_P g(x) - \mathrm{E}_Q g(y)| + |\mathrm{E}_Q g(y) - \mathrm{E}_Q f(y)| . \end{aligned}$$

The first and third terms satisfy

$$|\mathrm{E}_P f(x) - \mathrm{E}_P g(x)| \leq \mathrm{E}_P |f(x) - g(x)| \leq \epsilon.$$

Characteristic kernels (via Universality)

Proof:

First, it is clear that $P = Q$ implies $MMD(P, Q; \mathcal{F})$ is zero.

Converse: by the universality of $\mathcal{F}$, for any given $\epsilon > 0$ and $f \in C(\mathcal{X})$, $\exists g \in \mathcal{F}$

$$\|f - g\|_\infty \leq \epsilon.$$

We next make the expansion

$$\begin{aligned}
&|\mathrm{E}_P f(x) - \mathrm{E}_Q f(y)| \\
&\leq |\mathrm{E}_P f(x) - \mathrm{E}_P g(x)| + |\mathrm{E}_P g(x) - \mathrm{E}_Q g(y)| + |\mathrm{E}_Q g(y) - \mathrm{E}_Q f(y)| .
\end{aligned}$$

The first and third terms satisfy

$$|\mathrm{E}_P f(x) - \mathrm{E}_P g(x)| \leq \mathrm{E}_P |f(x) - g(x)| \leq \epsilon.$$

Characteristic kernels (via Universality)

Proof (continued):

$$E_P g(x) - E_Q g(y) = \langle g(\cdot), \mu_P - \mu_Q \rangle_{\mathcal{F}} = 0,$$

since $MMD(P, Q; \mathcal{F}) = 0$ implies $\mu_P = \mu_Q$. Hence

$$|E_P f(x) - E_Q f(y)| \leq 2\epsilon$$

for all $f \in C(\mathcal{X})$ and $\epsilon > 0$, which implies $P = Q$.

How to choose the best kernel (2) optimising the kernel parameters

The best test for the job

- A test's power depends on $k(x, x')$, P, and Q (and n)
- With characteristic kernel, MMD test has power $\to 1$ as $n \to \infty$ for any (fixed) problem
  - But, for many P and Q, will have terrible power with reasonable n!

The best test for the job

- A test's power depends on $k(x, x')$, P, and Q (and n)
- With characteristic kernel, MMD test has power $\to 1$ as $n \to \infty$ for any (fixed) problem
  - But, for many P and Q, will have terrible power with reasonable n!
- You *can* choose a good kernel for a given problem
- You *can't* get one kernel that has good finite-sample power for all problems

Choosing a kernel for the test

- Simple choice: exponentiated quadratic
$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$
- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$

Choosing a kernel for the test

- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

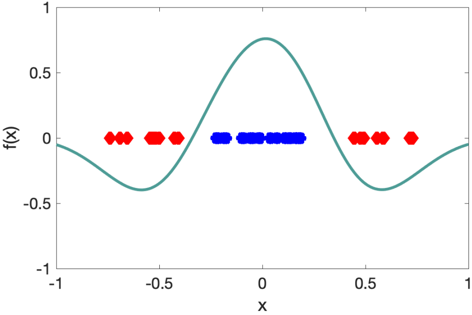
- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...

Choosing a kernel for the test

- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...

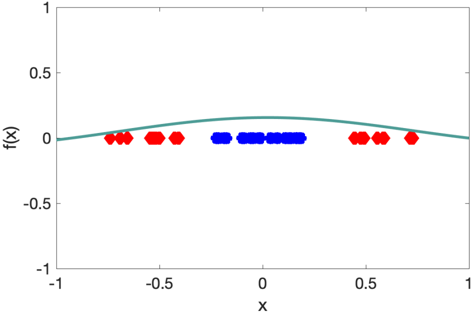


Choosing a kernel for the test

- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...

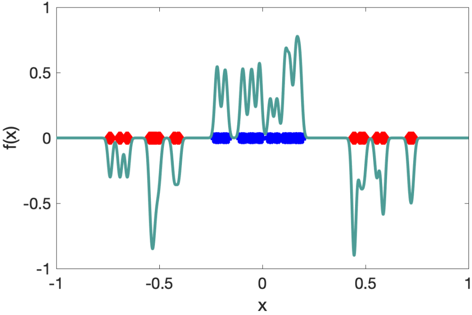


Choosing a kernel for the test

- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...



Choosing a kernel for the test

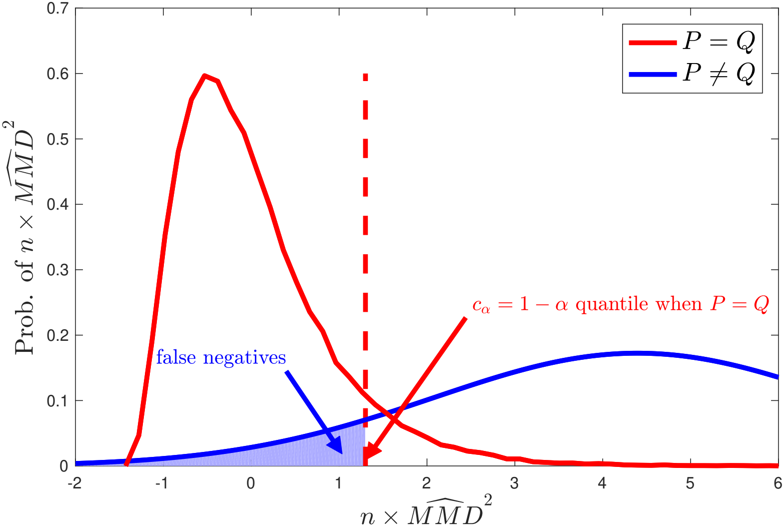
- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...
- ...and some problems (e.g. images) might have no good choice for σ

Graphical illustration

- Maximising test power same as minimizing false negatives



Optimizing kernel for test power

The power of our test (Pr_1 denotes probability under $P \neq Q$):

$$\mathrm{Pr}_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$

Optimizing kernel for test power

The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$
$$\to \Phi\left(\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} - \frac{c_\alpha}{n\sqrt{V_n(P,Q)}}\right)$$

where

- Φ is the CDF of the standard normal distribution.
- $\hat{c}_\alpha$ is an estimate of c_α test threshold.

Optimizing kernel for test power

The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

$$\begin{aligned} &\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right) \\ &\to \Phi\left(\underbrace{\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}}_{O(n^{1/2})} - \underbrace{\frac{c_\alpha}{n\sqrt{V_n(P,Q)}}}_{O(n^{-1/2})}\right)\end{aligned}$$

For large n, second term negligible!

Optimizing kernel for test power

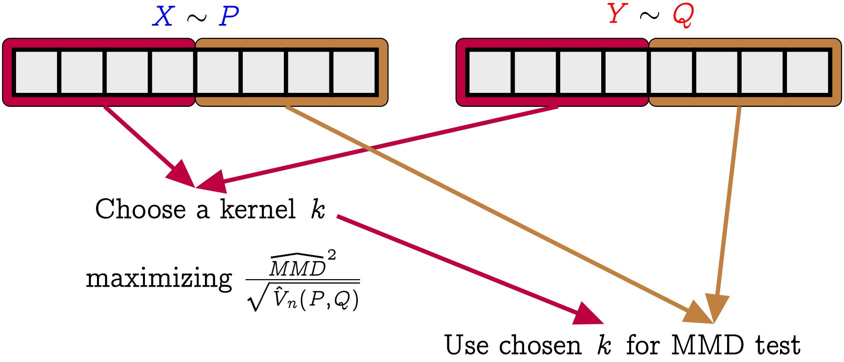
The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$
$$\to \Phi\left(\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} - \frac{c_\alpha}{n\sqrt{V_n(P,Q)}}\right)$$

To maximize test power, maximize

$$\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}$$

Data splitting

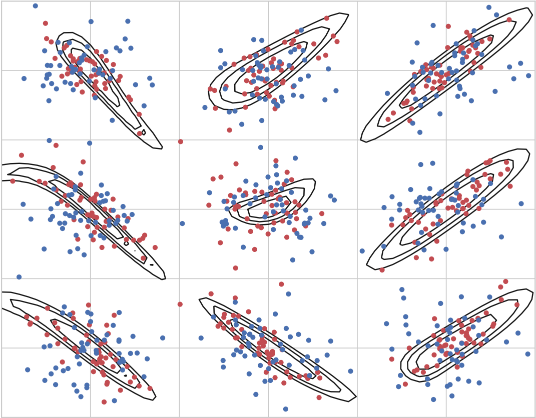


Learning a kernel helps a lot

Kernel with deep learned features:

$$k_\theta(x, y) = [(1-\epsilon)\kappa(\Phi_\theta(x), \Phi_\theta(y)) + \epsilon]\ q(x, y)$$

κ and q are Gaussian kernels



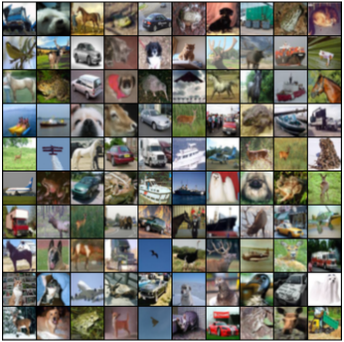
Learning a kernel helps a lot

Kernel with deep learned features:

$$k_\theta(x, y) = [(1-\epsilon)\kappa(\Phi_\theta(x), \Phi_\theta(y)) + \epsilon]\, q(x, y)$$

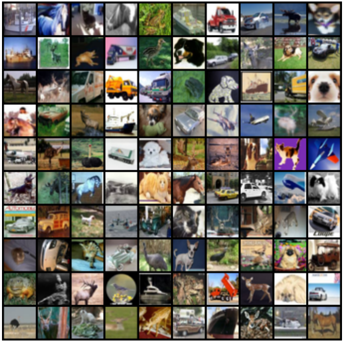
κ and q are Gaussian kernels

- CIFAR-10 vs CIFAR-10.1, null rejected 75% of time



CIFAR-10 test set (Krizhevsky 2009)

$X \sim P$



CIFAR-10.1 (Recht+ ICML 2019)

$Y \sim Q$

Learning a kernel helps a lot

Kernel with deep learned features:

$$k_\theta(x, y) = [(1-\epsilon)\kappa(\Phi_\theta(x), \Phi_\theta(y)) + \epsilon]\, q(x, y)$$

κ and q are Gaussian kernels

- CIFAR-10 vs CIFAR-10.1, null rejected 75% of time

arXiv.org > stat > arXiv:2002.09116

Statistics > Machine Learning

[Submitted on 21 Feb 2020]

Learning Deep Kernels for Non-Parametric Two-Sample Tests

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, D. J. Sutherland

ICML 2020

How to choose the best kernel (2)
minimax test without data splitting

Two-sample problem

Our aim: find a condition on $\|p - q\|_2$ to control Type II error β

$$\mathbb{P}_{p\times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \;\leq\; \beta$$

Remaining settings:

- samples $\mathbb{X}_m := (X_1, \ldots, X_m)$, $X_i \overset{\text{iid}}{\sim} p$ in $\mathbb{R}^d$
- samples $\mathbb{Y}_n := (Y_1, \ldots, Y_n)$, $Y_i \overset{\text{iid}}{\sim} q$ in $\mathbb{R}^d$

$\mathcal{H}_0 \colon p = q$	against	$\mathcal{H}_1 \colon p \neq q$
$\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1$	$\Longleftrightarrow$	reject $\mathcal{H}_0$
$\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0$	$\Longleftrightarrow$	fail to reject $\mathcal{H}_0$

Type I error: controlled by α by design

$$\mathbb{P}_{p\times p}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1) \;\leq\; \alpha$$

Two-sample problem

Our aim: find a condition on $\|p - q\|_2$ to control Type II error β

$$\mathbb{P}_{p \times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \;\leq\; \beta$$

Remaining settings:

- samples $\mathbb{X}_m := (X_1, \ldots, X_m)$, $X_i \overset{\text{iid}}{\sim} p$ in $\mathbb{R}^d$
- samples $\mathbb{Y}_n := (Y_1, \ldots, Y_n)$, $Y_i \overset{\text{iid}}{\sim} q$ in $\mathbb{R}^d$

$\mathcal{H}_0 \colon p = q$	against	$\mathcal{H}_1 \colon p \neq q$
$\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1$	$\Longleftrightarrow$	reject $\mathcal{H}_0$
$\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0$	$\Longleftrightarrow$	fail to reject $\mathcal{H}_0$

Type I error: controlled by α by design

$$\mathbb{P}_{p \times p}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1) \;\leq\; \alpha$$

Kernels and bandwidths

Kernel: $k_\lambda(x, y) := \prod_{i=1}^{d} K_i\left(\frac{x_i - y_i}{\lambda_i}\right)$ Bandwidth: $\lambda \in (0, \infty)^d$

Assumptions: $K_1, \ldots, K_d$ integrable and square integrable

Examples: Gaussian ($K_i(u) = e^{-u^2}$), Laplace ($K_i(u) = e^{-|u|}$), Matérn

Gaussian kernel: $k_\lambda(x, y) := \exp\left(-\sum_{i=1}^{d} \frac{(x_i - y_i)^2}{\lambda_i^2}\right)$

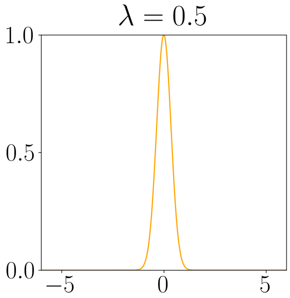
Kernels and bandwidths

Kernel: $k_{\lambda}(x, y) := \prod_{i=1}^{d} K_i\left(\frac{x_i - y_i}{\lambda_i}\right)$ Bandwidth: $\lambda \in (0, \infty)^d$

Assumptions: $K_1, \ldots, K_d$ integrable and square integrable

Examples: Gaussian ($K_i(u) = e^{-u^2}$), Laplace ($K_i(u) = e^{-|u|}$), Matérn

Gaussian kernel: $k_{\lambda}(x, y) := \exp\left(-\sum_{i=1}^{d} \frac{(x_i - y_i)^2}{{\lambda_i}^2}\right)$



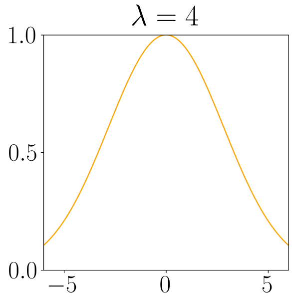
Kernels and bandwidths

Kernel: $k_\lambda(x, y) := \prod_{i=1}^{d} K_i\left(\frac{x_i - y_i}{\lambda_i}\right)$ Bandwidth: $\lambda \in (0, \infty)^d$

Assumptions: $K_1, \ldots, K_d$ integrable and square integrable

Examples: Gaussian ($K_i(u) = e^{-u^2}$), Laplace ($K_i(u) = e^{-|u|}$), Matérn

Gaussian kernel: $k_\lambda(x, y) := \exp\left(-\sum_{i=1}^{d} \frac{(x_i - y_i)^2}{\lambda_i^{\,2}}\right)$



MMDAgg for a *collection* of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}_{1-\alpha/|\Lambda|}^\lambda \text{ for some } \lambda \in \Lambda\right)$$

time complexity $\mathcal{O}\left(|\Lambda|\, B_1\, (m+n)^2\right)$

B_1 permutations for each $\lambda \in \Lambda$ to obtain thresholds $\widehat{q}_{1-\alpha/|\Lambda|}^\lambda$

MMDAgg for a *collection* of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}^{\,\lambda}_{1-\alpha/|\Lambda|} \text{ for some } \lambda \in \Lambda\right)$$

time complexity $\mathcal{O}\left(|\Lambda|\, B_1\, (m+n)^2\right)$

MMDAgg (MMD Aggregation): non-asymptotic level α

$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}^{\,\lambda}_{1-u_\alpha/|\Lambda} \text{ for some } \lambda \in \Lambda\right)$$

MMDAgg for a *collection* of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}_{1-\alpha/|\Lambda|}^\lambda \text{ for some } \lambda \in \Lambda\right)$$

time complexity $\mathcal{O}\left(|\Lambda|\, B_1\, (m+n)^2\right)$

MMDAgg (MMD Aggregation): non-asymptotic level α

$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}_{1-u_\alpha/|\Lambda}^\lambda \text{ for some } \lambda \in \Lambda\right)$$

Correction u_α defined as

$$\sup\left\{u > 0 : \mathbb{P}_{p\times p}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) - \widehat{q}_{1-u/|\Lambda}^\lambda\right) > 0\right) \leq \alpha\right\}$$

more powerful than Bonferroni correction as $u_\alpha \geq \alpha$

MMDAgg for a *collection* of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}^{\,\lambda}_{1-\alpha/|\Lambda|} \text{ for some } \lambda \in \Lambda\right)$$

time complexity $\mathcal{O}\Big(|\Lambda|\, B_1\, (m+n)^2\Big)$

MMDAgg (MMD Aggregation): non-asymptotic level α

$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}^{\,\lambda}_{1-u_\alpha/|\Lambda} \text{ for some } \lambda \in \Lambda\right)$$

Correction u_α defined as

$$\sup\left\{u > 0 : \mathbb{P}_{p\times p}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) - \widehat{q}^{\,\lambda}_{1-u/|\Lambda}\right) > 0\right) \leq \alpha\right\}$$

more powerful than Bonferroni correction as $u_\alpha \geq \alpha$

B_2 permutations to simulate draws from $\mathbb{P}_{p\times p}$, search over u

MMDAgg for a *collection* of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}^{\,\lambda}_{1-\alpha/|\Lambda|} \text{ for some } \lambda \in \Lambda\right)$$

time complexity $\mathcal{O}\left(|\Lambda|\, B_1\, (m+n)^2\right)$

MMDAgg (MMD Aggregation): non-asymptotic level α

$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}^{\,\lambda}_{1-u_\alpha/|\Lambda} \text{ for some } \lambda \in \Lambda\right)$$

Correction u_α defined as

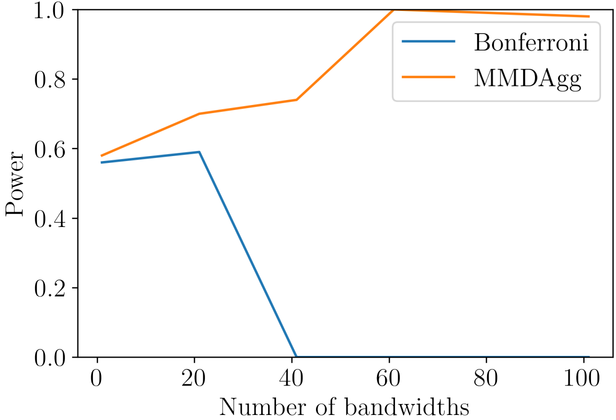
$$\sup\left\{u > 0 : \mathbb{P}_{p\times p}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) - \widehat{q}^{\,\lambda}_{1-u/|\Lambda}\right) > 0\right) \leq \alpha\right\}$$

more powerful than Bonferroni correction as $u_\alpha \geq \alpha$

Time complexity $\mathcal{O}\left(|\Lambda|\,(B_1 + B_2)\,(m+n)^2\right)$

Multiple testing correction comparison

Simple example: 3-d Gaussians with different means



$$\Lambda(i) := \left\{2^{\ell}\lambda_{\text{med}} : \ell \in \{-i, \dots, i\}\right\} \text{ for } i \in \{0, 10, 20, 30, 40, 50\}$$

MMDAgg parameter-free user-friendly implementation

mmdagg package: github.com/antoninschrab/mmdagg

```
from mmdagg import mmdagg                  # X shape (m, d)
output = mmdagg(X, Y) # 0 or 1             # Y shape (n, d)
```

JAX: runs on either CPU or GPU (significant speed improvements)

- JAX GPU runs 100 times faster than Numpy CPU

arXiv > stat > arXiv:2110.15073

Statistics > Machine Learning

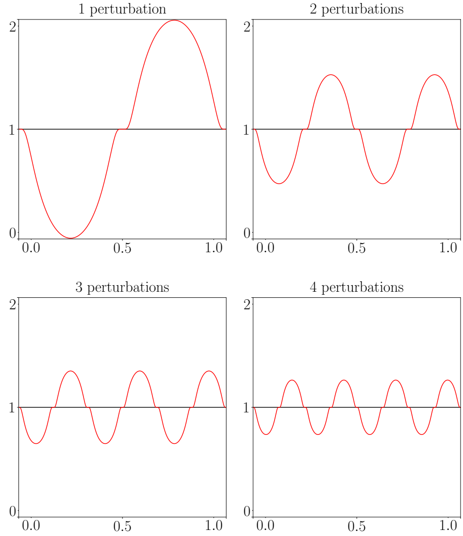
[Submitted on 28 Oct 2021 (v1), last revised 21 Aug 2023 (this version, v4)]

MMD Aggregated Two-Sample Test

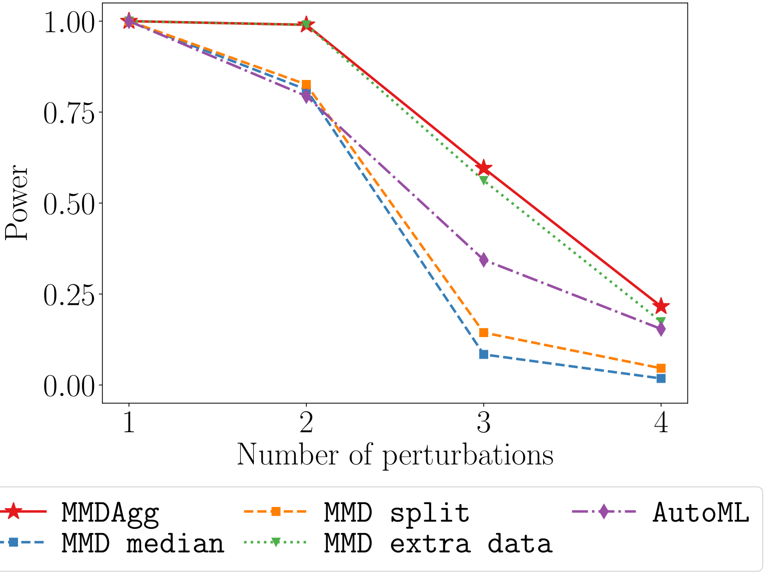
Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, Arthur Gretton

JMLR 2023

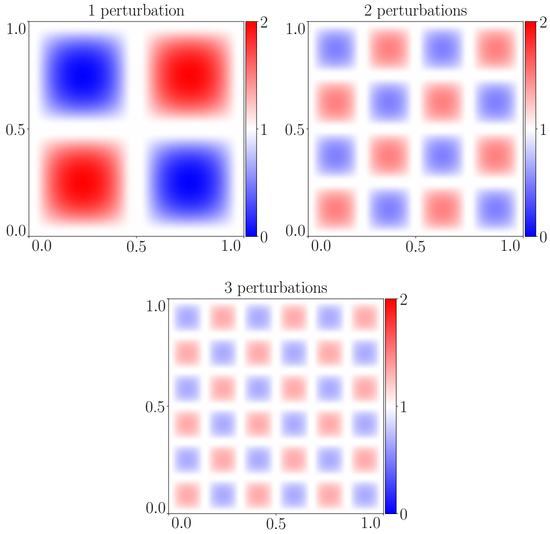
Experiment on perturbed uniform $d = 1$



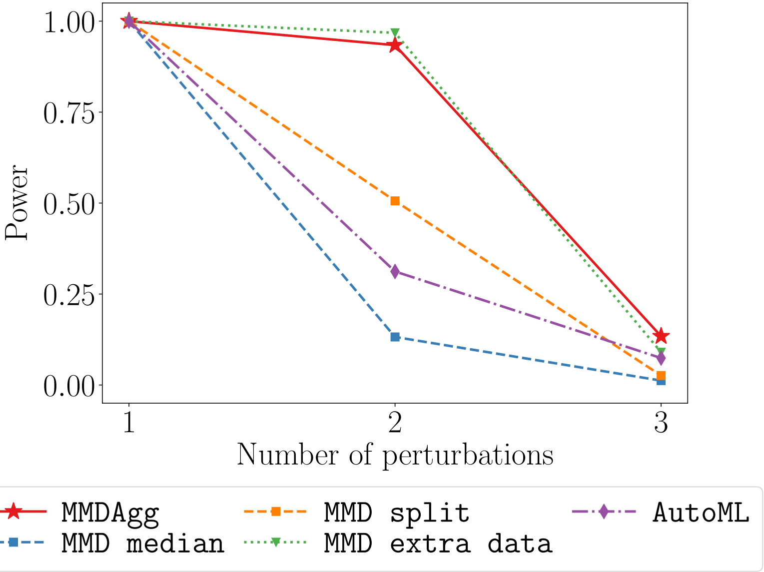
Experiment on perturbed uniform $d = 1$



Experiment on perturbed uniform $d = 2$

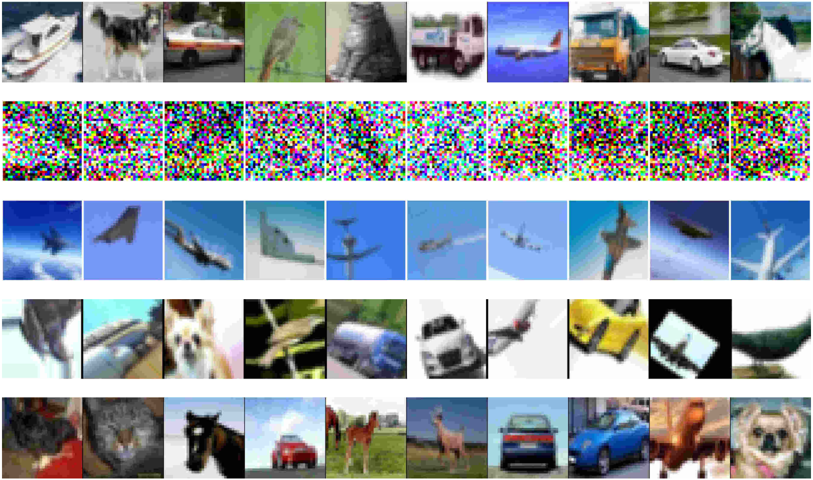


Experiment on perturbed uniform $d = 2$



Experiment on image shifts on MNIST & CIFAR-10

Failing Loudly Benchmark: Rabanser et al., 2019



Experiment on image shifts on MNIST & CIFAR-10