Representing and comparing probabilities with kernels: Part 2

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

MLSS Madrid, 2018

Comparing two samples

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

$$\sim P$$

$$\sim Q$$

Outline

Two sample testing

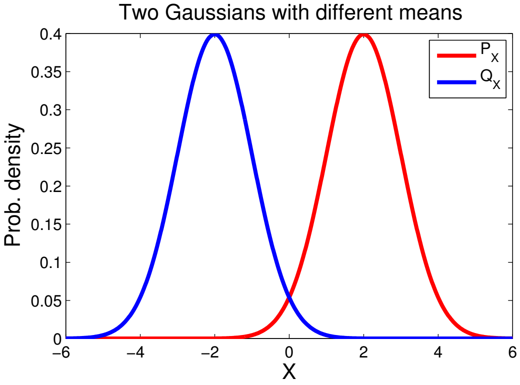
- Test statistic: Maximum Mean Discrepancy (MMD)...
  - ...as a difference in feature means
  - ...as an integral probability metric (not just a technicality!)
- Statistical testing with the MMD
- "How to choose the best kernel"

Training GANs with MMD

Maximum Mean Discrepancy

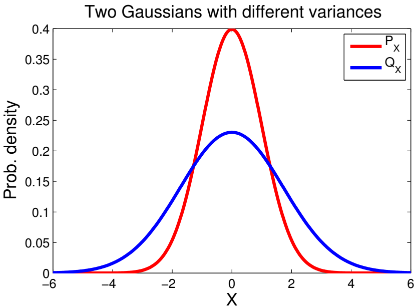
Feature mean difference

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



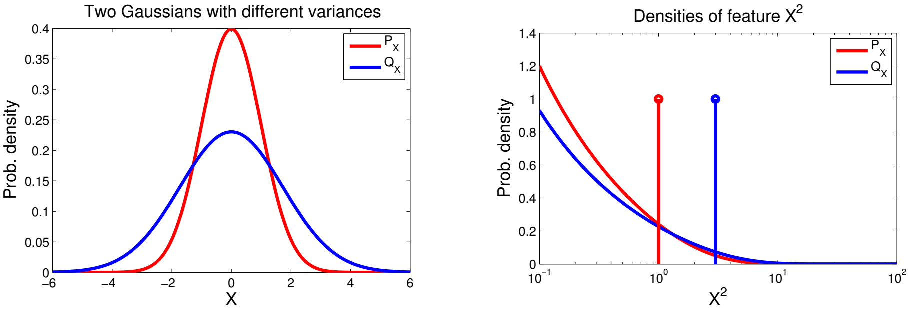
Feature mean difference

- 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 $\varphi(x) = x^2$



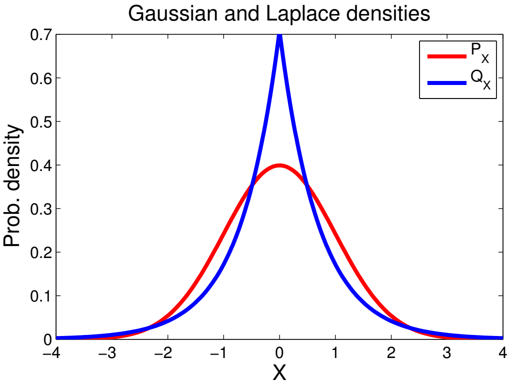
Feature mean difference

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

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

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!

Infinitely many features using kernels

Kernels: dot products of features

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

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

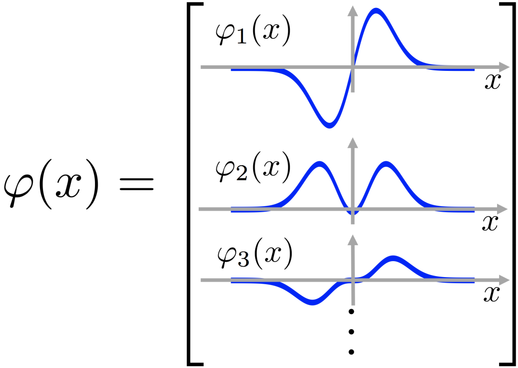
For positive definite k,

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

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

Exponentiated quadratic kernel

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



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,

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

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

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

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

Fine print: feature map $\varphi(x)$ must be Bochner integrable for all probability measures considered. Always true if kernel bounded.

Infinitely many features of *distributions*

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

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

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

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

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

Fine print: feature map $\varphi(x)$ must be Bochner integrable for all probability measures considered. Always true if kernel bounded.

The maximum mean discrepancy

The maximum mean discrepancy is the distance between **feature means**:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between **feature means**:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between **feature means**:

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

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

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 $k(\text{fish}_i, \text{fish}_j)$

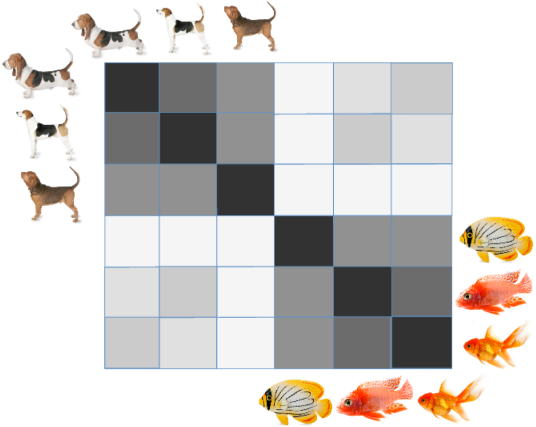
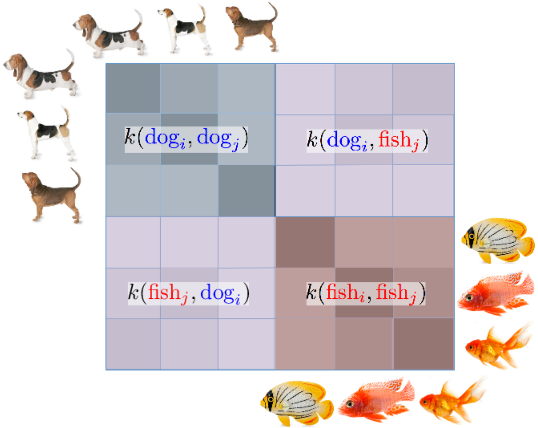


Illustration of MMD

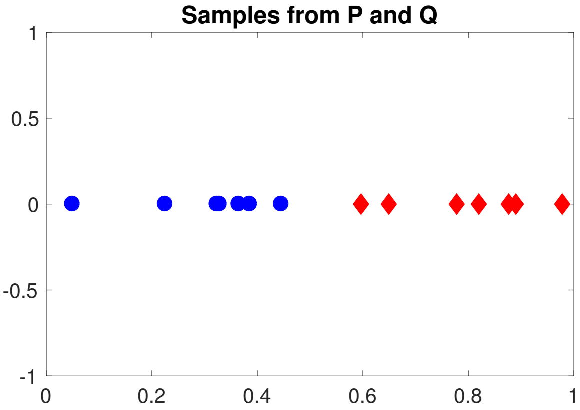
The maximum mean discrepancy:

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



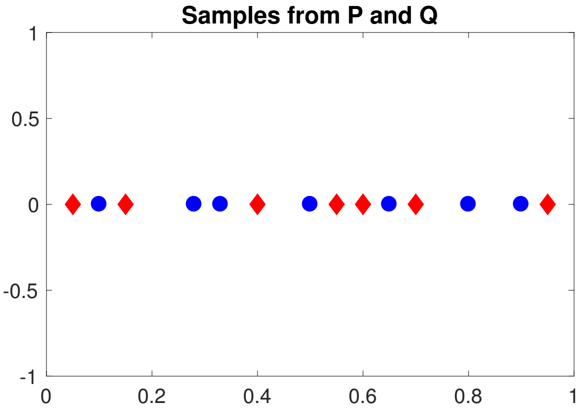
MMD as an integral probability metric

Are P and Q different?



MMD as an integral probability metric

Are P and Q different?

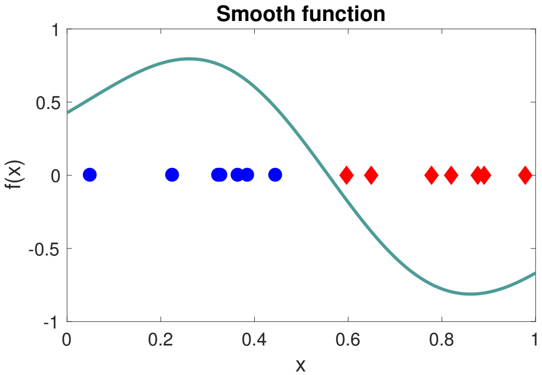


MMD as an integral probability metric

Integral probability metric:

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

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

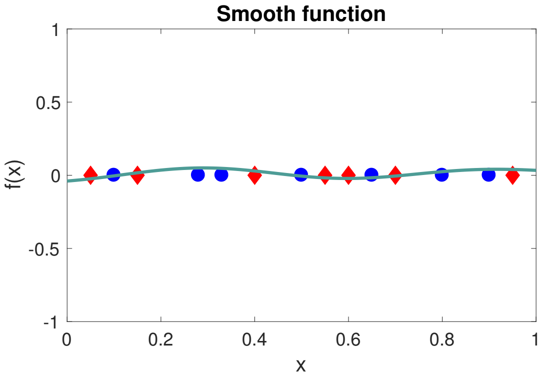


MMD as an integral probability metric

Integral probability metric:

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

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



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

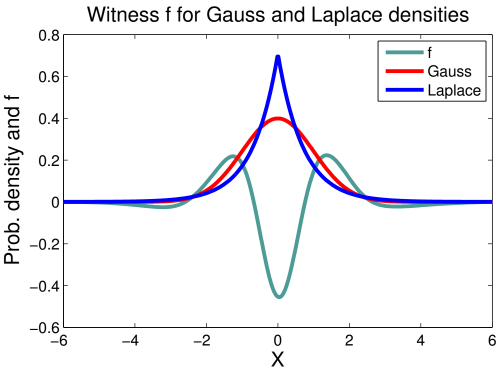
(F = unit ball in RKHS $\mathcal{F}$)

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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



MMD as an integral probability metric

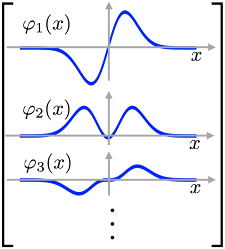
Maximum mean discrepancy: smooth function for P vs Q

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

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

Functions are linear combinations of features:

$$f(x) = \langle f, \varphi(x) \rangle_{\mathcal{F}} = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x) = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{bmatrix}^\top$$



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

Expectations of functions are linear combinations of expected features

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

(always true if kernel is bounded)

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

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

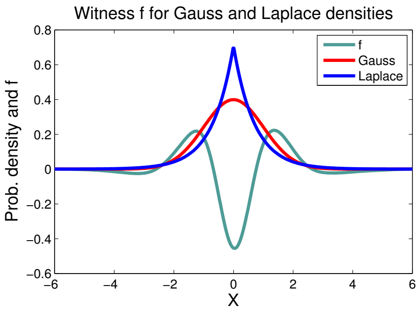
Other choices for witness function class:

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

Integral prob. metric vs feature difference

The MMD:

$$MMD(P, Q; F) = \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right]$$



Integral prob. metric vs feature difference

The MMD:

use

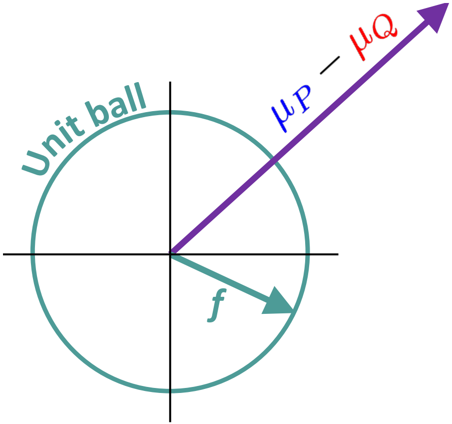
$$\mathbf{E}_P f(X) = \langle \mu_P, f \rangle_{\mathcal{F}}$$

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

Integral prob. metric vs feature difference

The MMD:

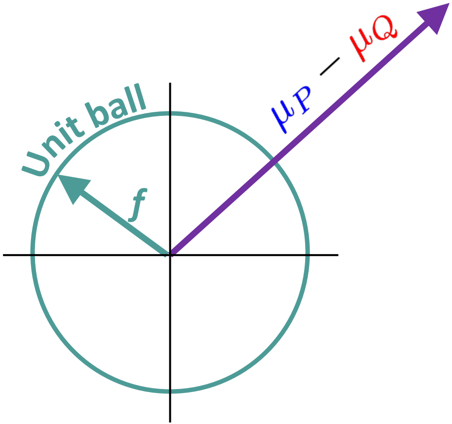
$$
\begin{aligned}
& MMD(P, Q; F) \\
& = \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)\right] \\
& = \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



Integral prob. metric vs feature difference

The MMD:

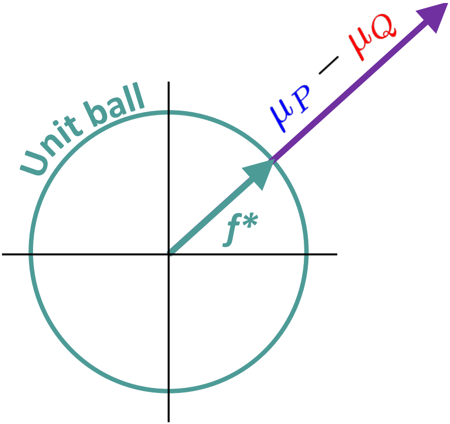
$$
\begin{aligned}
& MMD(P, Q; F) \\
&= \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right] \\
&= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



Integral prob. metric vs feature difference

The MMD:

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



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

Integral prob. metric vs feature difference

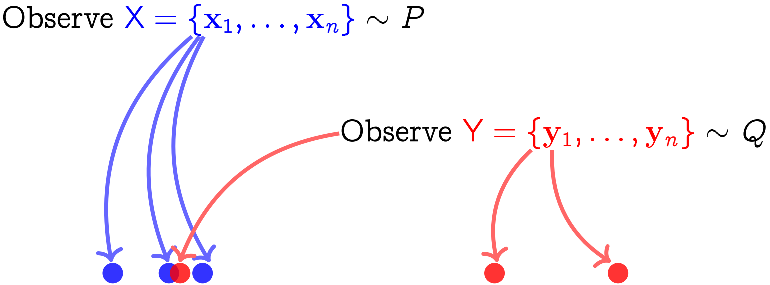
The MMD:

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

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

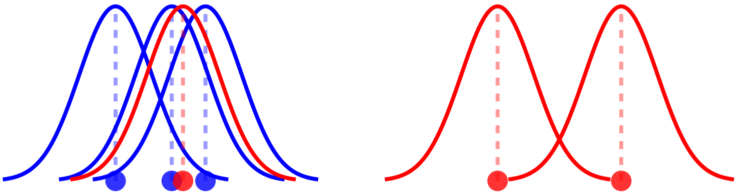
Construction of MMD witness

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



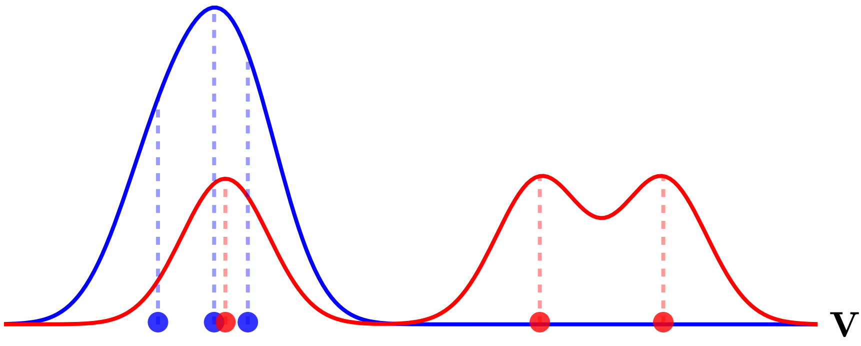
Construction of MMD witness

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



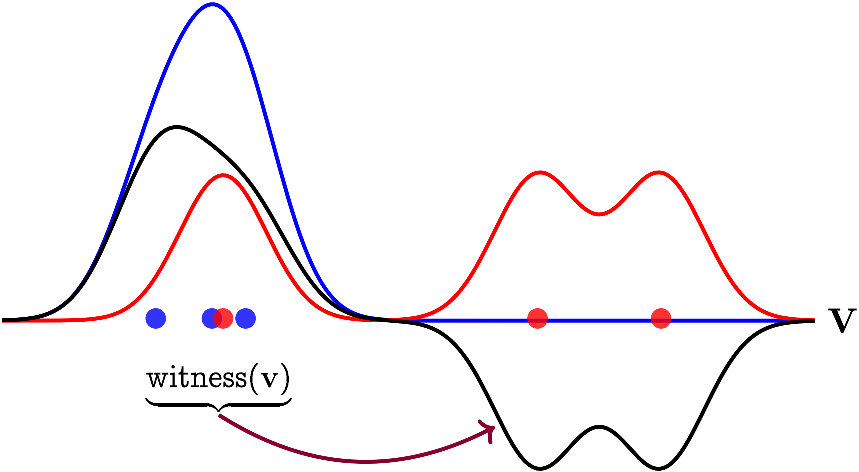
Construction of MMD witness

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

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

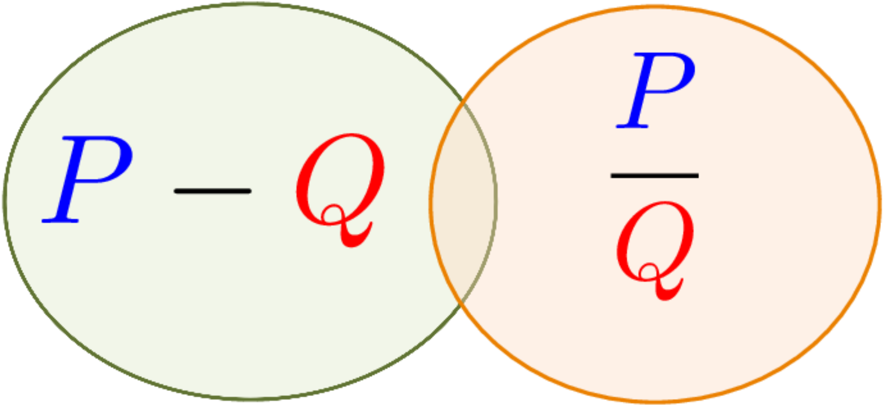
The empirical witness function at v

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

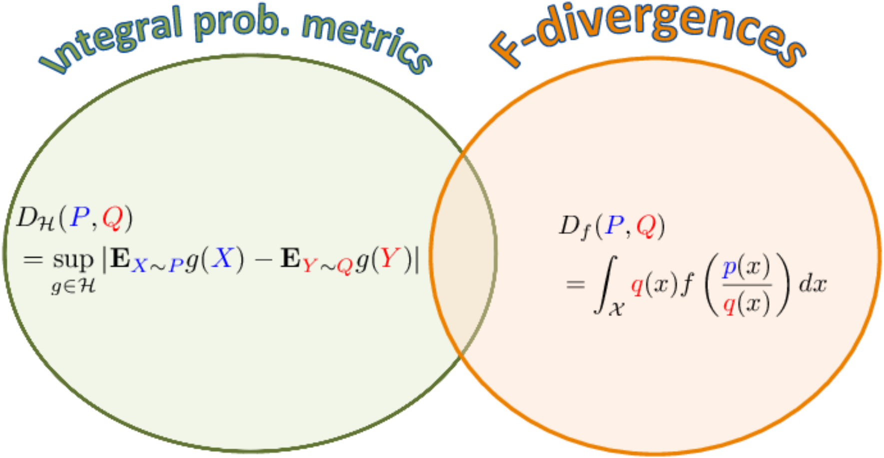
Don't need explicit feature coefficients $f^* := \begin{bmatrix} f_1^* & f_2^* & \dots \end{bmatrix}$

Interlude: divergence measures

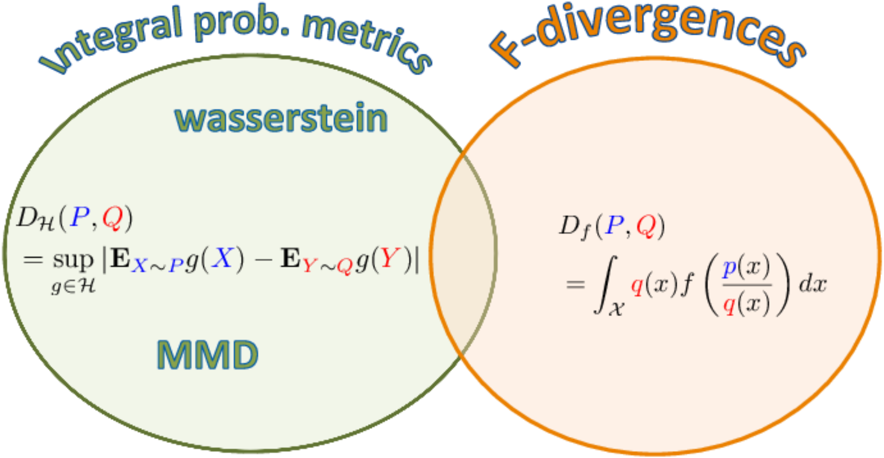
Divergences



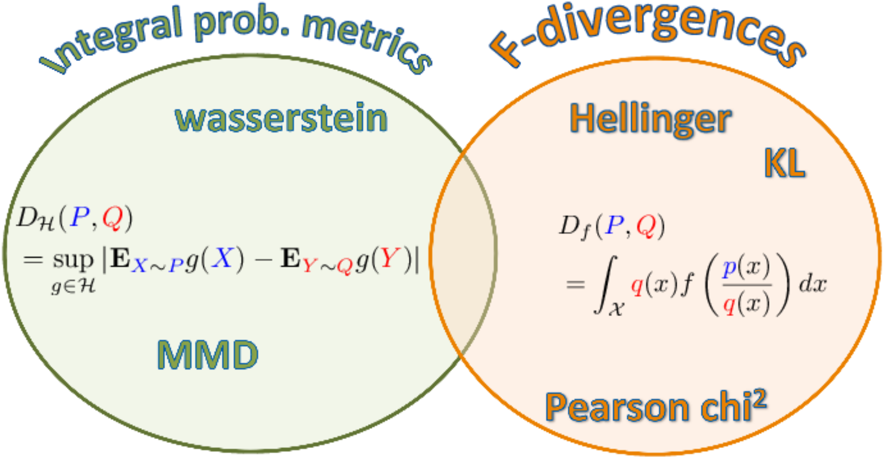
Divergences



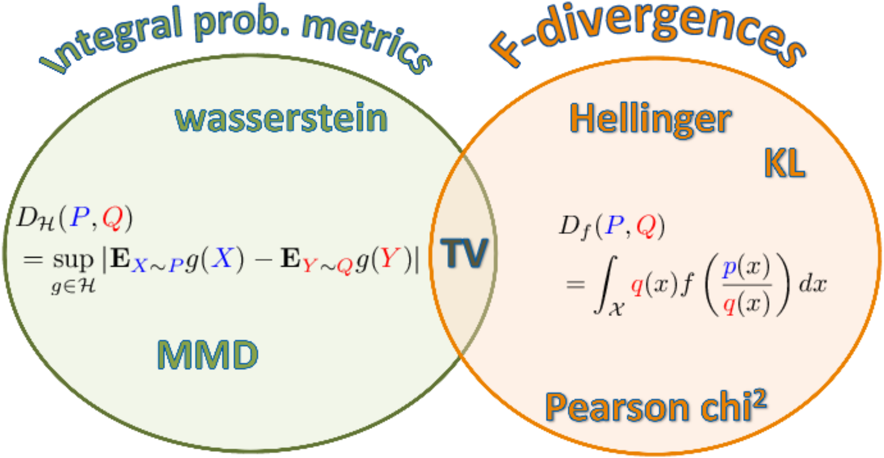
Divergences



Divergences



Divergences



Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet (2012)

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"

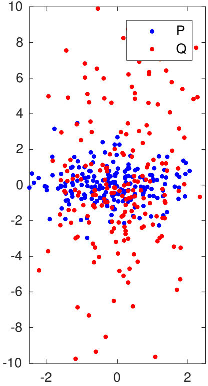
Want Threshold c_α for $\widehat{MMD}^2$ to get false positive rate α

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

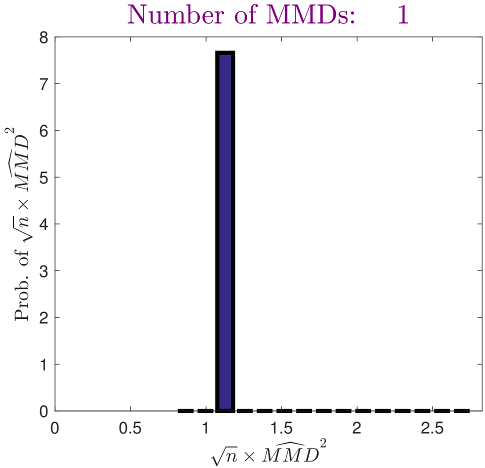
$\sqrt{n} \times \widehat{MMD}^2 = 1.2$

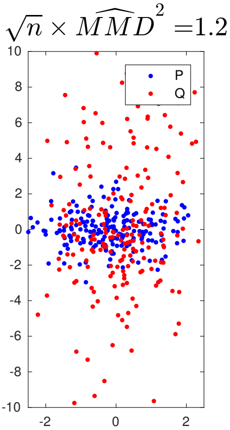


Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

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

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

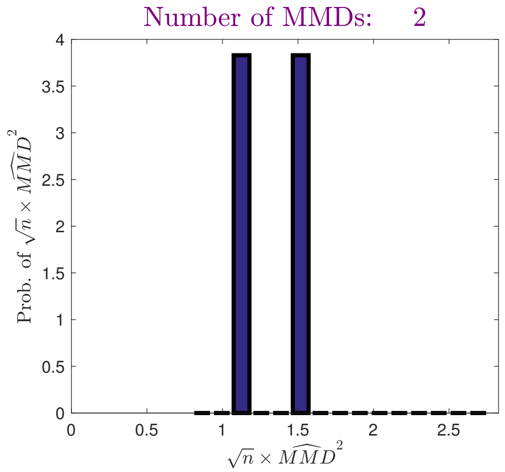


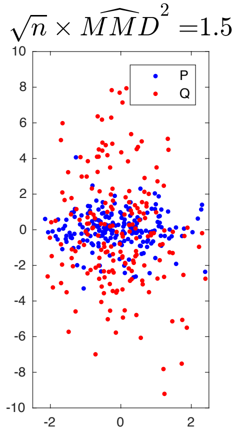


Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

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

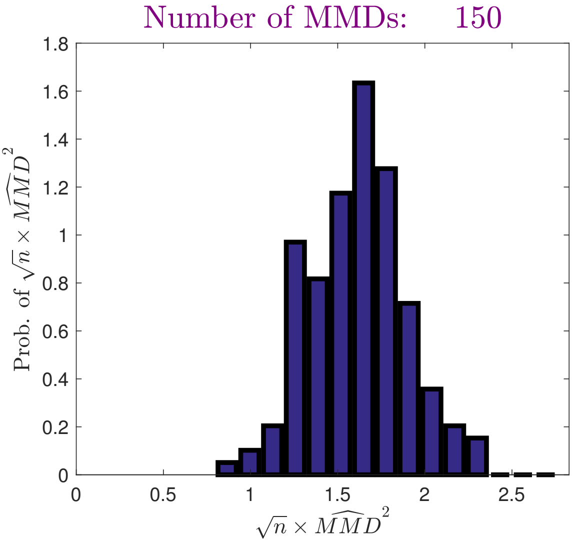
- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.5$





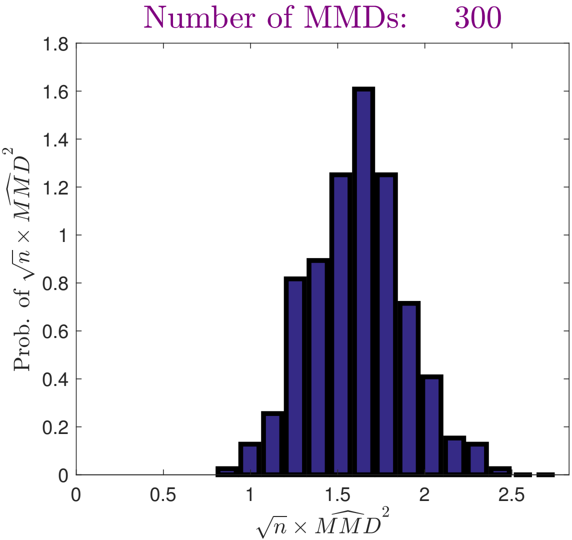
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 150 times ...



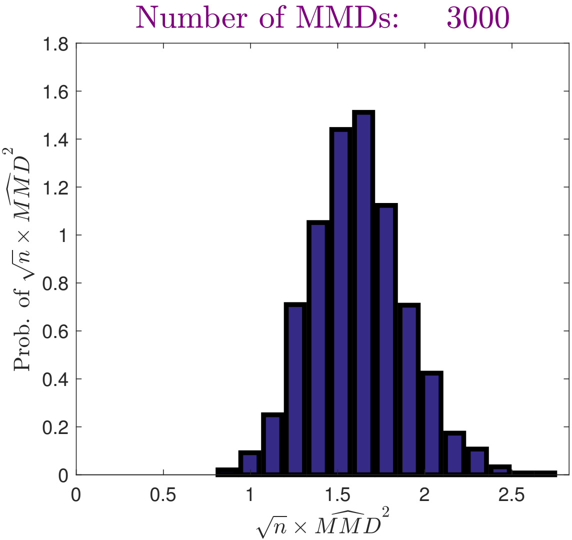
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 300 times ...



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

Repeat this 3000 times . . .

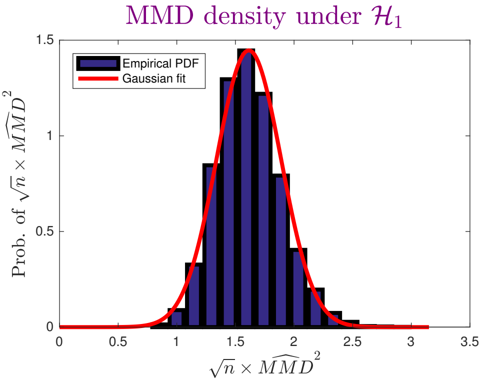


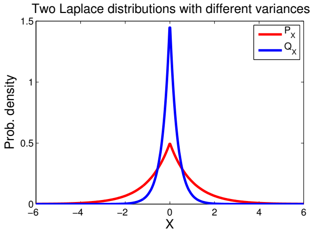
Asymptotics of $\widehat{MMD}^2$ when $P \neq Q$

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

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

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



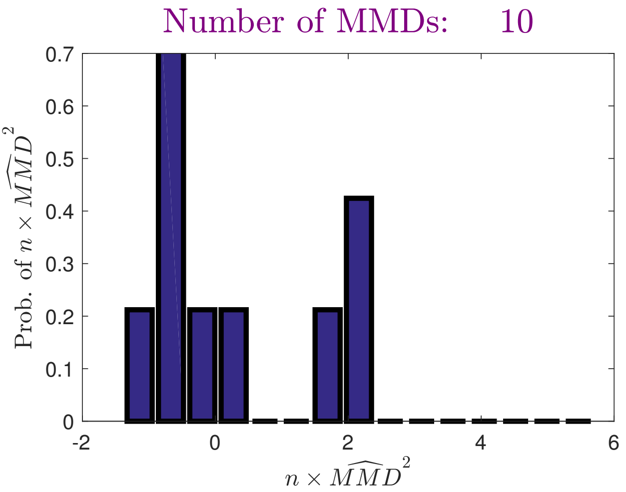


Behaviour of $\widehat{MMD}^2$ when $P = Q$

What happens when P and Q are the same?

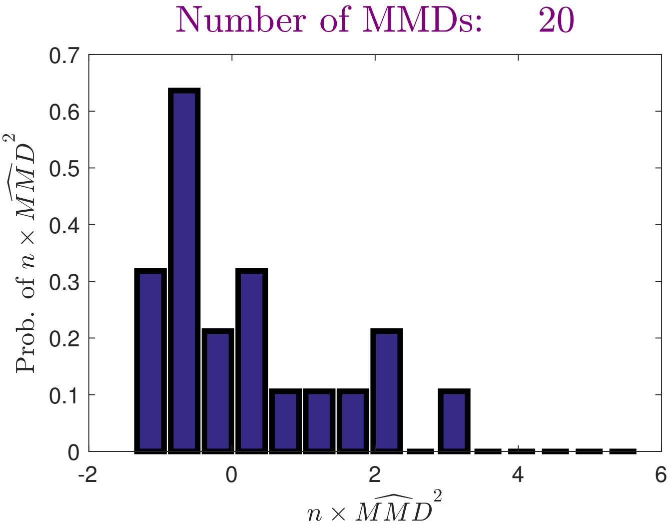
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



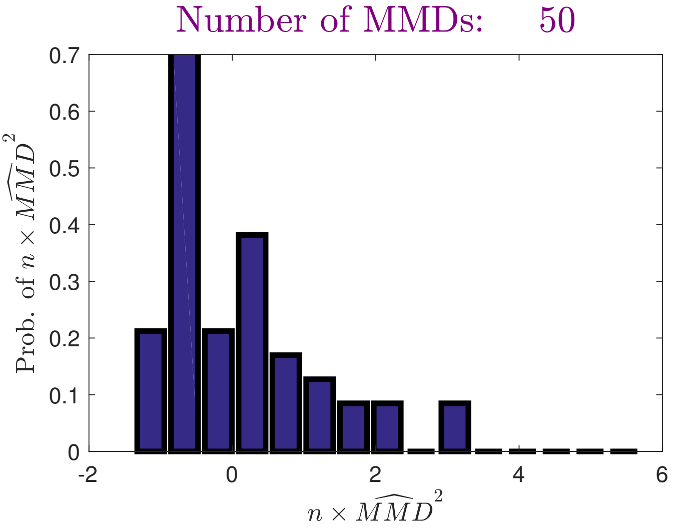
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



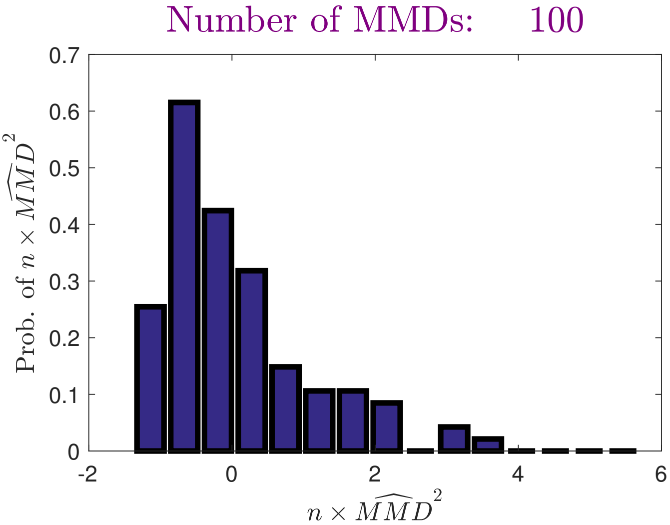
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



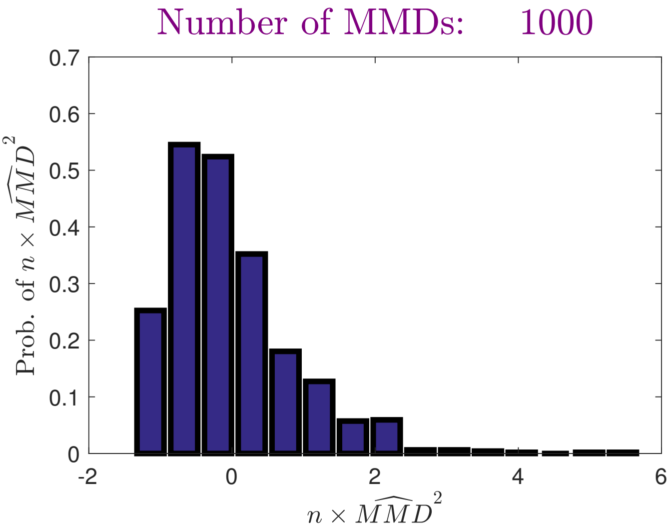
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



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

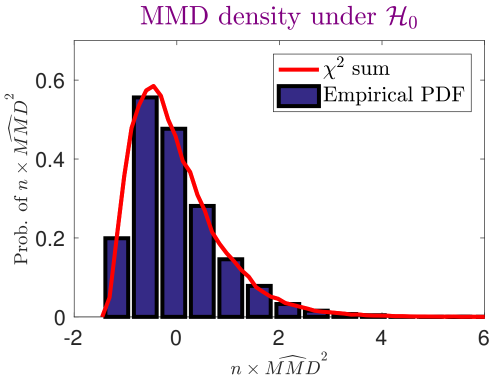
- Case of $P = Q = \mathcal{N}(0,1)$



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

Where $P = Q$, statistic has asymptotic distribution

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



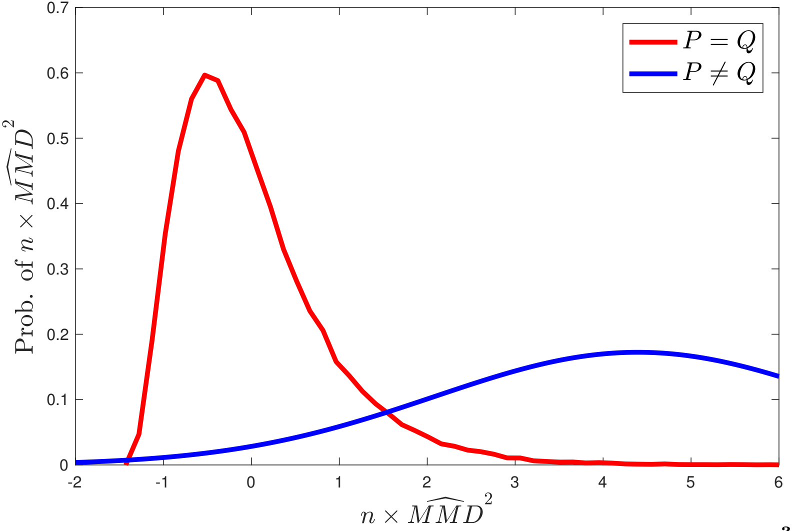
where

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

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

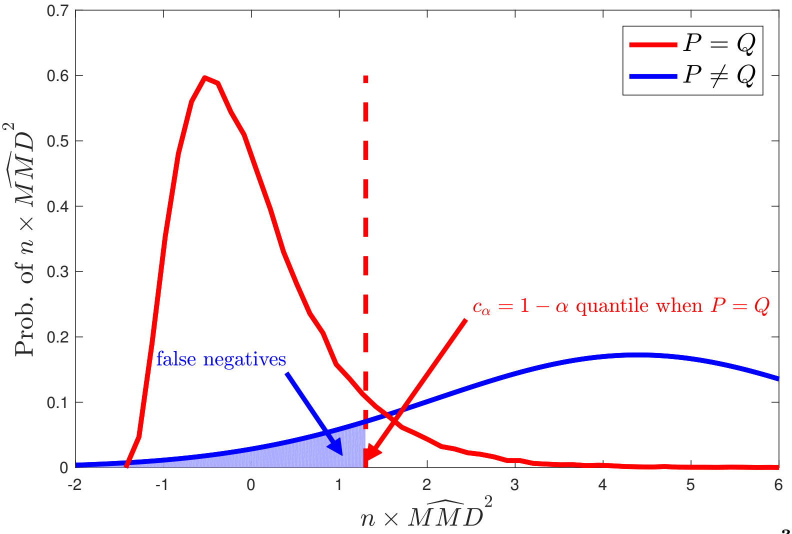
A statistical test

A summary of the asymptotics:



A statistical test

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



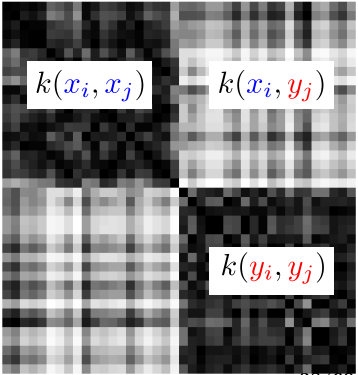
How do we get test threshold c_α?

Original empirical MMD for dogs and fish:

$X = [$  $\ldots\ ]$

$Y = [$  $\ldots\ ]$

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



How do we get test threshold c_α?

Permuted dog and fish samples (**merdogs**):

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

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

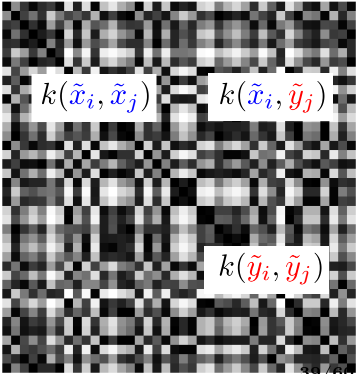
How do we get test threshold c_α?

Permuted dog and fish samples (**merdogs**):

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

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

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



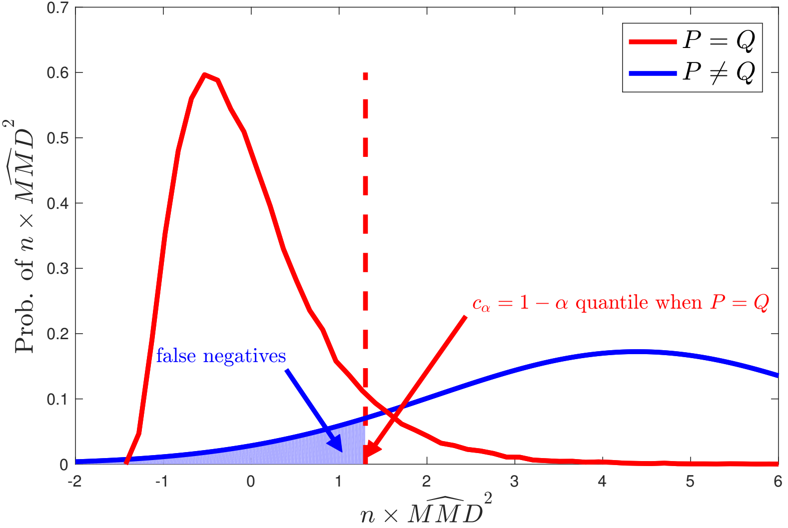
Permutation simulates

$P = Q$

How to choose the best kernel (1) optimising the kernel parameters

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 $P \neq Q$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$

Optimizing kernel for test power

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

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

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 $P \neq Q$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$

$$\to \Phi\Bigg(\underbrace{\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}}_{O(n^{1/2})} - \underbrace{\frac{c_\alpha}{n\sqrt{V_n(P,Q)}}}_{O(n^{-1/2})}\Bigg)$$

Variance under $\mathcal{H}_1$ decreases as $\sqrt{V_n(P,Q)} \sim O(n^{-1/2})$

For large n, second term negligible!

Optimizing kernel for test power

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

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

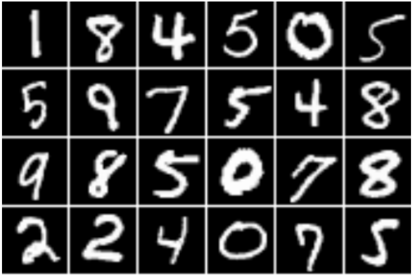
To maximize test power, maximize

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

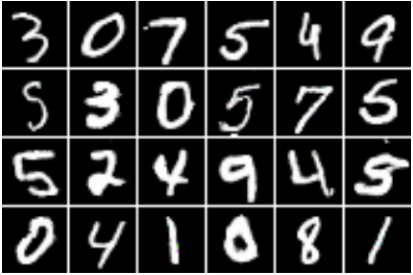
(Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017)

Code: github.com/dougalsutherland/opt-mmd

Troubleshooting for generative adversarial networks



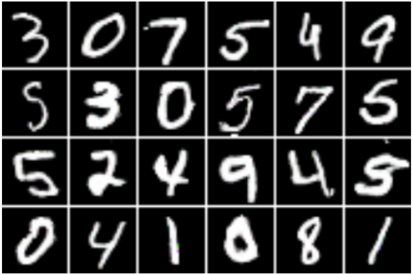
MNIST samples



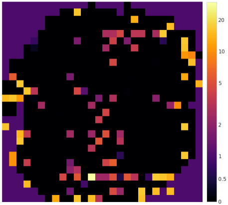
Samples from a GAN

Troubleshooting for generative adversarial networks

MNIST samples



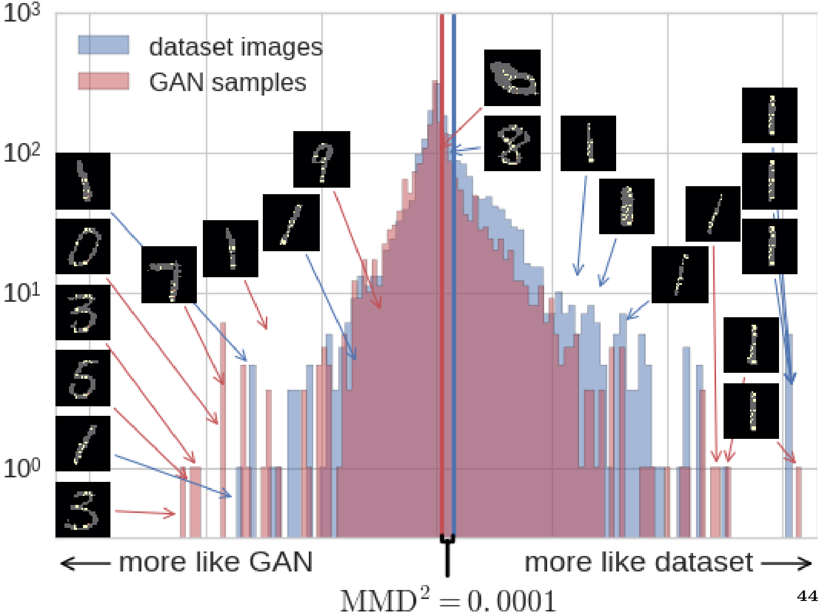
Samples from a GAN



ARD map

- Power for **optimzed ARD kernel**: 1.00 at $\alpha = 0.01$
- Power for optimized RBF kernel: 0.57 at $\alpha = 0.01$

Troubleshooting generative adversarial networks



How to choose the best kernel (2)
characteristic kernels

Characteristic kernels

Characteristic: MMD a metric $MMD = 0$ iff $P = Q$)
[NIPS07b, JMLR10]

In the next slides:

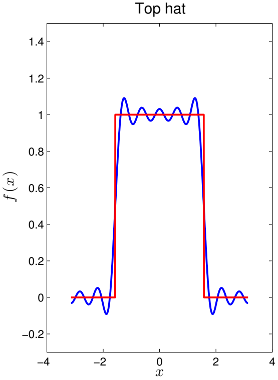
- Characteristic property on $[-\pi, \pi]$ with periodic boundary
- Characteristic property on $\mathbb{R}^d$

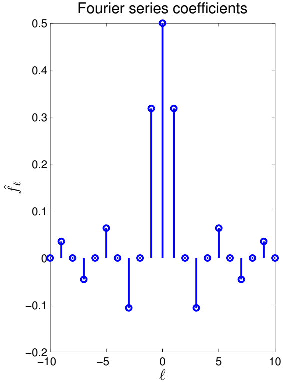
Characteristic kernels on $[-\pi, \pi]$

Reminder: Fourier series

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

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



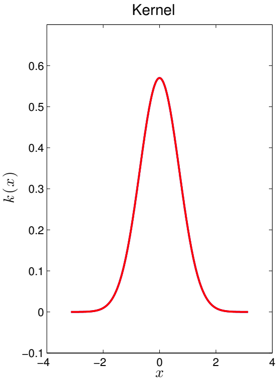


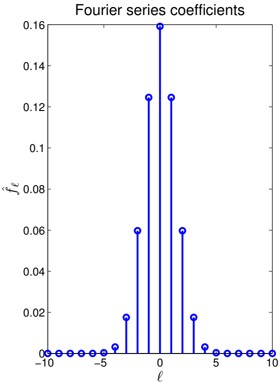
Characteristic kernels on $[-\pi, \pi]$

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)

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

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)

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

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)

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

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)

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

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)

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

MMD can be written in terms of Fourier series:

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

A simpler Fourier representation for MMD

From previous slide,

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

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:

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

A simpler Fourier representation for MMD

From previous slide,

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

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

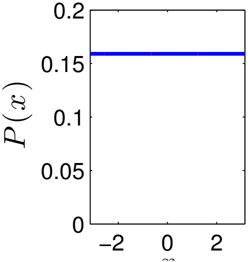
$$\|f\|_{\mathcal{F}}^2 = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_\ell|^2}{\hat{k}_\ell}.$$

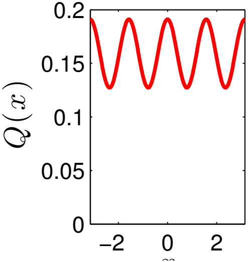
Simple, interpretable expression for squared MMD:

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

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

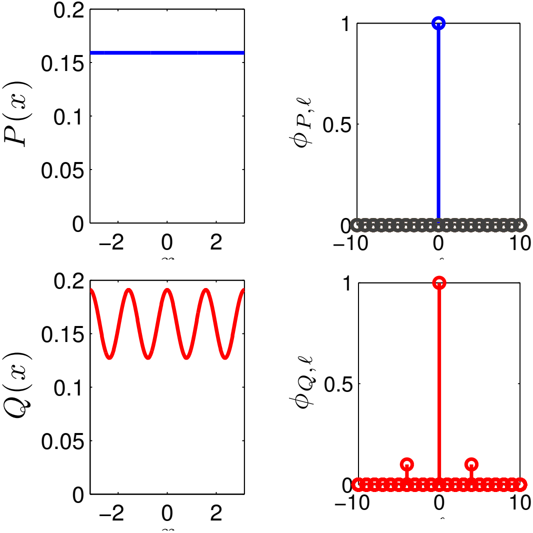
Example: P differs from Q at one frequency:





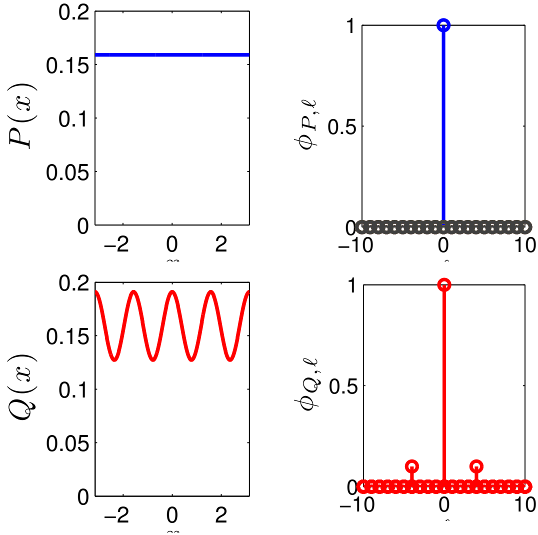
Characteristic kernels on $[-\pi, \pi]$

Example: P differs from Q at one frequency:

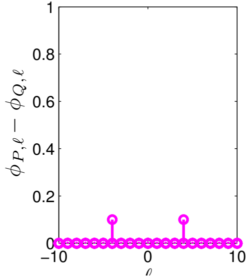


Characteristic kernels on $[-\pi, \pi]$

Example: P differs from Q at one frequency:

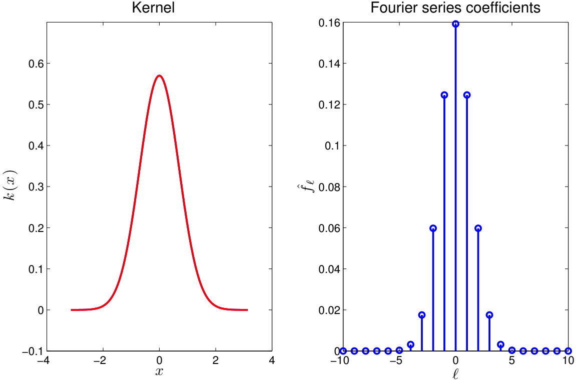


Characteristic function difference



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

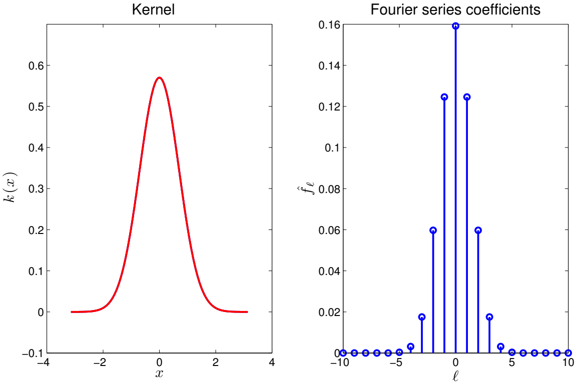
Is the Gaussian spectrum kernel characteristic?



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

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

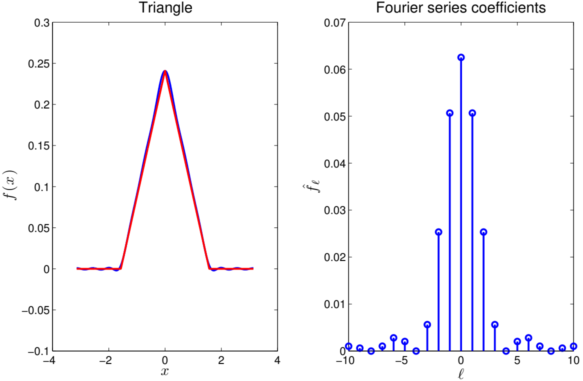
Is the Gaussian spectrum kernel characteristic? YES



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

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

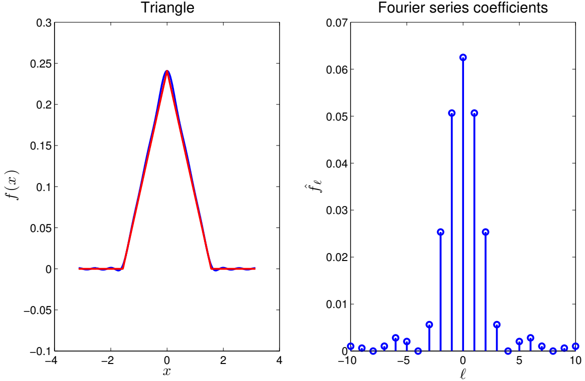
Is the triangle kernel characteristic?



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

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: an exercise! But recall the Fourier series case for $[-\pi, \pi]$:

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

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

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

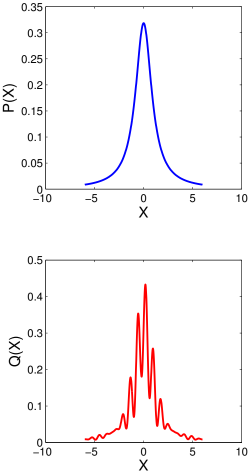
$$MMD^2(P, Q; F) = \int |\varphi_P(\omega) - \varphi_Q(\omega)|^2 \, d\Lambda(\omega)$$

Proof: an exercise! But recall the Fourier series case for $[-\pi, \pi]$:

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

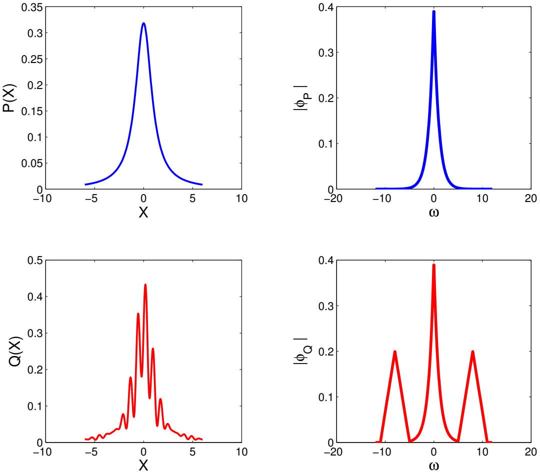
Characteristic kernels on $\mathbb{R}^d$

Example: P differs from Q at roughly one frequency:



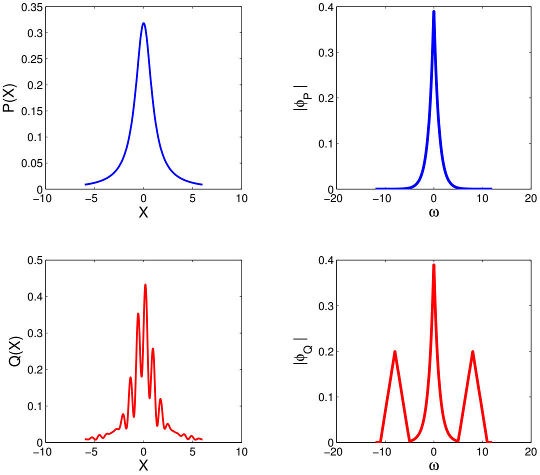
Characteristic kernels on $\mathbb{R}^d$

Example: P differs from Q at roughly one frequency:

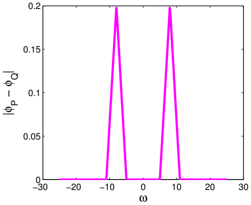


Characteristic kernels on $\mathbb{R}^d$

Example: P differs from Q at roughly one frequency:



Characteristic function difference

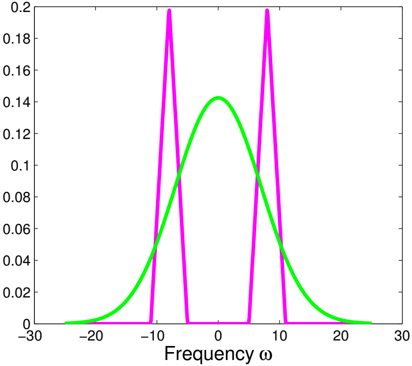


Characteristic kernels on $\mathbb{R}^d$

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

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

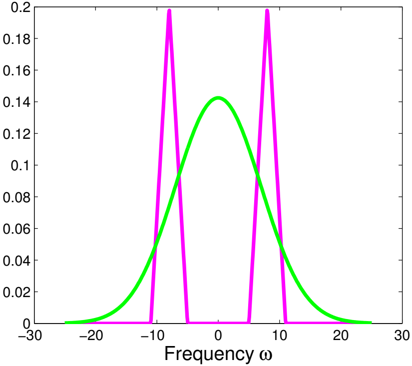
Difference $|\varphi_P - \varphi_Q|$



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

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

Characteristic

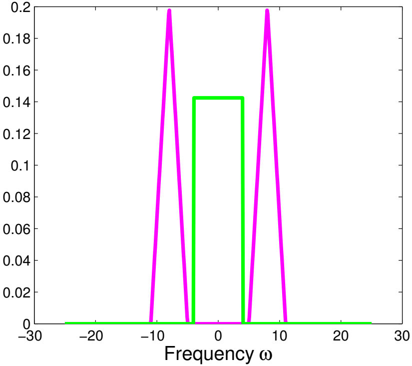


Characteristic kernels on $\mathbb{R}^d$

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

Sinc kernel spectrum $\Lambda(\omega)$

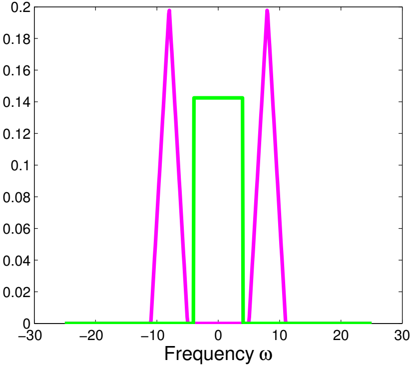
Difference $|\varphi_P - \varphi_Q|$



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

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

Not characteristic

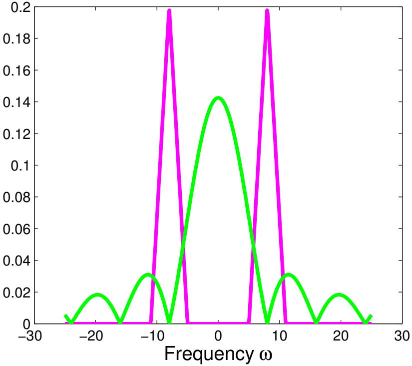


Characteristic kernels on $\mathbb{R}^d$

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

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

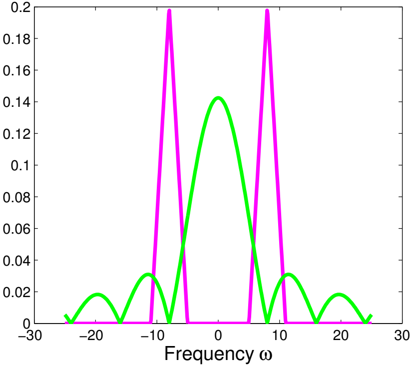
Difference $|\phi_P - \phi_Q|$



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

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

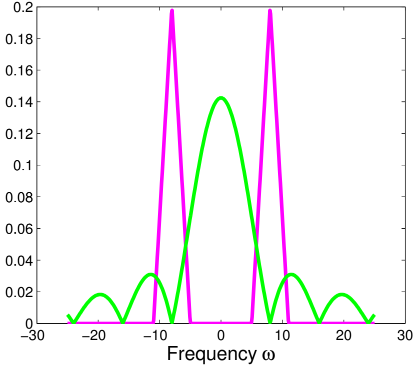
???



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

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

Characteristic



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]

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]