Kernel Methods for Two-Sample and Goodness-Of-Fit Testing

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

PHYSTAT, 2023

A motivation: comparing two samples

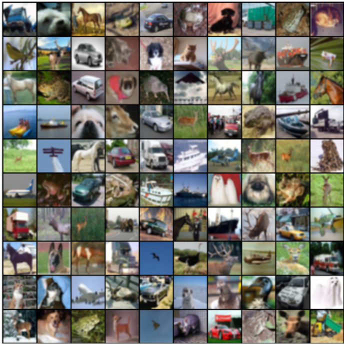
- Given: Samples from unknown distributions P and Q.
- Goal: do P and Q differ?

$$\sim P$$

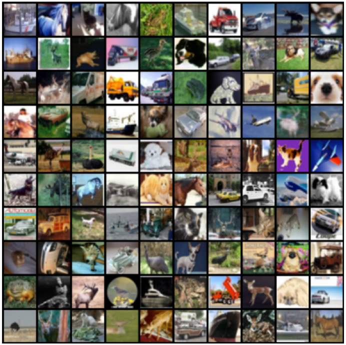
$$\sim Q$$

A real-life example: two-sample tests

- Goal: do P and Q differ?



CIFAR 10 samples



Cifar 10.1 samples

Significant difference?

G,Borgwardt, Rasch, Schoelkopf, Smola. A kernel two-sample test. JMLR 2012.

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

A second task: dependence testing

- Given: Samples from a distribution P_{XY}

X	Y
	A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.
	Their noses guide them through life, and they're never happier than when following an interesting scent.
	A responsive, interactive pet, one that will blow in your ear and follow you everywhere.

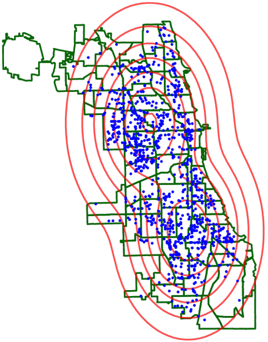
Text from dogtime.com and petfinder.com

G., Fukumizu, Teo, Song, Schoelkopf, Smola. A Kernel Statistical Test of Independence. NeurIPS 2007

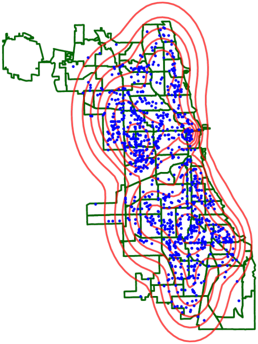
Chwialkopski, G. A kernel independence test for random processes. ICML 2023

A third task: model comparison

- Have: two candidate models P and Q, and samples $\{x_i\}_{i=1}^n$ from reference distribution R
- Goal: which of P and Q is better?



P : two components

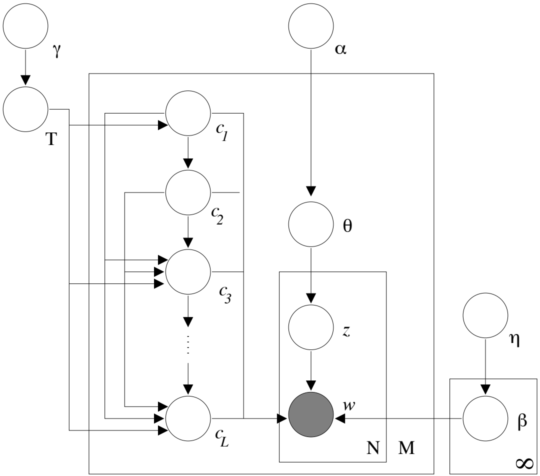


Q : ten components

Kanagawa, Jitkrittum, Mackey, Fukumizu, G., A Kernel Stein Test for comparing Latent Variable Models. JRSS B 2023.

Most interesting models have latent structure

Graphical model representation of hierarchical LDA with a nested CRP prior, Blei et al. (2003)



Outline

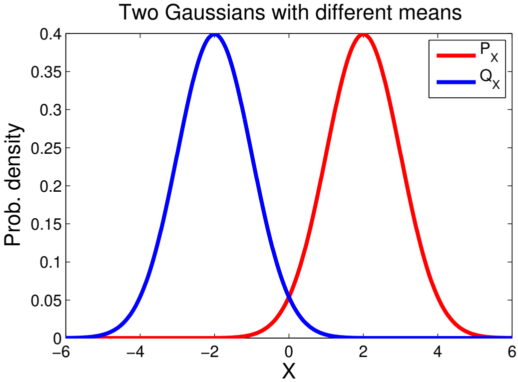
- Maximum Mean Discrepancy (MMD)...
  - ...as a difference in feature means
  - ...as an integral probability metric (not just a technicality!)

- A statistical test based on the MMD
  - learn adaptive NN features
  - learn interpretable features with maximium testing power

The MMD

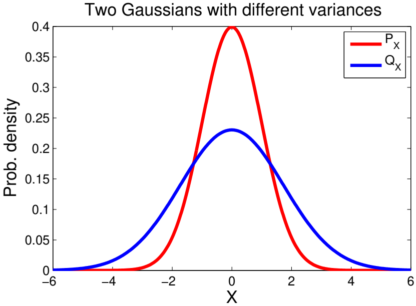
Feature mean difference

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



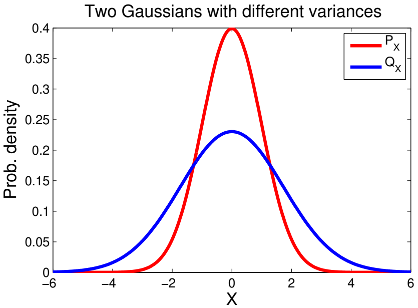
Feature mean difference

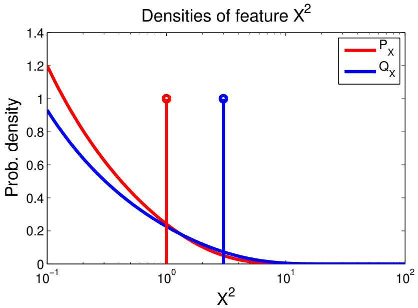
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$



Feature mean difference

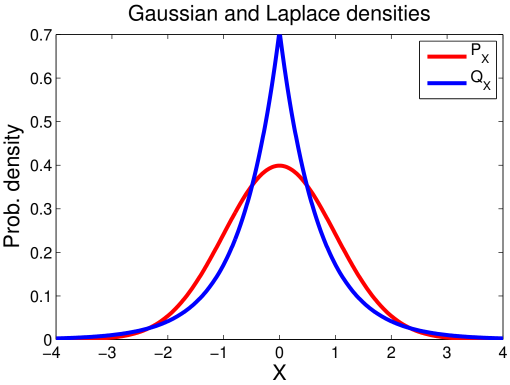
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$





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,

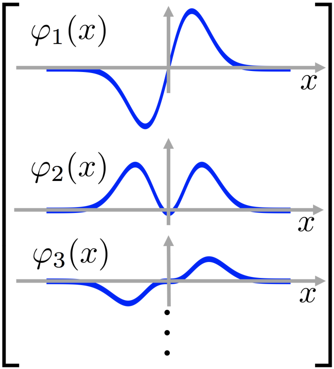
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

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

Exponentiated quadratic kernel

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

$$\varphi(x) =$$



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 $y \sim Q$.

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,

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

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 $y \sim Q$.

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:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\
&= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

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

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

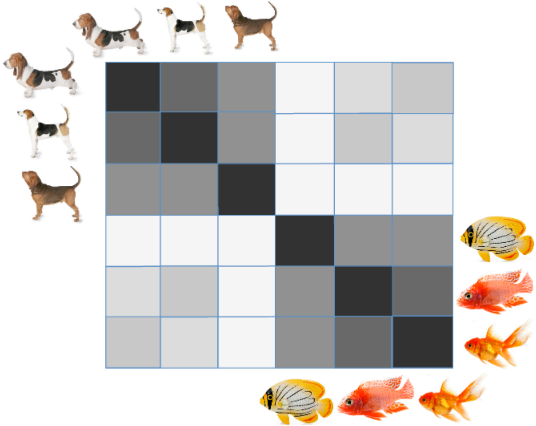
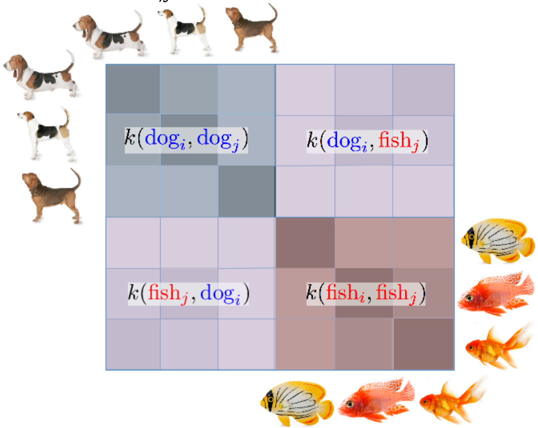


Illustration of MMD

The maximum mean discrepancy:

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

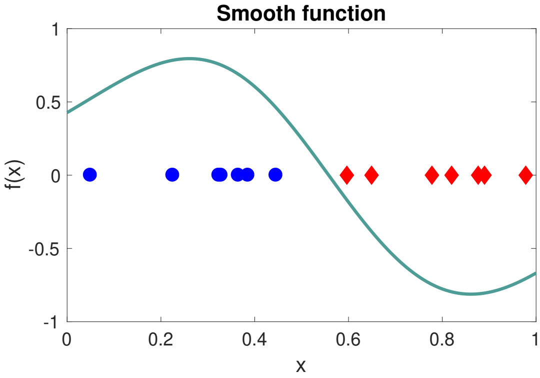


MMD as an integral probability metric

Integral probability metric:

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

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

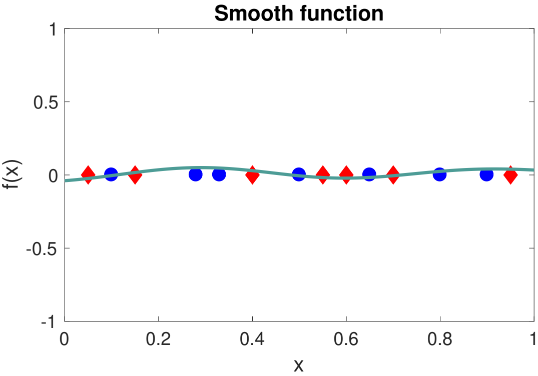


MMD as an integral probability metric

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

Maximum mean discrepancy: smooth function for P vs Q

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

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

MMD as an integral probability metric

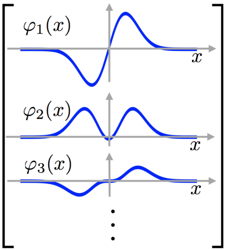
Maximum mean discrepancy: smooth function for P vs Q

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

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

Functions are linear combinations of features:

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



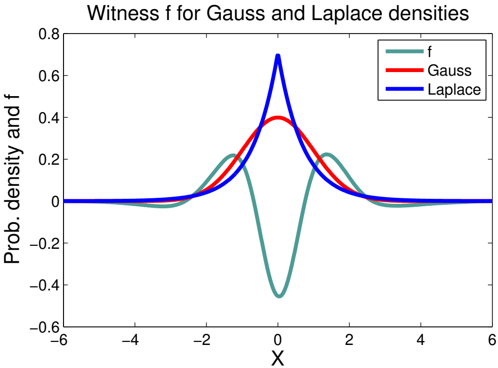
$$\|f\|_{\mathcal{F}}^2 := \sum_{i=1}^{\infty} f_i^2 \leq 1$$

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

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

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

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

Expectations of functions are linear combinations of expected features

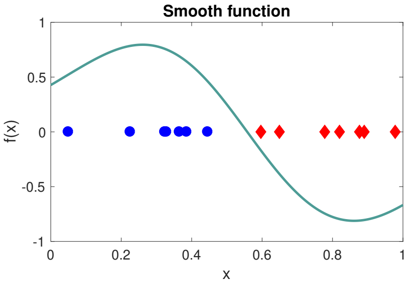
$$\mathrm{E}_P(f(X)) = \langle f, \mathrm{E}_P \varphi(X) \rangle_{\mathcal{F}} = \langle f, \mu_P \rangle_{\mathcal{F}}$$

(always true if kernel is bounded)

Integral prob. metric vs feature mean difference

The MMD:

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



Integral prob. metric vs feature mean difference

The MMD:

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

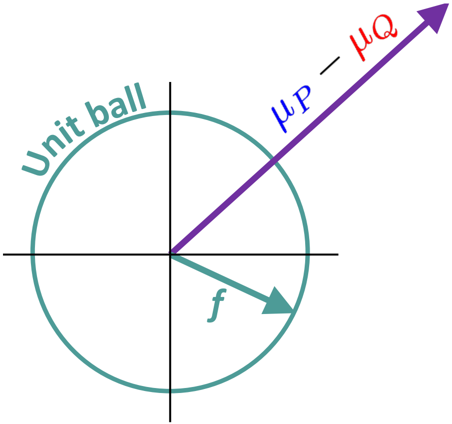
use

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

Integral prob. metric vs feature mean difference

The MMD:

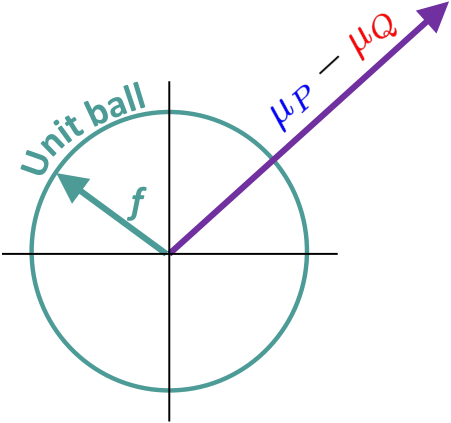
$$MMD(P, Q; F)$$
$$= \sup_{\|f\| \leq 1} [\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y)]$$
$$= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature mean difference

The MMD:

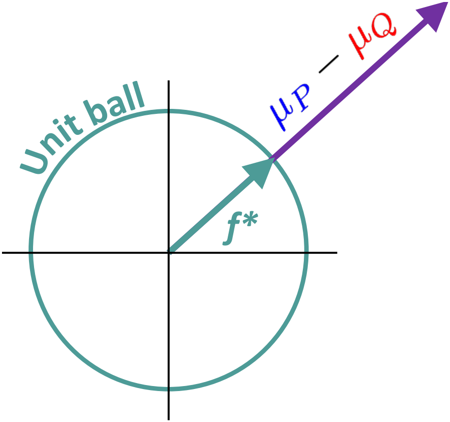
$$
\begin{aligned}
& MMD(P, Q; F) \\
&= \sup_{\|f\| \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right] \\
&= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



Integral prob. metric vs feature mean difference

The MMD:

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



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

Integral prob. metric vs feature mean difference

The MMD:

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

IPM view equivalent to feature mean difference (kernel case only)

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"

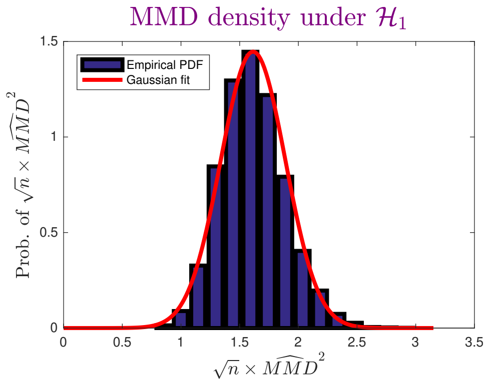
Want Threshold c_α for $\widehat{MMD}^2$ to get false positive rate α

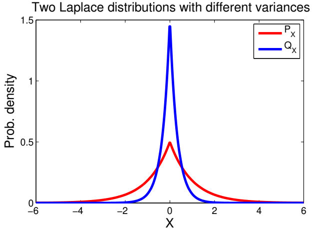
Asymptotics of $\widehat{MMD}^2$ when $P \neq Q$

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





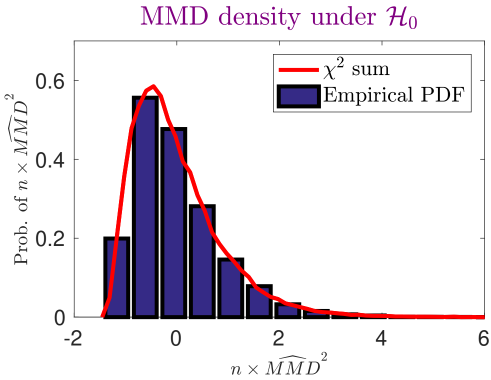
Behaviour of $\widehat{MMD}^2$ when $P = Q$

What happens when P and Q are the same?

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

Where $P = Q$, statistic has asymptotic distribution

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



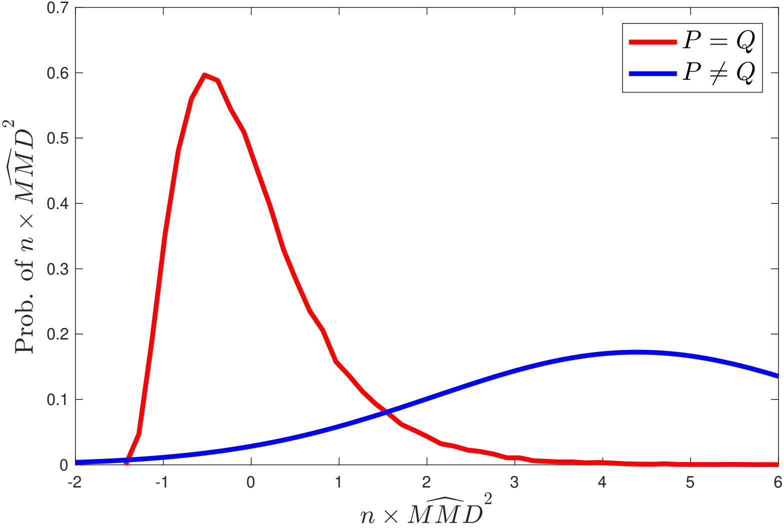
where

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

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

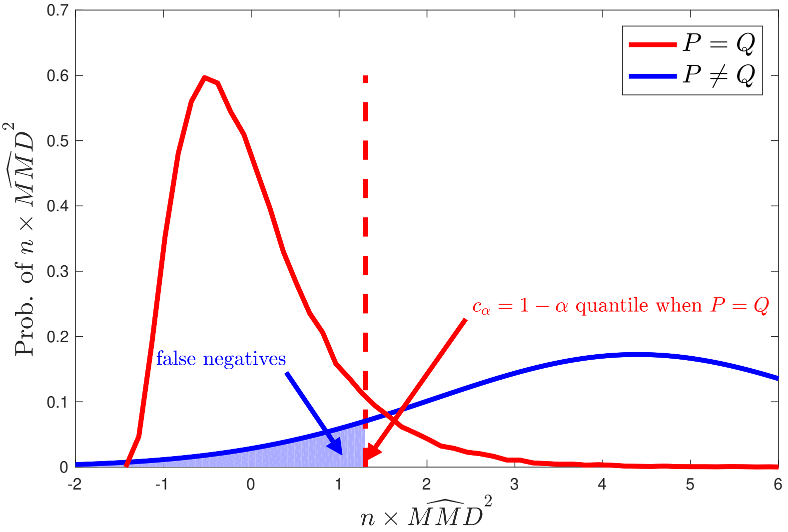
A statistical test

A summary of the asymptotics:



A statistical test

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



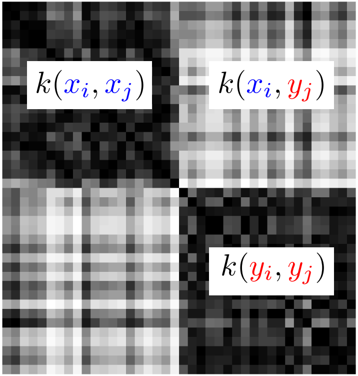
How do we get test threshold c_α?

Original empirical MMD for dogs and fish:

$X = [$  $\ldots]$

$Y = [$  $\ldots]$

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

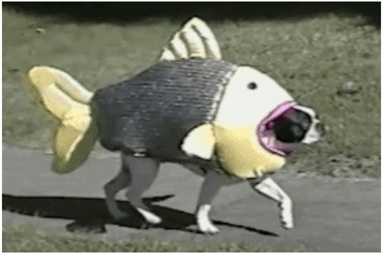


How do we get test threshold c_α?

Permuted dog and fish samples (merdogs):

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

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



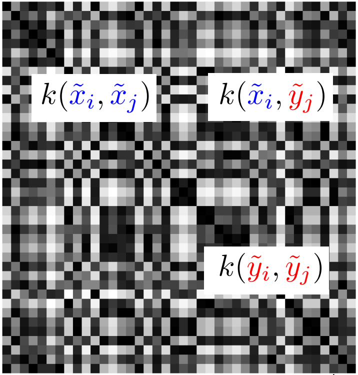
How do we get test threshold c_α?

Permuted dog and fish samples (merdogs):

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

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

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



Permutation simulates
$P = Q$

How do we get test threshold c_α?

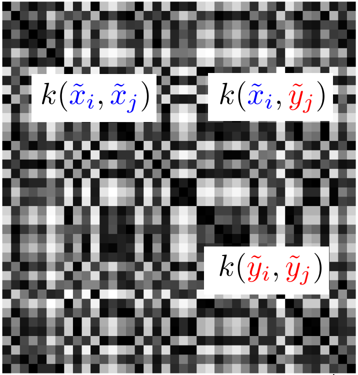
Permuted dog and fish samples (merdogs):

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

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

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



How to choose the best kernel: optimising the kernel parameters

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

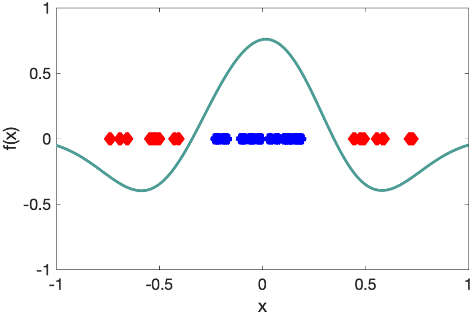
- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...

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...

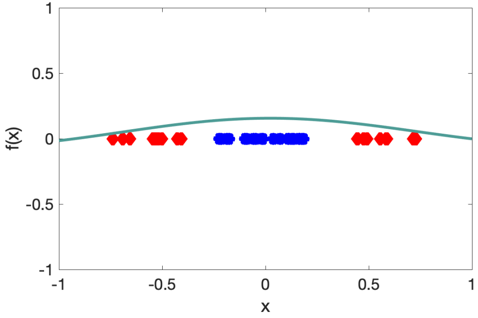


Choosing a kernel for the test

- 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...

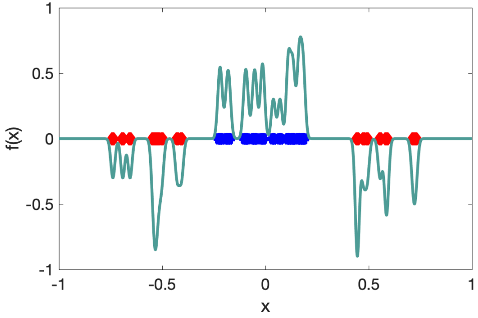


Choosing a kernel for the test

- 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

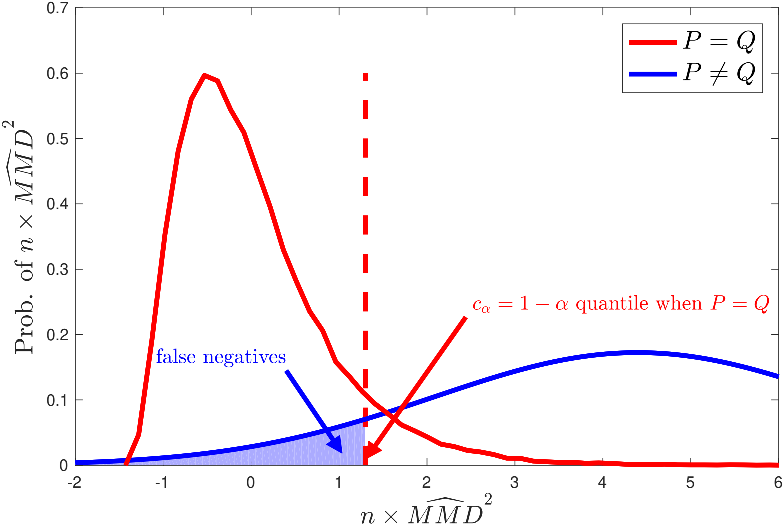
- Simple choice: exponentiated quadratic

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

- *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 $P \neq Q$):

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

Optimizing kernel for test power

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

$$
\begin{aligned}
&\mathrm{Pr}_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right) \\
&\to \Phi\left(\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} - \frac{c_\alpha}{n\sqrt{V_n(P,Q)}}\right)
\end{aligned}
$$

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 $P \neq Q$):

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

For large n, second term negligible!

Optimizing kernel for test power

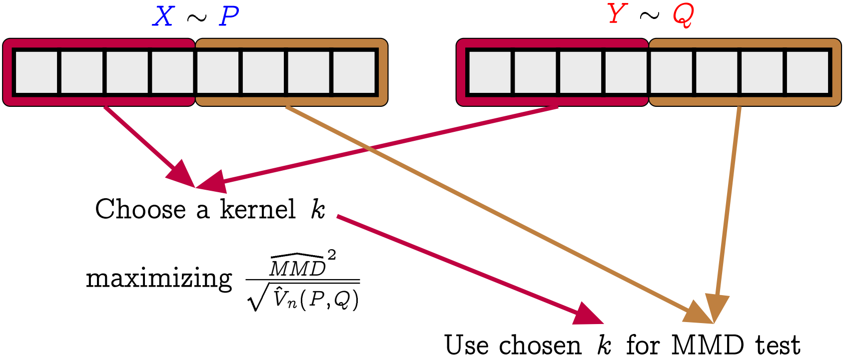
The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

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

To maximize test power, maximize

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

Data splitting

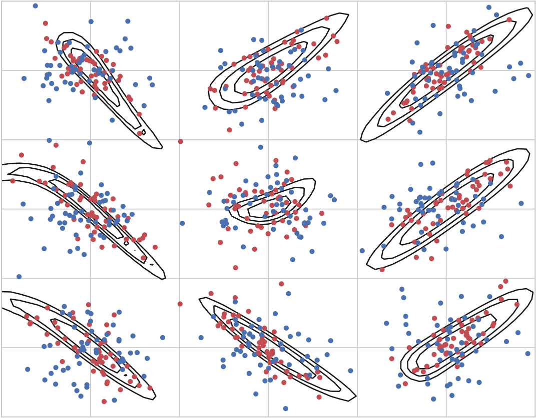


Learning a kernel helps a lot

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



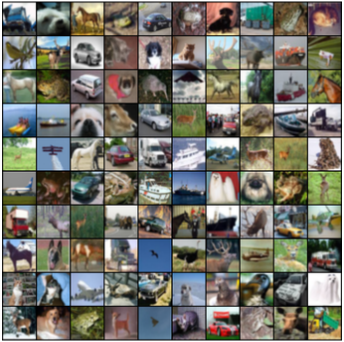
Learning a kernel helps a lot

Kernel with deep learned features:

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

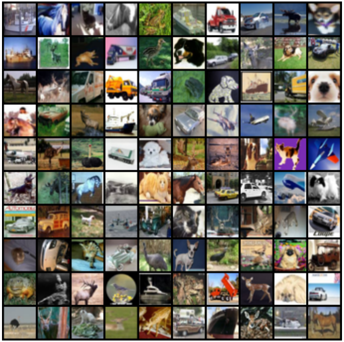
κ and q are Gaussian kernels

- 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]

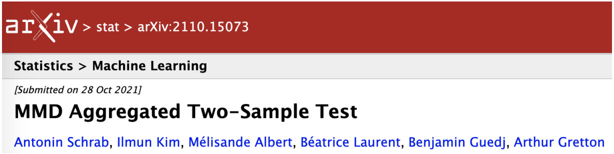
Learning Deep Kernels for Non-Parametric Two-Sample Tests

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

ICML 2020

Code: https://github.com/fengliu90/DK-for-TST

Adaptive testing without data splitting


arXiv > stat > arXiv:2110.15073

Statistics > Machine Learning

[Submitted on 28 Oct 2021]

MMD Aggregated Two-Sample Test

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

In revision, JMLR

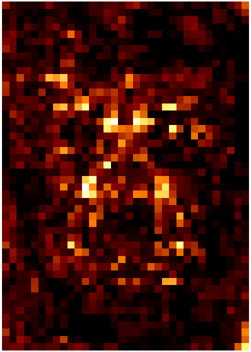
Code: https://github.com/antoninschrab/mmdagg-paper

Interpretable test features

From the two collections

$\{$ , ,  $, \ldots\}$ and $\{$ , ,  $, \ldots\}$,

produce a new point indicating where to look for the differences



Interpretable test features

From the two collections

{ , ,  , . . .} and { , ,  , . . .},

produce a new point indicating where to look for the differences

Interpretable test features

arXiv > stat > arXiv:1605.06796

Statistics > Machine Learning

[Submitted on 22 May 2016 (v1), last revised 28 Oct 2016 (this version, v2)]

Interpretable Distribution Features with Maximum Testing Power

Wittawat Jitkrittum, Zoltan Szabo, Kacper Chwialkowski, Arthur Gretton

NeurIPS 2016

Code: https://github.com/wittawatj/interpretable-test

Research support

Work supported by:

The Gatsby Charitable Foundation

Deepmind

Questions?

- A brief introduction to RKHS
- Maximum Mean Discrepancy (MMD)...
  - ...as a difference in feature means
  - ...as an integral probability metric (not just a technicality!)
- Statistical tests based on the MMD

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 $f^* := \begin{bmatrix} f_1^* & f_2^* & \ldots \end{bmatrix}$

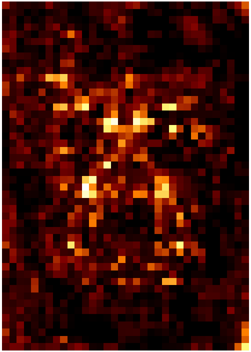
Interpretable test features

Overview

From the two collections

{ , ,  , . . . } and { , ,  , . . . },

produce a new point indicating where to look for the differences



Overview

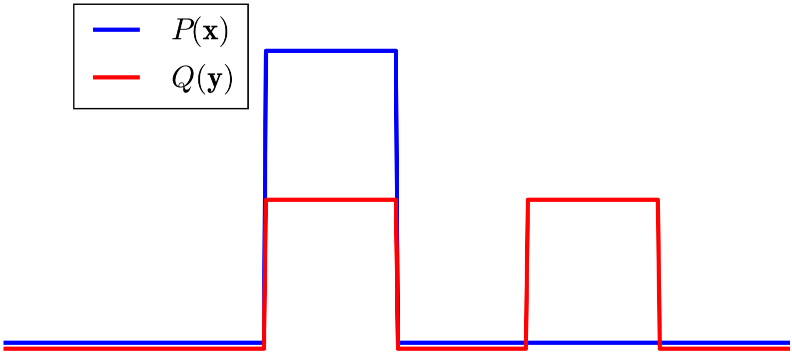
From the two collections

{ , ,  , . . .} and { , ,  , . . .},

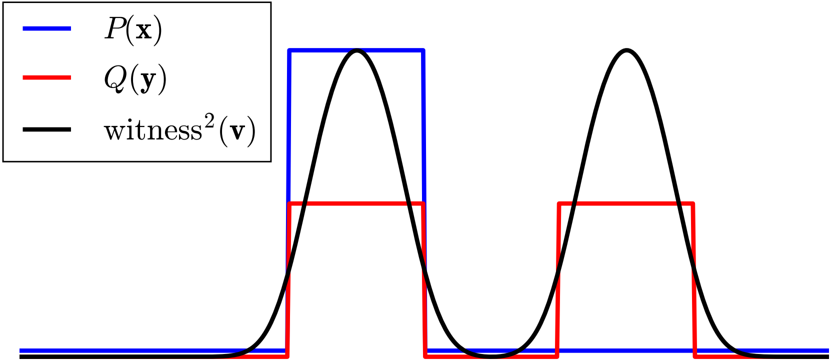
produce a new point indicating where to look for the differences

Distinguishing Feature(s)

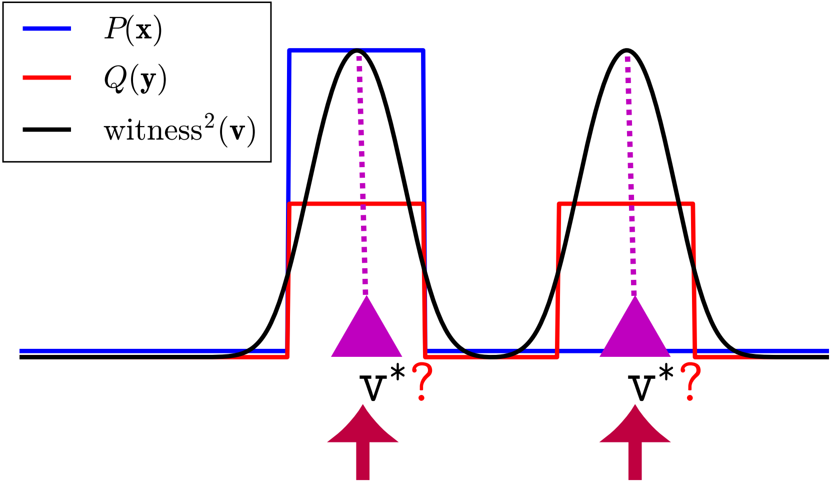
Where is the best location to observe the difference of $P(\mathbf{x})$ and $Q(\mathbf{y})$?



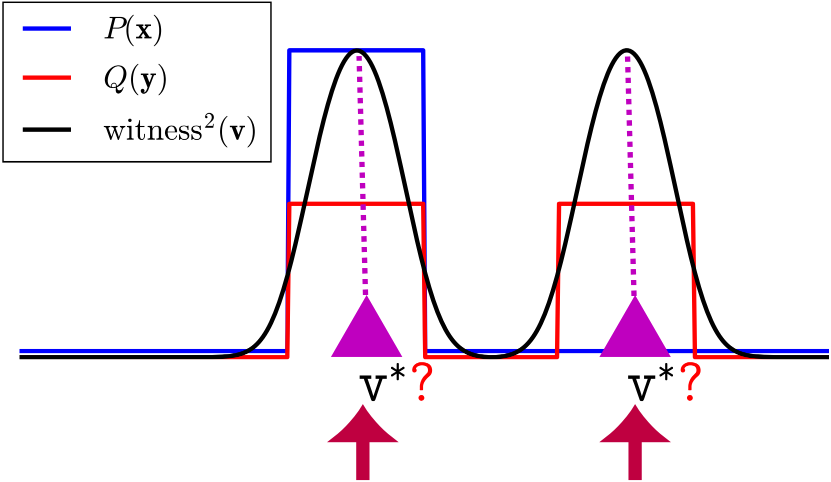
Maximum of the witness function?



Maximum of the witness function?



Maximum of the witness function?



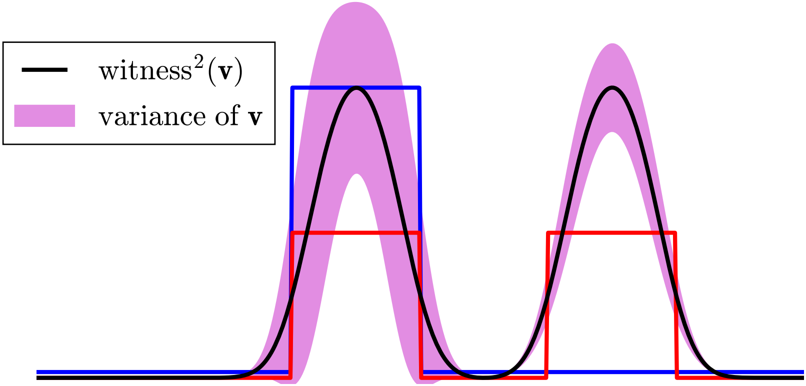
- $\text{witness}^2(\mathrm{v})$ only cares about the “signal”.
- Not the “noise” (variability) at each feature.

Signal-to-noise of witness function maximizes power

- Variance of v = variance of v from X + variance of v from Y.
- ME Statistic: $\hat{\lambda}_n(\mathrm{v}) := n \frac{\mathrm{witness}^2(\mathrm{v})}{\text{variance of v}}$.

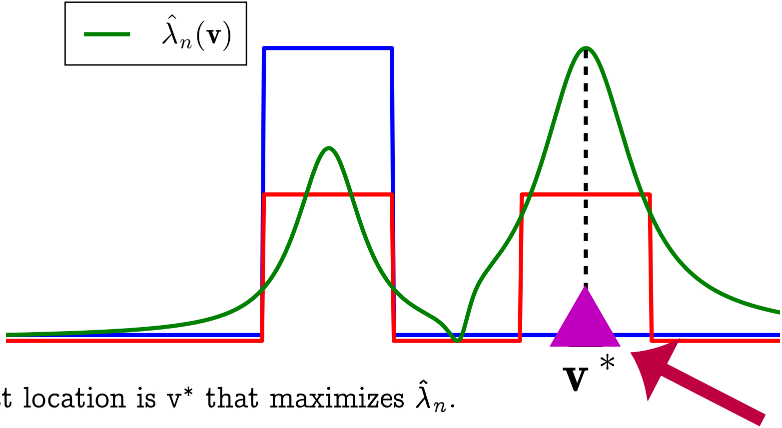
Signal-to-noise of witness function maximizes power

- Variance of v = variance of v from X + variance of v from Y.
- ME Statistic: $\hat{\lambda}_n(\mathrm{v}) := n \frac{\text{witness}^2(\mathrm{v})}{\text{variance of v}}$.



Signal-to-noise of witness function maximizes power

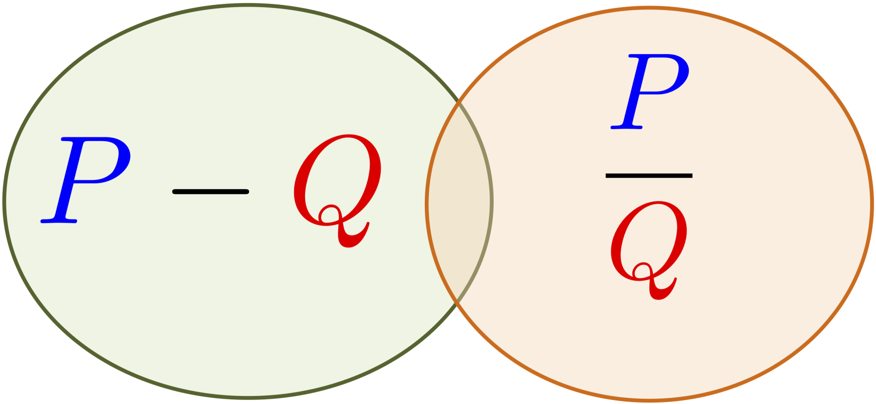
- Variance of v = variance of v from X + variance of v from Y.
- ME Statistic: $\hat{\lambda}_n(\mathrm{v}) := n \frac{\mathrm{witness}^2(\mathrm{v})}{\text{variance of v}}$.



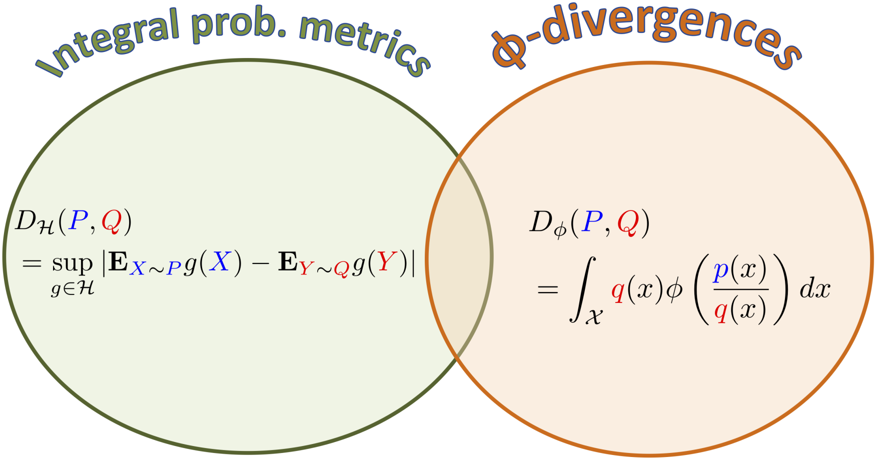
- Best location is v^* that maximizes $\hat{\lambda}_n$.

Divergence measures

Divergences



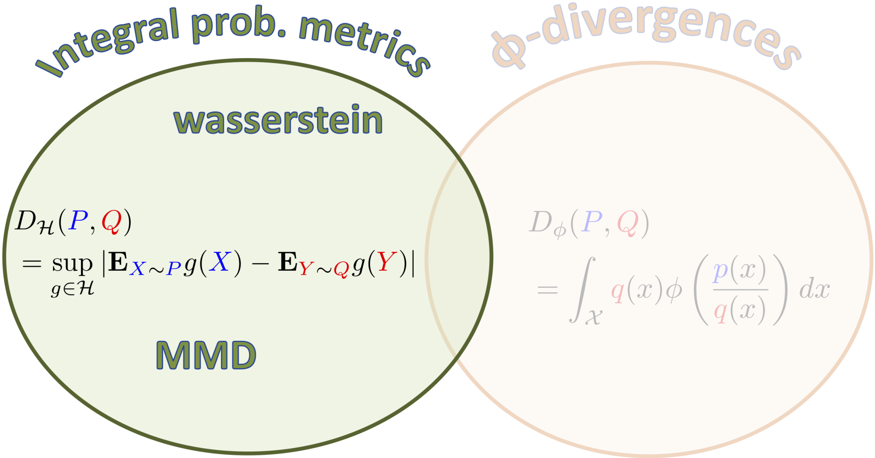
Divergences



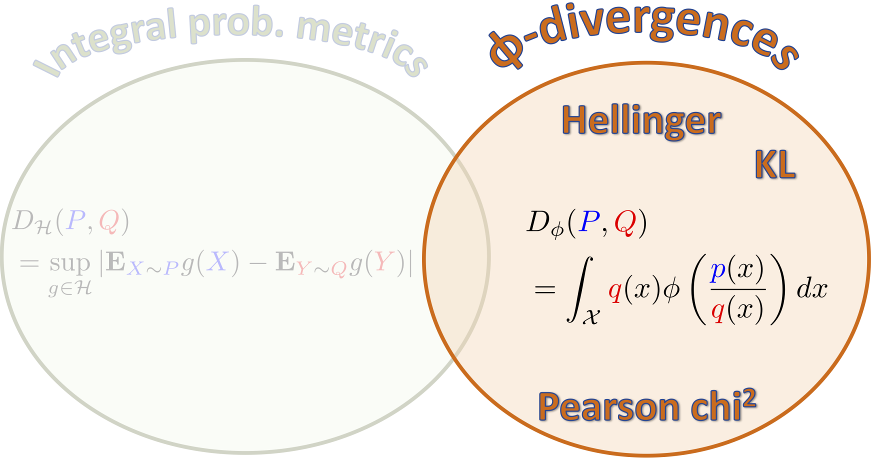
$$D_{\mathcal{H}}(P, Q) = \sup_{g \in \mathcal{H}} \left| \mathbf{E}_{X \sim P}\, g(X) - \mathbf{E}_{Y \sim Q}\, g(Y) \right|$$

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

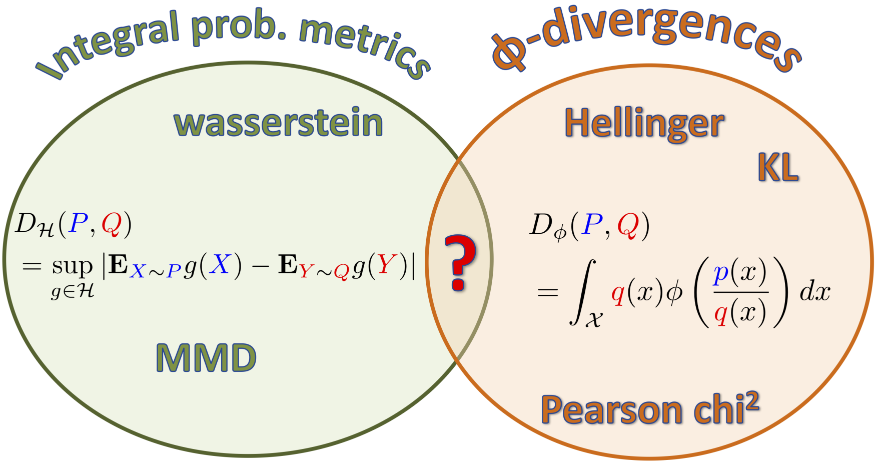
The integral probability metrics



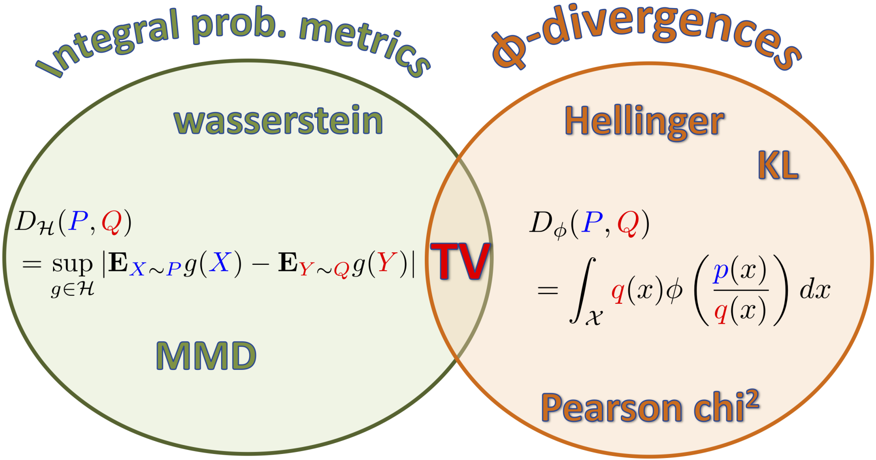
The ϕ-divergences



Divergences



Divergences



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