A motivation: comparing two samples

- Given: Samples from unknown distributions $P$ and $Q$.
- Goal: do $P$ and $Q$ differ?
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
A second task: dependence testing

- Given: Samples from a distribution $P_{XY}$

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Dog" /></td>
<td>A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.</td>
</tr>
<tr>
<td><img src="image2.png" alt="Dog" /></td>
<td>Their noses guide them through life, and they're never happier than when following an interesting scent.</td>
</tr>
<tr>
<td><img src="image3.png" alt="Cat" /></td>
<td>A responsive, interactive pet, one that will blow in your ear and follow you everywhere.</td>
</tr>
</tbody>
</table>

Text from dogtime.com and petfinder.com


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

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

![Two Gaussians with different variances](image-url)
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 \textit{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

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

Features: Gaussian Processes for Machine learning, Rasmussen and Williams, Ch. 4.
Given $P$ a Borel probability measure on $X$, define feature map of probability $P$,

$$\mu_P = [\ldots E_P [\varphi_i(X)] \ldots]$$

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

$$\langle \mu_P, \mu_Q \rangle_F = E_{P, Q} k(x, y)$$

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 \textit{distributions}

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

$$\mu_P = [\ldots E_P[\varphi_i(X)]\ldots]$$

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

$$\langle \mu_P, \mu_Q \rangle_F = E_{P, Q} k(x, y)$$

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

\textit{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:

\[ MMD^2(P, Q) = \| \mu_P - \mu_Q \|_F^2 \]
\[ = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_F \]
The maximum mean discrepancy

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

\[
MMD^2(P, Q) = \|\mu_P - \mu_Q\|^2_F = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_F = \langle \mu_P, \mu_P \rangle_F + \langle \mu_Q, \mu_Q \rangle_F - 2\langle \mu_P, \mu_Q \rangle_F
\]
The maximum mean discrepancy

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

\[
MMD^2(P, Q) = ||\mu_P - \mu_Q||_F^2 \\
= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_F \\
= E_P k(X, X') + E_Q k(Y, Y') - 2E_{P,Q} k(X, Y)
\]

(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:

\[
\overline{\text{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i\neq j} k(\text{dog}_i, \text{dog}_j) + \frac{1}{n(n-1)} \sum_{i\neq j} k(\text{fish}_i, \text{fish}_j) \\
- \frac{2}{n^2} \sum_{i,j} k(\text{dog}_i, \text{fish}_j)
\]
MMD as an integral probability metric

Integral probability metric:
Find a "well behaved function" $f(x)$ to maximize

$$\mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)$$
MMD as an integral probability metric

Integral probability metric:
Find a "well behaved function" $f(x)$ to maximize

$$E_P f(X) - E_Q f(Y)$$
MMD as an integral probability metric

Maximum mean discrepancy: smooth function for $P$ vs $Q$

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

($F$ = unit ball in RKHS $\mathcal{F}$)
MMD as an integral probability metric

Maximum mean discrepancy: smooth function for $P$ vs $Q$

$$MMD(P, Q; F) := \sup_{\|f\|_F \leq 1} [E_P f(X) - E_Q f(Y)]$$

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

Functions are linear combinations of features:

$$f(x) = \langle f, \varphi(x) \rangle_F = \sum_{\ell=1}^{\infty} f_{\ell} \varphi_{\ell}(x) = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{bmatrix}^T \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \varphi_3(x) \\ \vdots \end{bmatrix}$$

$$\|f\|_F^2 := \sum_{i=1}^{\infty} f_i^2 \leq 1$$
MMD as an integral probability metric

Maximum mean discrepancy: smooth function for $P$ vs $Q$

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

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

Witness $f$ for Gauss and Laplace densities

![Graph showing the probability density and $f$ for Gauss and Laplace densities](image)
MMD as an integral probability metric

Maximum mean discrepancy: smooth function for $P$ vs $Q$

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

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

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

Other choices for witness function class:

- Bounded continuous [Dudley, 2002]
- Bounded variation 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 $Q$

\[ \text{MMD}(P, Q; F) := \sup_{\|f\|_F \leq 1} [E_P f(X) - E_Q f(Y)] \]

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

Expectations of functions are linear combinations of expected features

\[ E_P(f(X)) = \langle f, E_P \varphi(X) \rangle_\mathcal{F} = \langle f, \mu_P \rangle_\mathcal{F} \]

(always true if kernel is bounded)
The MMD:

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

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

use

\[
\mathbb{E}_P f(X) = \langle \mu_P, f \rangle_F
\]

\[
\mathbb{E}_Q f(Y) = \langle \mu_Q, f \rangle_F
\]

\[
= \sup_{f \in F, \|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_F
\]
The MMD:

\[ MMD(P, Q; F) \]
\[ = \sup_{\|f\| \leq 1} \left[ \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \right] \]
\[ = \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \]
Integral prob. metric vs feature mean difference

The MMD:

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

\[
= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_F
\]
The MMD:

\[ MMD(P, Q; F) = \sup_{\|f\| \leq 1} \left| E_P f(X) - E_Q f(Y) \right| \]

\[ = \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_F \]

\[ f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|} \]
Integral prob. metric vs feature mean difference

The MMD:

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

\[ = \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \]

\[ = \|\mu_P - \mu_Q\|_{\mathcal{F}} \]

IPM view equivalent to feature mean difference (kernel case only)
Two-Sample Testing with MMD
A statistical test using MMD

The empirical MMD:

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

How does this help decide whether \( P = Q \)?
A statistical test using MMD

The empirical MMD:

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

Perspective from statistical hypothesis testing:

- Null hypothesis \( \mathcal{H}_0 \) when \( P = Q \)
  - should see \( \hat{\text{MMD}}^2 \) “close to zero”.
- Alternative hypothesis \( \mathcal{H}_1 \) when \( P \neq Q \)
  - should see \( \hat{\text{MMD}}^2 \) “far from zero”
A statistical test using MMD

The empirical MMD:

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

Perspective from statistical hypothesis testing:

- **Null hypothesis** \( \mathcal{H}_0 \) when \( P = Q \)
  - should see \( \widehat{\text{MMD}}^2 \) “close to zero”.
- **Alternative hypothesis** \( \mathcal{H}_1 \) when \( P \neq Q \)
  - should see \( \widehat{\text{MMD}}^2 \) “far from zero”

Want **Threshold** \( c_\alpha \) for \( \widehat{\text{MMD}}^2 \) to get **false positive rate** \( \alpha \)
Asymptotics of $\widehat{\text{MMD}}^2$ when $P \neq Q$

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

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

where variance $V_n(P, Q) = O(n^{-1})$.
Behaviour of $\overline{\text{MMD}}^2$ when $P = Q$

What happens when $P$ and $Q$ are the same?
Asymptotics of $\widehat{\text{MMD}}^2$ 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

$$\lambda_l \psi_i(x') = \int_{\mathcal{X}} \tilde{k}(x, x') \psi_i(x') dP(x)$$

is centred

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

A summary of the asymptotics:

\begin{align*}
\text{Prob. of } n \times \hat{MMD}^2 \quad &\text{ vs } \quad n \times \hat{MMD}^2 \\
\end{align*}

- Red line: \(P = Q\)
- Blue line: \(P \neq Q\)
A statistical test

Test construction: (G., Borgwardt, Rasch, Schoelkopf, and Smola, JMLR 2012)
How do we get test threshold $c_\alpha$?

Original empirical MMD for dogs and fish:

$$
X = \begin{bmatrix}
\text{[\text{dogs}]} & \text{[\text{dogs}]} & \text{[\text{dogs}]} & \ldots
\end{bmatrix}
$$

$$
Y = \begin{bmatrix}
\text{[\text{fishes}]} & \text{[\text{fishes}]} & \text{[\text{fishes}]} & \ldots
\end{bmatrix}
$$

$$
\hat{\text{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j)
$$

$$
+ \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j)
$$

$$
- \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)
$$
How do we get test threshold $c_\alpha$?

Permuted dog and fish samples (merdogs):

$$\tilde{X} = [\text{fish} \quad \text{dog} \quad \text{fish} \quad \ldots]$$

$$\tilde{Y} = [\text{dog} \quad \text{fish} \quad \text{dog} \quad \ldots]$$
How do we get test threshold $c_\alpha$?

Permuted dog and fish samples (merdogs):

$$\tilde{X} = [\text{fish}, \text{dog}, \text{dog}, \ldots]$$

$$\tilde{Y} = [\text{dog}, \text{fish}, \text{dog}, \ldots]$$

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

$$+ \frac{1}{n(n-1)} \sum_{i \neq j} k(\tilde{y}_i, \tilde{y}_j)$$

$$- \frac{2}{n^2} \sum_{i,j} k(\tilde{x}_i, \tilde{y}_j)$$

Permutation simulates $P = Q$
How do we get test threshold $c_\alpha$?

Permuted dog and fish samples (merdogs):

$$\tilde{X} = [\text{dog, fish, ...}]$$

$$\tilde{Y} = [\text{dog, fish, ...}]$$

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

Proposition 1, Schrab, Kim, Albert, Laurens, 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 \( \sigma \): for any \( P \) and \( Q \), power \( \rightarrow 1 \) as \( n \rightarrow \infty \)
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 \( \sigma \): for any \( P \) and \( Q \), power \( \to 1 \) as \( n \to \infty \)
- But choice of \( \sigma \) 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 \( \sigma \): for any \( P \) and \( Q \), power \( \to 1 \) as \( n \to \infty \)
- But choice of \( \sigma \) 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 \( \sigma \): for any \( P \) and \( Q \), power \( \to 1 \) as \( n \to \infty \)
- But choice of \( \sigma \) 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 $\sigma$: for any $P$ and $Q$, power $\to 1$ as $n \to \infty$
- But choice of $\sigma$ 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 \( \sigma \): for any \( P \) and \( Q \), power \( \rightarrow 1 \) as \( n \rightarrow \infty \)
- But choice of \( \sigma \) is very important for finite \( n \)...
- ...and some problems (e.g. images) might have no good choice for \( \sigma \)
Graphical illustration

- Maximising test power same as minimizing false negatives

\[ c_\alpha = 1 - \alpha \] quantile when \( P = Q \)
Optimizing kernel for test power

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

$$\Pr_1 \left( n\hat{\text{MMD}}^2 > \hat{c}_\alpha \right)$$
Optimizing kernel for test power

The power of our test (Pr$_1$ denotes probability under $P \neq Q$):

$$
\text{Pr}_1 \left( n \text{MMD}^2 \geq \hat{c}_\alpha \right) \\
\rightarrow \Phi \left( \frac{\text{MMD}^2(P, Q)}{\sqrt{V_n(P, Q)}} - \frac{c_\alpha}{n\sqrt{V_n(P, Q)}} \right)
$$

where

- $\Phi$ is the CDF of the standard normal distribution.
- $\hat{c}_\alpha$ is an estimate of $c_\alpha$ test threshold.
Optimizing kernel for test power

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

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

For large $n$, second term negligible!
Optimizing kernel for test power

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

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

$$\rightarrow \Phi \left( \frac{\text{MMD}^2(P, Q)}{\sqrt{V_n(P, Q)}} - \frac{c_\alpha}{n \sqrt{V_n(P, Q)}} \right)$$

To maximize test power, maximize

$$\frac{\text{MMD}^2(P, Q)}{\sqrt{V_n(P, Q)}}$$
Data splitting

Choose a kernel $k$ maximizing

$$\frac{\text{MMD}^2}{\sqrt{\hat{V}_n(P,Q)}}$$

Use chosen $k$ for MMD test
Learning a kernel helps a lot

Kernel with deep learned features:

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

\( \kappa \) and \( q \) are Gaussian kernels

CIFAR-10 vs CIFAR-10.1, null rejected 75% of time
Learning a kernel helps a lot

Kernel with deep learned features:

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

\( \kappa \) 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:
\[ k_\theta(x, y) = \left[ (1 - \epsilon) \kappa(\Phi_\theta(x), \Phi_\theta(y)) + \epsilon \right] q(x, y) \]
\( \kappa \) and \( q \) are Gaussian kernels

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

ICML 2020

Code: https://github.com/fengliu90/DK-for-TST
Adaptive testing without data splitting

In revision, JMLR

Code: https://github.com/antoninschrab/mmdagg-paper
Interpretable test features

From the two collections:

\[
\{ \text{\begin{subfigure}{0.1\textwidth}
\end{subfigure}, \begin{subfigure}{0.1\textwidth}
\end{subfigure}, \ldots \} \text{ and } \{ \begin{subfigure}{0.1\textwidth}
\end{subfigure}, \begin{subfigure}{0.1\textwidth}
\end{subfigure}, \begin{subfigure}{0.1\textwidth}
\end{subfigure}, \ldots \} \}
\]

produce a new point indicating where to look for the differences.
Interpretable test features

From the two collections

\{ \text{[Images of faces]} \} \text{ and } \{ \text{[Images of faces]} \},

produce a new point indicating where to look for the differences
Interpretable test features

Interpretable Distribution Features with Maximum Testing Power

Wittawat Jitkrittum, Zoltan Szabo, Kacper Chwialkowski, Arthur Gretton

NeurIPS 2016

Code: https://github.com/wittawatj/interpretable-test
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 \( P \)

\[ \hat{\mu}_P := \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \]
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 \( \nu \)

\[ f^*(\nu) = \langle f^*, \varphi(\nu) \rangle_{\mathcal{F}} \]
Derivation of empirical witness function

Recall the witness function expression

\[ f^* \propto \mu_P - \mu_Q \]

The empirical feature mean for \( P \)

\[ \hat{\mu}_P := \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \]

The empirical witness function at \( v \)

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

\[ \propto \langle \hat{\mu}_P - \hat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \]
Derivation of empirical witness function

Recall the witness function expression

\[ f^* \propto \mu_P - \mu_Q \]

The empirical feature mean for \( P \)

\[ \hat{\mu}_P := \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \]

The empirical witness function at \( v \)

\[
\begin{align*}
    f^*(v) &= \langle f^*, \varphi(v) \rangle_F \\
    &\propto \langle \hat{\mu}_P - \hat{\mu}_Q, \varphi(v) \rangle_F \\
    &= \frac{1}{n} \sum_{i=1}^{n} k(x_i, v) - \frac{1}{n} \sum_{i=1}^{n} k(y_i, v)
\end{align*}
\]

Don’t need explicit feature coefficients \( f^* := [f_1^*, f_2^*, \ldots] \)
Interpretable test features
Overview

From the two collections

\{ \text{ }, \text{ }, \text{ } \text{ }, \ldots \} \text{ and } \{ \text{ }, \text{ }, \text{ } \text{ }, \ldots \},

produce a new point indicating where to look for the differences.
Overview

From the two collections

\[
\{ \text{female}, \text{male}, \ldots \} \quad \text{and} \quad \{ \text{angry}, \text{afraid}, \ldots \},
\]

produce a new point indicating where to look for the differences.
Distinguishing Feature(s)

Where is the best location to observe the difference of $P(x)$ and $Q(y)$?
Maximum of the witness function?

$P(x)$

$Q(y)$

$\text{witness}^2(v)$
Maximum of the witness function?

\[ P(x) \]
\[ Q(y) \]
\[ \text{witness}^2(v) \]
Maximum of the witness function?

- $P(x)$
- $Q(y)$
- witness$^2(v)$

- witness$^2(v)$ only cares about the “signal”.
- Not the “noise” (variability) at each feature.
Signal-to-noise of witness function maximizes power

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

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

- Variance of \( v \) = variance of \( v \) from \( X \) + variance of \( v \) from \( Y \).

- ME Statistic: \( \hat{\lambda}_n(v) := n \frac{\text{witness}^2(v)}{\text{variance of } v} \).

- Best location is \( v^* \) that maximizes \( \hat{\lambda}_n \).
Divergence measures
Divergences

\[ P \quad Q \]

\[ \frac{P}{Q} \]
Divergences

Integral prob. metrics

\( D_\mathcal{H}(P, Q) \)
\[
= \sup_{g \in \mathcal{H}} |\mathbb{E}_{X \sim P} g(X) - \mathbb{E}_{Y \sim Q} g(Y)|
\]

\( \phi \)-divergences

\( D_\phi(P, Q) \)
\[
= \int_X q(x) \phi \left( \frac{p(x)}{q(x)} \right) dx
\]
The integral probability metrics

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

\[ D_\phi(P, Q) = \int_X q(x) \phi \left( \frac{p(x)}{q(x)} \right) dx \]
The $\phi$-divergences

Integral prob. metrics

$$D_{\mathcal{H}}(P, Q) = \sup_{g \in \mathcal{H}} |E_{X \sim P} g(X) - E_{Y \sim Q} g(Y)|$$

$\phi$-divergences

Hellinger

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

KL

Pearson chi$^2$
Divergences

**Integral prob. metrics**

\[ D_{\mathcal{H}}(P, Q) = \sup_{g \in \mathcal{H}} |E_{X \sim P} g(X) - E_{Y \sim Q} g(Y)| \]

**MMD**

\[ D_{\phi}(P, Q) = \int_X q(x) \phi \left( \frac{p(x)}{q(x)} \right) dx \]

**\(\phi\)-divergences**

- **Hellinger**
- **KL**
- **Pearson chi\(^2\)**
Divergences

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

\[ D_{\phi}(P, Q) \]

\[ = \int g(x) \phi \left( \frac{p(x)}{q(x)} \right) dx \]

Integral prob. metrics

\text{wasserstein}

MMD

\text{TV}

\text{φ-divergences}

\text{Hellinger}

KL

\text{Pearson chi}^2

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