Comparing two samples

Arthur Gretton

Gatsby Computational Neuroscience Unit,
Deepmind

Columbia Statistics, 2023
Comparing two samples

- Given: Samples from unknown distributions \( P \) and \( Q \).
- Goal: do \( P \) and \( Q \) differ?
A real-life example: two-sample tests

The problem: Do local field potential (LFP) signals change when measured near a spike burst?
A real-life example: two-sample tests

The problem: Do local field potential (LFP) signals change when measured near a spike burst?
A real-life example: two-sample tests

■ The problem: Do local field potential (LFP) signals change when measured near a spike burst?
A real-life example: two-sample tests

Goal: do $P$ and $Q$ differ?

CIFAR 10 samples

Cifar 10.1 samples

Significant difference?

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

Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017.
A real-life example: discrete domains

How do you compare distributions in a discrete domain?

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

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

\[ P_X = Q_Y \]

\( Y_1 \): Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

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

\[ \ldots \]

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

Two sample testing

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

- Statistical testing with the MMD

- “How to choose the best kernel”
  - when are feature means unique?
  - what kernel gives the most powerful test?
Maximum Mean Discrepancy
Feature mean difference

- Simple example: 2 Gaussians with different means
- Answer: $t$-test
Feature mean difference

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

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

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

Kernels: dot products of features

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

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

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

$$\varphi(x) = [\ldots \varphi_i(x) \ldots] \in l_2$$

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.
Infinitely many features of distributions

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

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

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

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

for $x \sim P$ and $y \sim Q$. 
Given $P$ a Borel probability measure on $\mathcal{X}$, define feature map of probability $P$,

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

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

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

for $x \sim P$ and $y \sim Q$. 
Expectations of RKHS functions

**Function evaluation in an RKHS:**

\[ f(x) = \langle f, \varphi_x \rangle_F \]

**Expectation evaluation in an RKHS:**

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

\[ \mu_P \text{ gives you expectations of all RKHS functions} \]

**Empirical mean embedding:**

\[ \hat{\mu}_P = \frac{1}{m} \sum_{i=1}^{m} \varphi(x_i) \quad x_i \overset{\text{i.i.d.}}{\sim} P \]

... does this reasoning work in infinite dimensions?
Expectations of RKHS functions

Function evaluation in an RKHS:

\[ f(x) = \langle f, \varphi_x \rangle_F \]

Expectation evaluation in an RKHS:

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

\( \mu_P \) gives you expectations of all RKHS functions

Empirical mean embedding:

\( \hat{\mu}_P = \frac{1}{m} \sum_{i=1}^{m} \varphi(x_i) \quad x_i \sim P \quad \text{i.i.d.} \)

... does this reasoning work in infinite dimensions?
Does the feature space mean exist?

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

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

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

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

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

$$A f = \langle f(\cdot), g_A(\cdot) \rangle_\mathcal{F}$$
Does the feature space mean exist?

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

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

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

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

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

$$Af = \langle f(\cdot), g_A(\cdot) \rangle_{\mathcal{F}}$$
Does the feature space mean exist?

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

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

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

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

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

$$Af = \langle f(\cdot), g_A(\cdot) \rangle_{\mathcal{F}}$$
Does the feature space mean exist?

Existence of mean embedding: If $E_P \sqrt{k(x, x)} = E_P \|\varphi(x)\|_F < \infty$ then $\exists \mu_P \in \mathcal{F}$. 
Does the feature space mean exist?

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

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

$$|T_P f| = |E_P f(x)|.$$

$$\leq E_P |f(x)|$$
Does the feature space mean exist?

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

Proof:

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

$$|T_P f| = |E_P f(x)| \leq E_P |f(x)| = E_P |\langle f, \varphi(x) \rangle_\mathcal{F}|$$
Does the feature space mean exist?

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

Proof:

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

$$|T_P f| = |\mathbb{E}_P f(x)| \leq \mathbb{E}_P |f(x)| = \mathbb{E}_P |\langle f, \varphi(x) \rangle_\mathcal{F}| \leq \mathbb{E}_P \left( \sqrt{k(x, x)} \|f\|_\mathcal{F} \right)$$
Does the feature space mean exist?

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

Proof:

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

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

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

$$T_P f = \langle f, \mu_P \rangle_\mathcal{F}.$$
\( \mu_P \) as a function in the RKHS

Embedding of \( P \) to feature space

- Mean embedding \( \mu_P \in \mathcal{F} \),

\[
\langle \mu_P, f \rangle_{\mathcal{F}} = \mathbb{E}_P f(x).
\]

- What does prob. feature map look like?

\[
\mu_P(t) = \langle \mu_P, \varphi(t) \rangle_{\mathcal{F}} \\
= \langle \mu_P, k(\cdot, t) \rangle_{\mathcal{F}} \\
= \mathbb{E}_{x \sim P} k(x, t)
\]

Expectation of kernel!
$\mu_P$ as a function in the RKHS

Embedding of $P$ to feature space

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

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

- What does prob. feature map look like?

$$\mu_P(t) = \langle \mu_P, \varphi(t) \rangle_{\mathcal{F}}$$

$$= \langle \mu_P, k(\cdot, t) \rangle_{\mathcal{F}}$$

$$= E_{x \sim P} k(x, t)$$

Expectation of kernel!
\( \mu_P \) as a function in the RKHS

Embedding of \( P \) to feature space

- Mean embedding \( \mu_P \in \mathcal{F} \),

\[
\langle \mu_P, f \rangle_{\mathcal{F}} = \mathbb{E}_P f(x).
\]

- What does prob. feature map look like?

\[
\mu_P(t) = \langle \mu_P, \varphi(t) \rangle_{\mathcal{F}}
\]

\[
= \langle \mu_P, k(\cdot, t) \rangle_{\mathcal{F}}
\]

\[
= \mathbb{E}_{x \sim P} k(x, t)
\]

Expectation of kernel!
The maximum mean discrepancy is the distance between feature means:

\[ MMD^2(P, Q) = \|\mu_P - \mu_Q\|^2 \]

\[ = E_P k(x, x') + E_Q k(y, y') - 2E_{P,Q} k(x, y) \]

(a) within distrib. similarity, (b) cross-distrib. similarity.
The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

\[ MMD^2(P, Q) = \| \mu_P - \mu_Q \|^2_F \]
\[ = \underbrace{E_P k(x, x')}_{(a)} + \underbrace{E_Q k(y, y')}_{(a)} - 2E_{P,Q} k(x, y) \]

Proof:

\[ \| \mu_P - \mu_Q \|^2_F = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_F \]
\[ = \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2 \langle \mu_P, \mu_Q \rangle \]
The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

\[ \text{MMD}^2(P, Q) = \| \mu_P - \mu_Q \|^2_F \]

\[ = \mathbb{E}_P k(x, x') + \mathbb{E}_Q k(y, y') - 2 \mathbb{E}_{P, Q} k(x, y) \]

(a) \hspace{1cm} (a) \hspace{1cm} (b)

Proof:

\[ \| \mu_P - \mu_Q \|^2_F = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_F \]

\[ = \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2 \langle \mu_P, \mu_Q \rangle \]
The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

\[
MMD^2(P, Q) = \|\mu_P - \mu_Q\|^2_\mathcal{F}
= \mathbb{E}_P k(x, x') + \mathbb{E}_Q k(y, y') - 2\mathbb{E}_{P, Q} k(x, y)
\]

Proof:

\[
\|\mu_P - \mu_Q\|^2_\mathcal{F} = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_\mathcal{F}
= \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2\langle \mu_P, \mu_Q \rangle
= \mathbb{E}_P \mu_P(x) + \ldots
\]
The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

\[
MMD^2(P, Q) = \|\mu_P - \mu_Q\|_F^2 \\
= \mathbb{E}_P k(x, x') + \mathbb{E}_Q k(y, y') - 2\mathbb{E}_{P, Q} k(x, y)
\]

(a) \hspace{2cm} (a) \hspace{2cm} (b)

Proof:

\[
\|\mu_P - \mu_Q\|_F^2 = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_F \\
= \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2\langle \mu_P, \mu_Q \rangle \\
= \mathbb{E}_P [\mu_P(x)] + \ldots \\
= \mathbb{E}_P \langle \mu_P, k(x, \cdot) \rangle + \ldots
\]
The maximum mean discrepancy

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

\[ MMD^2(P, Q) = \|\mu_P - \mu_Q\|_F^2 \]

\[ = \mathbb{E}_P k(x, x') + \mathbb{E}_Q k(y, y') - 2\mathbb{E}_{P,Q} k(x, y) \]

**Proof:**

\[ \|\mu_P - \mu_Q\|_F^2 = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_F \]

\[ = \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2\langle \mu_P, \mu_Q \rangle \]

\[ = \mathbb{E}_P[\mu_P(x)] + \ldots \]

\[ = \mathbb{E}_P \langle \mu_P, k(x, \cdot) \rangle + \ldots \]

\[ = \mathbb{E}_P k(x, x') + \mathbb{E}_Q k(y, y') - 2\mathbb{E}_{P,Q} k(x, y) \]
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:

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

$$-\frac{2}{n^2} \sum_{i,j} k(\text{dog}_i, \text{fish}_j)$$
MMD as an integral probability metric

Are $P$ and $Q$ different?
MMD as an integral probability metric

Are $P$ and $Q$ different?

![Samples from P and Q](image)
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

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

What if the function is not smooth?

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

What if the function is not smooth?

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

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

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

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

![Witness f for Gauss and Laplace densities](image)

- **f**
- **Gauss**
- **Laplace**
MMD as an integral probability metric

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

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

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

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

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

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

A reminder for the proof on the next slide:

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

(always true if kernel is bounded)
The MMD:

$$MMD(P, Q; F) = \sup_{\|f\|_{\mathcal{F}} \leq 1} [E_P f(X) - E_Q f(Y)]$$
Integral prob. metric vs feature difference

The MMD:

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

use

\[
E_P f(X) = \langle \mu_P, f \rangle_F
\]

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

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

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

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

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

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

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

The MMD:

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

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

\[
= \|\mu_P - \mu_Q\|
\]

Function view and feature view equivalent
Construction of MMD witness

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

Observe $X = \{x_1, \ldots, x_n\} \sim P$

Observe $Y = \{y_1, \ldots, y_n\} \sim Q$
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!)

\[ \text{witness}(v) \]
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 \( P \)

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

The empirical witness function at \( v \)

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

\[ f^*(v) = \langle f^*, \varphi(v) \rangle_F \]

\[ \propto \langle \hat{\mu}_P - \hat{\mu}_Q, \varphi(v) \rangle_F \]

\[ = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v) - \frac{1}{n} \sum_{i=1}^{n} k(y_i, v) \]

Don’t need explicit feature coefficients \( f^* := \begin{bmatrix} f_1^* & f_2^* & \ldots \end{bmatrix} \)
Interlude: divergence measures
Divergences
Divergences

**Integral prob. metrics**

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

**F-divergences**

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

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

\[ D_f(P, Q) = \int_X q(x) f \left( \frac{p(x)}{q(x)} \right) dx \]
**Divergences**

**Integral prob. metrics**

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

- MMD

**F-divergences**

- Hellinger
  \[ D_f(P, Q) = \int_X q(x) f \left( \frac{p(x)}{q(x)} \right) dx \]

- KL

- Pearson chi²
Divergences

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

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

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

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

$$\quad - \frac{2}{n^2} \sum_{i, j} k(x_i, y_j)$$

Perspective from statistical hypothesis testing:

- **Null hypothesis** $\mathcal{H}_0$ when $P = Q$
  - should see $\hat{MMD}^2$ “close to zero”.
- **Alternative hypothesis** $\mathcal{H}_1$ when $P \neq Q$
  - should see $\hat{MMD}^2$ “far from zero”
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 $H_0$ when $P = Q$
  - should see $\hat{\text{MMD}}^2$ “close to zero”.
- Alternative hypothesis $H_1$ when $P \neq Q$
  - should see $\hat{\text{MMD}}^2$ “far from zero”

Want Threshold $c_\alpha$ for $\hat{\text{MMD}}^2$ to get false positive rate $\alpha$
Behaviour of $\hat{MMD}^2$ when $P \neq Q$

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

- Laplace with different y-variance.

- $\sqrt{n} \times \hat{MMD}^2 = 1.2$
Behaviour of $\hat{\text{MMD}}^2$ when $P \neq Q$

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

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

Number of MMDs: 1
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

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

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

![Graph showing the number of MMDs: 2]
Behaviour of $\sqrt{n} \times \tilde{MMD}^2$ when $P \neq Q$

Repeat this 150 times …

Number of MMDs: 150
Behaviour of $\hat{MMD}^2$ when $P \neq Q$

Repeat this 300 times …

Number of MMDs: 300
Behaviour of $\hat{MMD}^2$ when $P \neq Q$

Repeat this 3000 times …

Number of MMDs: 3000
Asymptotics of $\hat{\text{MMD}}^2$ when $P \neq Q$

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

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

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

MMD density under $\mathcal{H}_1$
Behaviour of $MMD^2$ when $P = Q$

What happens when $P$ and $Q$ are the same?
Behaviour of $\sqrt{\text{MMD}^2}$ when $P = Q$

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

![Bar graph showing the probability distribution of $n \times \sqrt{\text{MMD}^2}$]
Behaviour of $\widehat{\text{MMD}}^2$ when $P = Q$

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

Number of MMDs: 20
Behaviour of $\hat{MMD}^2$ when $P = Q$

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

Number of MMDs: 50
Behaviour of $\overline{MMD}^2$ when $P = Q$

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

Number of MMDs: 100
Behaviour of $\hat{MMD}^2$ when $P = Q$

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

Number of MMDs: 1000
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

$$\lambda_i \psi_i(x') = \int_{X} \bar{k}(x, x') \psi_i(x) dP(x)$$

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

A summary of the asymptotics:

\begin{center}
\begin{tikzpicture}
    \begin{axis}[
        width=\textwidth,
        axis lines=left,
        xlabel={$n \times \hat{MMD}^2$},
        ylabel={Prob. of $n \times \hat{MMD}^2$},
        xmin=-2, xmax=6,
        ymin=0, ymax=0.7,
        xtick={-2,-1,0,1,2,3,4,5,6},
        ytick={0,0.1,0.2,0.3,0.4,0.5,0.6,0.7},
        legend pos=north east,
    ]
        \addplot[red,mark=none] coordinates {
            (-2,0)
            (-1,0)
            (0,0.5)
            (1,0.6)
            (2,0.5)
            (3,0.3)
            (4,0.2)
            (5,0.1)
            (6,0.05)
        };
        \addplot[blue,mark=none] coordinates {
            (-2,0)
            (-1,0)
            (0,0.1)
            (1,0.2)
            (2,0.3)
            (3,0.4)
            (4,0.5)
            (5,0.6)
            (6,0.7)
        };
        \legend{$P = Q$, $P \neq Q$}
    \end{axis}
\end{tikzpicture}
\end{center}
A statistical test

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

The figure shows the probability of $n \times \hat{MMD}^2$ as a function of $n \times \hat{MMD}^2$. The red line represents $P = Q$, and the blue line represents $P \neq Q$. The red vertical line indicates the $c_\alpha = 1 - \alpha$ quantile when $P = Q$. False negatives are indicated by a shaded area below the $P = Q$ line.
How do we get test threshold $c_\alpha$?

Permuted dog and fish samples (merdogs):

$$\tilde{X} = \begin{bmatrix} \text{fish} & \text{dog} & \text{fish} & \ldots \end{bmatrix}$$

$$\tilde{Y} = \begin{bmatrix} \text{dog} & \text{fish} & \text{dog} & \ldots \end{bmatrix}$$
How do we get test threshold $c_\alpha$?

Permuted dog and fish samples (merdogs):

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

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

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

Permuted dog and fish samples (merdogs):

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

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

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, Laurent, Guedj, Gretton (2021), MMD Aggregated Two-Sample Test, arXiv:2110.15073
Approx. null distribution of $MMD^2$ via permutation

Null distribution estimated from 500 permutations

Example: $P = Q = \mathcal{N}(0, 1)$
Consistent test w/o bootstrap

Maximum mean discrepancy (MMD):

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

Is \( MMD^2 \) significantly \( > 0? \)

\[ P = Q, \text{ null distrib. of } \overline{MMD}: \]

\[ n \overline{MMD} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l(z_l^2 - 2), \]

\( \lambda_l \) is \( l \)th eigenvalue of centered kernel \( \tilde{k}(x_i, x_j) \)

Use Gram matrix spectrum for \( \hat{\lambda}_l \):
consistent test without bootstrap
How to choose the best kernel (1) characteristic kernels
Characteristic kernels

Characteristic: MMD a metric $MMD = 0$ iff $P = Q$)
[NeurIPS07b, JMLR10]

In the next slides:

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

Reminder: Fourier series

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

$$f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(i\ell x) = \sum_{l=-\infty}^{\infty} \hat{f}_l (\cos(\ell x) + i \sin(\ell x)).$$
Characteristic kernels on \([-\pi, \pi]\)

**Jacobi theta kernel** (close to exponentiated quadratic):

\[
k(x - y) = \frac{1}{2\pi} \vartheta \left( \frac{x - y}{2\pi}, \frac{i\sigma^2}{2\pi} \right), \quad \hat{k}_\ell = \frac{1}{2\pi} \exp \left( -\frac{\sigma^2 \ell^2}{2} \right).
\]

\(\vartheta\) is the Jacobi theta function, close to Gaussian when \(\sigma^2\) 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)

$$\mu_P(x) = \langle \mu_P, k(\cdot, x) \rangle_F$$

$$= E_{t \sim P} k(t - x)$$

$$= \int_{-\pi}^{\pi} k(t - x) dP(t) \quad \hat{\mu}_{P,\ell} = \hat{k}_\ell \times \bar{\varphi}_{P,\ell}$$
The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

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

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

Maximum mean embedding via Fourier series:

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

(convolution theorem)

\[
\mu_P(x) = \langle \mu_P, k(\cdot, x) \rangle \mathcal{F} = \mathbb{E}_{t \sim P} k(t - x) = \int_{-\pi}^{\pi} k(t - x) dP(t) \\
\hat{\mu}_{P,\ell} = \hat{k}_{\ell} \times \hat{\varphi}_{P,\ell}
\]
The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

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

\[
\mu_P(x) = \langle \mu_P, k(\cdot, x) \rangle_{\mathcal{F}} \\
= E_{t \sim P} k(t - x) \\
= \int_{-\pi}^{\pi} k(t - x) dP(t) \\
\hat{\mu}_{P,\ell} = \hat{k}_\ell \times \bar{\varphi}_{P,\ell}
\]
The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

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

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

MMD can be written in terms of Fourier series:

$$MMD(P, Q; \mathcal{F}) = \| \mu_P - \mu_Q \|_{\mathcal{F}}$$
$$= \left\| \sum_{\ell=-\infty}^{\infty} \left[ (\varphi_{P,\ell} - \varphi_{Q,\ell}) \hat{k}_\ell \right] \exp(i\ell x) \right\|_{\mathcal{F}}$$
A simpler Fourier representation for MMD

From previous slide,

\[
MMD(P, Q; \mathcal{F}) = \left\| \sum_{\ell=-\infty}^{\infty} \left[ (\varphi_P,\ell - \varphi_Q,\ell) \widehat{k_\ell} \right] \exp(i\ell x) \right\|_{\mathcal{F}}
\]

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

\[
||f||^2_{\mathcal{F}} = \sum_{\ell=-\infty}^{\infty} \left| \frac{\hat{f}_\ell}{\hat{k_\ell}} \right|^2.
\]

Simple, interpretable expression for squared MMD:

\[
MMD^2(P, Q; \mathcal{F}) = \sum_{\ell=-\infty}^{\infty} \frac{|\varphi_P,\ell - \varphi_Q,\ell|^2}{\hat{k_\ell}^2} = \sum_{\ell=-\infty}^{\infty} |\varphi_P,\ell - \varphi_Q,\ell|^2 \widehat{k_\ell}
\]
A simpler Fourier representation for MMD

From previous slide,

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

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

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

Simple, interpretable expression for squared MMD:

\[
MMD^2(P, Q; \mathcal{F}) = \sum_{\ell=-\infty}^{\infty} \frac{|\varphi_P,\ell - \varphi_Q,\ell|^2 \hat{k}_\ell^2}{\hat{k}_\ell} = \sum_{\ell=-\infty}^{\infty} |\varphi_P,\ell - \varphi_Q,\ell|^2 \hat{k}_\ell
\]
Characteristic kernels on $[-\pi, \pi]$

Example: $P$ differs from $Q$ at one frequency:

\begin{align*}
P(x) &\quad \text{and} \quad Q(x) \quad \text{for} \quad x \in [-2, 2] \\
\end{align*}
Characteristic kernels on $[-\pi, \pi]$

Example: $P$ differs from $Q$ at one frequency:

\[ P(x) \]
\[ Q(x) \]
\[ \phi_{P,\ell} \]
\[ \phi_{Q,\ell} \]
Characteristic kernels on $[-\pi, \pi]$

Example: $P$ differs from $Q$ at one frequency:

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

Is the Gaussian spectrum kernel characteristic?

\[
MMD^2(P, Q; F) = \sum_{\ell=-\infty}^{\infty} \left| \varphi_{P,\ell} - \varphi_{Q,\ell} \right|^2 \hat{k}_\ell
\]
Characteristic kernels on \([-\pi, \pi]\)

Is the Gaussian spectrum kernel characteristic? **YES**

\[
MMD^2(P, Q; F) = \sum_{\ell=-\infty}^{\infty} |\varphi_{P,\ell} - \varphi_{Q,\ell}|^2\hat{k}_\ell
\]
Characteristic kernels on \([-\pi, \pi]\)

Is the triangle kernel characteristic?

\[
MMD^2(P, Q; F) = \sum_{l=-\infty}^{\infty} |\varphi_{P,l} - \varphi_{Q,l}|^2 \hat{k}_l
\]
Characteristic kernels on $[-\pi, \pi]$

Is the triangle kernel characteristic? NO

\[ MMD^2(P, Q; F) = \sum_{l=-\infty}^{\infty} |\varphi_{P,l} - \varphi_{Q,l}|^2 \hat{\kappa}_l \]
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:

$$MMD^2(P, Q; F)$$

$$:= E_P k(x - x') + E_Q k(y - y') - 2 E_{P,Q} k(x, y)$$

$$= \int \int \left[ k(s - t) \, d(P - Q)(s) \right] \, d(P - Q)(t)$$

$$\overset{(a)}{=} \int \int \int_{\mathbb{R}^d} e^{-i(s-t)^T\omega} \, d\Lambda(\omega) \, d(P - Q)(s) \, d(P - Q)(t)$$

$$\overset{(b)}{=} \int \int_{\mathbb{R}^d} e^{-is^T\omega} \, d(P - Q)(s) \int_{\mathbb{R}^d} e^{it^T\omega} \, d(P - Q)(t) \, d\Lambda(\omega)$$

$$= \int_{\mathbb{R}^d} |\varphi_P(\omega) - \varphi_Q(\omega)|^2 \, d\Lambda(\omega)$$
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:

$$MMD^2(P, Q; F)$$

$$:= E_P k(x - x') + E_Q k(y - y') - 2E_{P,Q} k(x, y)$$

$$= \int \int \left[ k(s - t) \, d(P - Q)(s) \right] \, d(P - Q)(t)$$

$$= \int \int \int_{\mathbb{R}^d} e^{-i(s-t)^T \omega} \, d\Lambda(\omega) \, d(P - Q)(s) \, d(P - Q)(t) \quad (a)$$

$$= \int \int_{\mathbb{R}^d} e^{-is^T \omega} \, d(P - Q)(s) \int_{\mathbb{R}^d} e^{it^T \omega} \, d(P - Q)(t) \, d\Lambda(\omega) \quad (b)$$

$$= \int_{\mathbb{R}^d} |\phi_P(\omega) - \phi_Q(\omega)|^2 \, d\Lambda(\omega)$$
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:

$$MMD^2(P, Q; F) := E_P k(x - x') + E_Q k(y - y') - 2E_{P,Q} k(x, y)$$

$$= \int \int \left[ k(s - t) \, d(P - Q)(s) \right] \, d(P - Q)(t)$$

$$(a) \quad \int \int \int_{\mathbb{R}^d} e^{-i(s-t)^T \omega} \, d\Lambda(\omega) \, d(P - Q)(s) \, d(P - Q)(t)$$

$$(b) \quad \int \int_{\mathbb{R}^d} e^{-is^T \omega} \, d(P - Q)(s) \int_{\mathbb{R}^d} e^{it^T \omega} \, d(P - Q)(t) \, d\Lambda(\omega)$$

$$= \int_{\mathbb{R}^d} |\varphi_P(\omega) - \varphi_Q(\omega)|^2 \, d\Lambda(\omega)$$

(a) Using Bochner’s theorem...
Characteristic kernels on $\mathbb{R}^d$

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

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

Proof:

$$MMD^2(P, Q; F)$$

$$:= E_P k(x - x') + E_Q k(y - y') - 2E_{P,Q} k(x, y)$$

$$= \int \int \left[ k(s - t) \, d(P - Q)(s) \right] \, d(P - Q)(t)$$

$$\overset{(a)}{=} \int \int \int_{\mathbb{R}^d} e^{-i(s-t)^T \omega} \, d\Lambda(\omega) \, d(P - Q)(s) \, d(P - Q)(t)$$

$$\overset{(b)}{=} \int \int_{\mathbb{R}^d} e^{-is^T \omega} \, d(P - Q)(s) \int_{\mathbb{R}^d} e^{it^T \omega} \, d(P - Q)(t) \, d\Lambda(\omega)$$

$$= \int_{\mathbb{R}^d} |\varphi_P(\omega) - \varphi_Q(\omega)|^2 \, d\Lambda(\omega)$$

(a) Using Bochner’s theorem......(b) and using Fubini’s theorem.
Characteristic kernels on $\mathbb{R}^d$

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

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

Proof:

$$MMD^2(P, Q; F)$$

$$:= E_P k(x - x') + E_Q k(y - y') - 2 E_{P,Q} k(x, y)$$

$$= \int \int \left[ k(s - t) \, d(P - Q)(s) \right] \, d(P - Q)(t)$$

$$\stackrel{(a)}{=} \int \int \int_{\mathbb{R}^d} e^{-i(s-t)^T \omega} \, d\Lambda(\omega) \, d(P - Q)(s) \, d(P - Q)(t)$$

$$\stackrel{(b)}{=} \int \int_{\mathbb{R}^d} e^{-i s^T \omega} \, d(P - Q)(s) \int_{\mathbb{R}^d} e^{i t^T \omega} \, d(P - Q)(t) \, d\Lambda(\omega)$$

$$= \int_{\mathbb{R}^d} |\phi_L(\omega) - \phi_R(\omega)|^2 \, d\Lambda(\omega)$$

(a) Using Bochner’s theorem......(b) and using Fubini’s theorem.
Characteristic kernels on $\mathbb{R}^d$

Example: $P$ differs from $Q$ at roughly one frequency:
Characteristic kernels on $\mathbb{R}^d$

Example: $P$ differs from $Q$ at roughly one frequency:
Characteristic kernels on $\mathbb{R}^d$

Example: $P$ differs from $Q$ at roughly one frequency:
Characteristic kernels on $\mathbb{R}^d$

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

Exponentiated quadratic 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 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 $|\phi_P - \phi_Q|$
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](image-url)
Choosing the best kernel (Fourier)

**Exponentiated quadratic kernel:**

![Graphs of Q(X) and \(|\phi_\omega Q|\) for different values of \(\omega\).]

MMD vs frequency of perturbation to \(P\)
Choosing the best kernel (Fourier)

**B-Spline kernel:**

MMD vs frequency of perturbation to $P$
MMD decay with increasing perturbation freq.

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

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

where \(\hat{k}_\ell\) decays as \(\ell\) grows.

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

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

has similar behaviour.
Summary: characteristic kernels on $\mathbb{R}^d$

**Characteristic kernel:** $\text{MMD} = 0$ 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 

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

(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:** $\text{MMD} = 0$ 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

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

(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

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

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

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

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

Characteristic kernels: $\text{MMD} = 0$ iff $P = Q$

Classical result:
$P = Q$ if and only if $E_P(f(x)) = E_Q(f(y))$ for all $f \in C(\mathcal{X})$, the space of bounded continuous functions on $\mathcal{X}$ \cite{Dudley2002}

Universal RKHS:
$k(x, x')$ continuous, $\mathcal{X}$ compact, and $\mathcal{F}$ dense in $C(\mathcal{X})$ with respect to $L_\infty$ \cite{Steinwart2001}

If $\mathcal{F}$ universal, then $\text{MMD}(P, Q; \mathcal{F}) = 0$ iff $P = Q$
Characteristic kernels (via Universality)

Characteristic kernels: \( MMD = 0 \) iff \( P = Q \)

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

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

If \( \mathcal{F} \) universal, then \( MMD(P, Q; \mathcal{F}) = 0 \) iff \( P = Q \)
Characteristic kernels (via Universality)

Characteristic kernels: \( MMD = 0 \) iff \( P = Q \)

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

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

If \( \mathcal{F} \) universal, then \( MMD(P, Q; \mathcal{F}) = 0 \) iff \( P = Q \)
Characteristic kernels (via Universality)

Characteristic kernels: \[ MMD = 0 \] iff \( P = Q \)

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

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

If \( \mathcal{F} \) universal, then \( MMD(P, Q; F) = 0 \) iff \( P = Q \)
Characteristic kernels (via Universality)

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

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

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

We next make the expansion

$$|E_P f(x) - E_Q f(y)|$$
$$\leq |E_P f(x) - E_P g(x)| + |E_P g(x) - E_Q g(y)| + |E_Q g(y) - E_Q f(y)|.$$ 

The first and third terms satisfy

$$|E_P f(x) - E_P g(x)| \leq E_P |f(x) - g(x)| \leq \epsilon.$$
Characteristic kernels (via Universality)

Proof:
First, it is clear that $P = Q$ implies $MMD(P, Q; \mathcal{F})$ is zero.
Converse: by the universality of $\mathcal{F}$, for any given $\epsilon > 0$ and $f \in C(\mathcal{X})$, $\exists g \in \mathcal{F}$

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

We next make the expansion

$$|\mathbb{E}_P f(x) - \mathbb{E}_Q f(y)|$$

$$\leq |\mathbb{E}_P f(x) - \mathbb{E}_P g(x)| + |\mathbb{E}_P g(x) - \mathbb{E}_Q g(y)| + |\mathbb{E}_Q g(y) - \mathbb{E}_Q f(y)|.$$ 

The first and third terms satisfy

$$|\mathbb{E}_P f(x) - \mathbb{E}_P g(x)| \leq \mathbb{E}_P |f(x) - g(x)| \leq \epsilon.$$
Proof (continued):

\[ E_P g(x) - E_Q g(y) = \langle g(\cdot), \mu_P - \mu_Q \rangle_F = 0, \]

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

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

for all \( f \in C(\mathcal{X}) \) and \( \epsilon > 0 \), which implies \( P = Q \).
How to choose the best kernel (2) optimising the kernel parameters
The best test for the job

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

With characteristic kernel, MMD test has power $\to 1$ as $n \to \infty$ for any (fixed) problem

- But, for many $P$ and $Q$, will have terrible power with reasonable $n$!

You *can* choose a good kernel for a given problem

You *can’t* get one kernel that has good finite-sample power for all problems
Choosing a kernel for the test

- Simple choice: exponentiated quadratic

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

- *Characteristic*: for any \( \sigma \): for any \( P \) and \( Q \), power \( \to 1 \) as \( n \to \infty \)
Choosing a kernel for the test

- Simple choice: exponentiated quadratic

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

- \textit{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 \( n \)...
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

![Graphical Illustration]

- $c_\alpha = 1 - \alpha$ quantile when $P = Q$

false negatives

$\text{Prob. of } n \times \hat{MMD}^2$

$n \times \hat{MMD}^2$
Optimizing kernel for test power

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

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

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

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

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

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 $P \neq Q$):

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

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

To maximize test power, maximize

$$\frac{\text{MMD}^2(P, Q)}{\sqrt{V_n(P, Q)}}$$
Choose a kernel $k$ maximizing $\frac{\overline{MMD}^2}{\sqrt{\hat{V}_n(P,Q)}}$.

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
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:

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

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

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

---


Statistics > Machine Learning

[Submitted on 21 Feb 2020]

**Learning Deep Kernels for Non-Parametric Two-Sample Tests**

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

ICML 2020
How to choose the best kernel (2)

test without data splitting
Two-sample problem

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

$$\mathbb{P}_{p \times q}(\Delta(X_m, Y_n) = 0) \leq \beta$$

Definitions:

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

$\mathcal{H}_0: p = q \quad \text{against} \quad \mathcal{H}_1: p \neq q$

$\Delta(X_m, Y_n) = 1 \quad \iff \quad \text{reject } \mathcal{H}_0$

$\Delta(X_m, Y_n) = 0 \quad \iff \quad \text{fail to reject } \mathcal{H}_0$

Type I error: controlled by $\alpha$ by design

$$\mathbb{P}_{p \times p}(\Delta(X_m, Y_n) = 1) \leq \alpha$$
Kernels and bandwidths

**Kernel:**  
\[ k_\lambda(x, y) := \prod_{i=1}^{d} K_i \left( \frac{x_i - y_i}{\lambda_i} \right) \]  

**Bandwidth:**  
\[ \lambda \in (0, \infty)^d \]

**Assumptions:**  
\( K_1, \ldots, K_d \) integrable and square integrable

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

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

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

Bandwidth: \( \lambda \in (0, \infty)^d \)

Assumptions: \( K_1, \ldots, K_d \) integrable and square integrable

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

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

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

Bandwidth: \( \lambda \in (0, \infty)^d \)

Assumptions: \( K_1, \ldots, K_d \) integrable and square integrable

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

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

- **Large bandwidth**: can only detect global differences
  - Global differences: detectable with small and large sample sizes
  - Risk: fails to detect local differences under $\mathcal{H}_1$

- **Small bandwidth**: can also detect local differences
  - Local differences: detectable only with large sample sizes
  - Risk: wrongly detects artificial local differences $\mathcal{H}_0$ (small sample sizes)

$\implies$ Bandwidths should decrease with the sample size

$\implies$ Aim: quantify at which rate $\lambda = (m + n)^{-r}$ to guarantee minimax optimal test power over a class of differences $p - q$.

- **Choice of bandwidth** is crucial for test power! Existing methods:
  - Median heuristic: no theoretical guarantees, fails in some settings
  - Data splitting: loss of power (fewer samples being used for testing)
Choice of bandwidth

- **Large bandwidth**: can only detect global differences
  - **Global differences**: detectable with small and large sample sizes
  - **Risk**: fails to detect local differences under $\mathcal{H}_1$
- **Small bandwidth**: can also detect local differences
  - **Local differences**: detectable only with large sample sizes
  - **Risk**: wrongly detects artificial local differences $\mathcal{H}_0$ (small sample sizes)

$\implies$ **Bandwidths** should decrease with the sample size

$\implies$ **Aim**: quantify at which rate $\lambda = (m + n)^{-r}$ to guarantee minimax optimal test power over a class of differences $p - q$.

- **Choice of bandwidth** is crucial for test power! Existing methods:
  - **Median heuristic**: no theoretical guarantees, fails in some settings
  - **Data splitting**: loss of power (fewer samples being used for testing)
Choice of bandwidth

- **Large bandwidth**: can only detect global differences
  - **Global differences**: detectable with small and large sample sizes
  - **Risk**: fails to detect local differences under $\mathcal{H}_1$

- **Small bandwidth**: can also detect local differences
  - **Local differences**: detectable only with large sample sizes
  - **Risk**: wrongly detects artificial local differences $\mathcal{H}_0$ (small sample sizes)

$\implies$ **Bandwidths** should decrease with the sample size

$\implies$ **Aim**: quantify at which rate $\lambda = (m + n)^{-r}$ to guarantee minimax optimal test power over a class of differences $p - q$.

- **Choice of bandwidth** is crucial for test power! Existing methods:
  - Median heuristic: no theoretical guarantees, fails in some settings
  - Data splitting: loss of power (fewer samples being used for testing)
**Choice of bandwidth**

- **Large bandwidth**: can only detect global differences
  - Global differences: detectable with small and large sample sizes
  - Risk: fails to detect local differences under $\mathcal{H}_1$

- **Small bandwidth**: can also detect local differences
  - Local differences: detectable only with large sample sizes
  - Risk: wrongly detects artificial local differences $\mathcal{H}_0$ (small sample sizes)

$\implies$ **Bandwidths** should decrease with the sample size

$\implies$ **Aim**: quantify at which rate $\lambda = (m + n)^{-r}$ to guarantee minimax optimal test power over a class of differences $p - q$.

- **Choice of bandwidth** is crucial for test power! Existing methods:
  - Median heuristic: no theoretical guarantees, fails in some settings
  - Data splitting: loss of power (fewer samples being used for testing)
MMD tests for fixed bandwidth $\lambda$

$$\Delta^\lambda_{\alpha}(X_m, Y_n) := \mathbb{1}\left(\overline{\text{MMD}}^2_{\lambda}(X_m, Y_n) > \hat{q}^\lambda_{1-\alpha}\right)$$

Quantile: $\hat{q}^\lambda_{1-\alpha}$ is the $[(B+1)(1-\alpha)]$-th largest value of $\overline{\text{MMD}}^2_{\lambda}(X_m, Y_n)$ and $B \mathcal{H}_0$-simulated test statistics

Permutations: $\overline{\text{MMD}}^2_{\lambda}(X^\sigma_m, Y^\sigma_n)$ where $(X^\sigma_m, Y^\sigma_n) = \sigma(X_m \cup Y_n)$
**MMD tests for fixed bandwidth \( \lambda \)**

\[
\Delta_\alpha^\lambda(X_m, Y_n) := 1 \left( \widehat{\text{MMD}}^2_\lambda(X_m, Y_n) > \widehat{q}_{1-\alpha}^\lambda \right)
\]

**Quantile:** \( \widehat{q}_{1-\alpha}^\lambda \) is the \([(B+1)(1-\alpha)]\)-th largest value of \( \widehat{\text{MMD}}^2_\lambda(X_m, Y_n) \) and \( B \) \( \mathcal{H}_0 \)-simulated test statistics

**Permutations:** \( \widehat{\text{MMD}}^2_\lambda(X_m, Y_n) \) where \( (X_m, Y_n) = \sigma(X_m \cup Y_n) \)

**Wild bootstrap:** case \( m = n, \varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} \text{Unif}\{-1, 1\} \) (Rademacher)

\[
\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \varepsilon_i \varepsilon_j \left( k_\lambda(X_i, X_j) - k_\lambda(X_i, Y_j) - k_\lambda(Y_i, X_j) + k_\lambda(Y_i, Y_j) \right)
\]
MMD tests for \textit{fixed} bandwidth \( \lambda \)

\[
\Delta^\lambda_\alpha(X_m, Y_n) := 1\left( \hat{\text{MMD}}^2_\lambda(X_m, Y_n) > \hat{q}^\lambda_{1-\alpha} \right)
\]

Quantile: \( \hat{q}^\lambda_{1-\alpha} \) is the \([B + 1](1 - \alpha)\)-th largest value of \( \hat{\text{MMD}}^2_\lambda(X_m, Y_n) \) and \( B \) \( \mathcal{H}_0 \)-simulated test statistics

Permutations: \( \hat{\text{MMD}}^2_\lambda(X^\sigma_m, Y^\sigma_n) \) where \( (X^\sigma_m, Y^\sigma_n) = \sigma(X_m \cup Y_n) \)

Wild bootstrap: case \( m = n, \epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} \text{Unif}\{-1, 1\} \) (Rademacher)

\[
\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \epsilon_i \epsilon_j \left( k_\lambda(X_i, X_j) - k_\lambda(X_i, Y_j) - k_\lambda(Y_i, X_j) + k_\lambda(Y_i, Y_j) \right)
\]

Non-asymptotic level (permutation and wild bootstrap):

\[
\mathbb{P}_{p \times p}(\Delta(X_m, Y_n) = 1) \leq \alpha, \text{ Time complexity: } \mathcal{O}(B(m + n)^2)
\]
MMD tests for fixed bandwidth $\lambda$

$$\Delta^\lambda_\alpha(X_m, Y_n) := \mathbb{1}\left(\widehat{\text{MMD}}^2_\lambda(X_m, Y_n) > \widehat{q}_{1-\alpha}^\lambda\right)$$

Quantile: $\widehat{q}_{1-\alpha}^\lambda$ is the $[(B+1)(1-\alpha)]$-th largest value of $\widehat{\text{MMD}}^2_\lambda(X_m, Y_n)$ and $B \mathcal{H}_0$-simulated test statistics

Permutations: $\widehat{\text{MMD}}^2_\lambda(X^\sigma_m, Y^\sigma_n)$ where $(X^\sigma_m, Y^\sigma_n) = \sigma(X_m \cup Y_n)$

Wild bootstrap: case $m = n$, $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} \text{Unif}\{-1, 1\}$ (Rademacher)

$$\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \epsilon_i \epsilon_j \left(k^\lambda(X_i, X_j) - k^\lambda(X_i, Y_j) - k^\lambda(Y_i, X_j) + k^\lambda(Y_i, Y_j)\right)$$

Non-asymptotic level (permutation and wild bootstrap):
$$\mathbb{P}_p \times_p (\Delta(X_m, Y_n) = 1) \leq \alpha$$

Time complexity: $O\left(B(m+n)^2\right)$

Power guarantee need smoothness assumption on $p-q$
Sobolev balls

Regularity/smoothness assumption: \( p - q \in S^s_d(R) \)

Sobolev balls:

\[
S^s_d(R) := \left\{ f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} ||\xi||_2^{2s} |\hat{f}(\xi)|^2 \, d\xi \leq (2\pi)^d R^2 \right\}
\]

radius \( R > 0 \)

smoothness parameter \( s > 0 \)

dimension \( d \)

Fourier transform \( \hat{f}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-ix^T\xi} \, dx \)

![Graph showing the influence of different smoothness parameters on the function's behavior.](image-url)
Sobolev balls

Regularity/smoothness assumption: $p - q \in S^s_d(R)$

Sobolev balls:

$$S^s_d(R) := \left\{ f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|\xi\|^{2s} |\hat{f}(\xi)|^2 \, d\xi \leq (2\pi)^d R^2 \right\}$$

radius $R > 0$

smoothness parameter $s > 0$

dimension $d$

Fourier transform $\hat{f}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-ix^T\xi} \, dx$
Sobolev balls

Regularity/smoothness assumption: $p - q \in S^s_d(R)$

Sobolev balls:

$$S^s_d(R) := \left\{ f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|\xi\|_2^{2s} |\hat{f}(\xi)|^2 \ d\xi \leq (2\pi)^d R^2 \right\}$$

radius $R > 0$
smoothness parameter $s > 0$

dimension $d$

Fourier transform $\hat{f}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-ix^T\xi} \ dx$
MMD test power, known smoothness

Theorem (MMD test minimax optimality)

For known smoothness $s$, assuming $p - q \in S_d^s(R)$ and setting

$$\lambda_i^* := (m + n)^{-2/(4s+d)}$$

for $i = 1, \ldots, d$, the condition

$$\|p - q\|_2 \geq \frac{C}{\sqrt{\beta}} (m + n)^{-2s/(4s+d)}$$

guarantees control of the type II error of the MMD test

$$\mathbb{P}_{p \times q} \left( \Delta_{\alpha}^\lambda (X_m, Y_n) = 0 \right) \leq \beta.$$ 

Minimax rate over Sobolev balls: $(m + n)^{-2s/(4s+d)}$

Can we be adaptive to the unknown smoothness $s$?
MMD test power, known smoothness

Theorem (MMD test minimax optimality)

For known smoothness $s$, assuming $p - q \in S^s_d(R)$ and setting

$$\lambda^*_i := (m + n)^{-2/(4s+d)}$$

for $i = 1, \ldots, d$, the condition

$$\|p - q\|_2 \geq \frac{C}{\sqrt{\beta}} (m + n)^{-2s/(4s+d)}$$

guarantees control of the type II error of the MMD test

$$\mathbb{P}_{p \times q} \left( \Delta_{\alpha}^* (X_m, Y_n) = 0 \right) \leq \beta.$$ 

Minimax rate over Sobolev balls: $(m + n)^{-2s/(4s+d)}$

Can we be adaptive to the unknown smoothness $s$?
MMDAgg for a *collection* of bandwidths $\Lambda$

Bonferroni multiple testing: non-asymptotic level $\alpha$

$$\Delta^\Lambda_\alpha(X_m, Y_n) := \mathbb{1} \left( \widehat{\text{MMD}}^2_\lambda(X_m, Y_n) > \widehat{q}^\lambda_{1-\alpha/|\Lambda|} \text{ for some } \lambda \in \Lambda \right)$$

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

MMDAgg (MMD Aggregation): non-asymptotic level $\alpha$

$$\Delta^\Lambda_\alpha(X_m, Y_n) := \mathbb{1} \left( \widehat{\text{MMD}}^2_\lambda(X_m, Y_n) > \widehat{q}^\lambda_{1-u_\alpha w_\lambda} \text{ for some } \lambda \in \Lambda \right)$$

positive weights $(w_\lambda)_{\lambda \in \Lambda}$ satisfying $\sum_{\lambda \in \Lambda} w_\lambda \leq 1$

Correction $u_\alpha$ defined as

$$\sup \left\{ u > 0 : \mathbb{P}_{p \times p} \left( \max_{\lambda \in \Lambda} \left( \widehat{\text{MMD}}^2_\lambda(X_m, Y_n) - \widehat{q}^\lambda_{1-u w_\lambda} \right) > 0 \right) \leq \alpha \right\}$$

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

Time complexity $\mathcal{O}(|\Lambda| (B_1 + B_2) (m + n)^2)$
**MMDAgg for a collection of bandwidths Λ**

**Bonferroni multiple testing: non-asymptotic level α**

\[
\Delta_\alpha^\Lambda (X_m, Y_n) := \mathbb{1} \left( \widehat{\text{MMD}}_\lambda^2 (X_m, Y_n) > \widehat{q}_{1-\alpha/|\Lambda|}^{\lambda} \right. \text{ for some } \lambda \in \Lambda \right)
\]

Time complexity \(O\left(|\Lambda| B_1 (m + n)^2\right)\)

**MMDAgg (MMD Aggregation): non-asymptotic level α**

\[
\Delta_\alpha^\Lambda (X_m, Y_n) := \mathbb{1} \left( \widehat{\text{MMD}}_\lambda^2 (X_m, Y_n) > \widehat{q}_{1-u_\alpha}^{\lambda} w_\lambda \right. \text{ for some } \lambda \in \Lambda \right)
\]

Positive weights \((w_\lambda)_{\lambda \in \Lambda}\) satisfying \(\sum_{\lambda \in \Lambda} w_\lambda \leq 1\)

**Correction** \(u_\alpha\) defined as

\[
\sup \left\{ u > 0 : \mathbb{P}_{p \times p} \left( \max_{\lambda \in \Lambda} \left( \widehat{\text{MMD}}_\lambda^2 (X_m, Y_n) - \widehat{q}_{1-u}^{\lambda} w_\lambda \right) > 0 \right) \leq \alpha \right\}
\]

More powerful than Bonferroni correction as \(u_\alpha \geq \alpha\)

Time complexity \(O\left(|\Lambda| (B_1 + B_2) (m + n)^2\right)\)
MMDAgg for a collection of bandwidths $\Lambda$

Bonferroni multiple testing: non-asymptotic level $\alpha$

\[ \Delta^\Lambda_\alpha(X_m, Y_n) := 1 \left( \widehat{\text{MMD}}^2_\lambda(X_m, Y_n) > \hat{q}^\lambda_{1-\alpha/|\Lambda|} \right) \text{ for some } \lambda \in \Lambda \]

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

MMDAgg (MMD Aggregation): non-asymptotic level $\alpha$

\[ \Delta^\Lambda_\alpha(X_m, Y_n) := 1 \left( \widehat{\text{MMD}}^2_\lambda(X_m, Y_n) > \hat{q}^\lambda_{1-\alpha} w_\lambda \right) \text{ for some } \lambda \in \Lambda \]

positive weights $(w_\lambda)_{\lambda \in \Lambda}$ satisfying $\sum_{\lambda \in \Lambda} w_\lambda \leq 1$

Correction $u_\alpha$ defined as

\[ \sup \left\{ u > 0 : \mathbb{P}_{p \times p} \left( \max_{\lambda \in \Lambda} \left( \widehat{\text{MMD}}^2_\lambda(X_m, Y_n) - \hat{q}^\lambda_{1-u} w_\lambda \right) > 0 \right) \leq \alpha \right\} \]

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

Time complexity $\mathcal{O}(|\Lambda| (B_1 + B_2) (m + n)^2)$
MMDAgg for a collection of bandwidths $\Lambda$

Bonferroni multiple testing: non-asymptotic level $\alpha$

$$\Delta^\Lambda_\alpha(X_m, Y_n) := \mathbb{1}\left(\overline{\text{MMD}}^2_{\lambda}(X_m, Y_n) > \hat{q}_{1-\alpha/|\Lambda|}^\lambda \text{ for some } \lambda \in \Lambda\right)$$

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

MMDAgg (MMD Aggregation): non-asymptotic level $\alpha$

$$\Delta^\Lambda_\alpha(X_m, Y_n) := \mathbb{1}\left(\overline{\text{MMD}}^2_{\lambda}(X_m, Y_n) > \hat{q}_{1-u_\alpha w_{\lambda}}^\lambda \text{ for some } \lambda \in \Lambda\right)$$

positive weights $(w_{\lambda})_{\lambda \in \Lambda}$ satisfying $\sum_{\lambda \in \Lambda} w_{\lambda} \leq 1$

Correction $u_\alpha$ defined as

$$\sup\left\{u > 0 : \mathbb{P}_{p \times p} \left(\max_{\lambda \in \Lambda} \left(\overline{\text{MMD}}^2_{\lambda}(X_m, Y_n) - \hat{q}_{1-u w_{\lambda}}^\lambda\right) > 0\right) \leq \alpha\right\}$$

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

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

Simple example: 3-d Gaussians with different means

\[ \Lambda(i) := \left\{ 2^l \lambda_{\text{med}} : l \in \{-i, \ldots, i\} \right\} \quad \text{for } i \in \{0, 10, 20, 30, 40, 50\} \]

\[ w_\lambda := 1 / |\Lambda| \]
MMDAgg test power guarantee

Theorem (MMDAgg minimax adaptivity)

\[
\Lambda^*: = \left\{ 2^{-\ell} 1_d : \ell \in \{1, \ldots, \left\lfloor \frac{2}{d} \log_2 \left( \frac{m + n}{\ln(\ln(m + n))} \right) \right\} \right\}, \ w_\lambda := \frac{6}{\pi^2 \ell^2}
\]

Assuming \( p - q \in S^s_d(R) \), the condition

\[
\|p - q\|_2 \geq \frac{C}{\sqrt{\beta}} \left( \frac{m + n}{\ln(\ln(m + n))} \right)^{-2s/(4s+d)}
\]

guarantees control of the type II error of MMDAgg

\[
\mathbb{P}_{p \times q} \left( \Delta^\Lambda^*_\alpha (X_m, Y_n) = 0 \right) \leq \beta.
\]

Minimax rate over Sobolev balls: \((m + n)^{-2s/(4s+d)}\)

Minimax adaptive over\( \{S^s_d(R) : s > 0, R > 0 \}\)

Unlike the MMD test \( \Delta^\lambda^*_\alpha \), MMDAgg \( \Delta^\Lambda^*_\alpha \) is independent of \( s \)
Radial basis function (RBF) kernel: \( k_\lambda(x, y) := K \left( \frac{||x - y||}{\lambda} \right) \)

Collection of bandwidths \( \Lambda \): discretisation of the interval \([\lambda_{\text{min}}, \lambda_{\text{max}}]\)

where \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \) are the (robust) minimum and maximum of

\[
\left\{ ||x - y|| : x \in X_m, y \in Y_n \right\}
\]

Possible to aggregate several kernels each with multiple bandwidths

Uniform weights: \( w_\lambda := 1 / |\Lambda| \)

Number of permutations / wild bootstraps: \( B_1 = B_2 = 2000 \)

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

- JAX GPU runs 100 times faster than Numpy CPU

mmdagg package: [github.com/antoninschrab/mmdagg](https://github.com/antoninschrab/mmdagg)

```python
from mmdagg import mmdagg  # X shape (m, d)
output = mmdagg(X, Y)  # 0 or 1  # Y shape (n, d)
```
**MMDAgg parameter-free user-friendly implementation**

Radial basis function (RBF) kernel: \( k_\lambda(x, y) := K \left( \frac{x - y}{\lambda} \right) \)

Collection of bandwidths \( \Lambda \): discretisation of the interval \([\lambda_{\text{min}}, \lambda_{\text{max}}]\)
where \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \) are the (robust) minimum and maximum of
\[
\left\{ \|x - y\| : x \in X_m, y \in Y_n \right\}
\]

Possible to aggregate several kernels each with multiple bandwidths

Uniform weights: \( w_\lambda := 1 / |\Lambda| \)

Number of permutations / wild bootstraps: \( B_1 = B_2 = 2000 \)

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

- JAX GPU runs 100 times faster than Numpy CPU

**mmdagg package**: [github.com/antoninschrab/mmdagg](https://github.com/antoninschrab/mmdagg)

```python
from mmdagg import mmdagg
output = mmdagg(X, Y) # 0 or 1
```

# X shape (m, d)
# Y shape (n, d)
MMDAgg parameter-free user-friendly implementation

Radial basis function (RBF) kernel: \( k_\lambda(x, y) := K \left( \frac{\|x - y\|}{\lambda} \right) \)

Collection of bandwidths \( \Lambda \): discretisation of the interval \([\lambda_{\text{min}}, \lambda_{\text{max}}]\)
where \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \) are the (robust) minimum and maximum of
\[ \left\{ \|x - y\| : x \in X_m, y \in Y_n \right\} \]

Possible to aggregate several kernels each with multiple bandwidths

Uniform weights: \( w_\lambda := 1 / |\Lambda| \)

Number of permutations / wild bootstraps: \( B_1 = B_2 = 2000 \)

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

- JAX GPU runs 100 times faster than Numpy CPU

mmdagg package: github.com/antoninschrab/mmdagg

```python
from mmdagg import mmdagg  # X shape (m, d)
output = mmdagg(X, Y)  # 0 or 1  # Y shape (n, d)
```
MMDAgg parameter-free user-friendly implementation

Radial basis function (RBF) kernel: \( k_\lambda(x, y) := K\left(\frac{\|x - y\|}{\lambda}\right) \)

Collection of bandwidths \( \Lambda \): discretisation of the interval \([\lambda_{\text{min}}, \lambda_{\text{max}}]\)
where \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \) are the (robust) minimum and maximum of
\[ \left\{ \|x - y\| : x \in \mathbb{X}_m, y \in \mathbb{Y}_n \right\} \]

Possible to aggregate several kernels each with multiple bandwidths

Uniform weights: \( w_\lambda := 1 / |\Lambda| \)

Number of permutations / wild bootstraps: \( B_1 = B_2 = 2000 \)

JAX: runs on either CPU or GPU (significant speed improvements)
- JAX GPU runs 100 times faster than Numpy CPU

mmdagg package: github.com/antoninschrab/mmdagg

```python
from mmdagg import mmdagg
output = mmdagg(X, Y) # 0 or 1
```

# X shape (m, d)
# Y shape (n, d)
MMDAgg parameter-free user-friendly implementation

Radial basis function (RBF) kernel: \( k_\lambda(x, y) := K \left( \frac{||x - y||}{\lambda} \right) \)

Collection of bandwidths \( \Lambda \): discretisation of the interval \([\lambda_{\min}, \lambda_{\max}]\)
where \( \lambda_{\min} \) and \( \lambda_{\max} \) are the (robust) minimum and maximum of
\[
\left\{ ||x - y|| : x \in \mathbb{X}_m, y \in \mathbb{Y}_n \right\}
\]
Possible to aggregate several kernels each with multiple bandwidths

Uniform weights: \( w_\lambda := 1 / |\Lambda| \)

Number of permutations / wild bootstraps: \( B_1 = B_2 = 2000 \)

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

JAX GPU runs 100 times faster than Numpy CPU

mmdagg package: github.com/antoninschrab/mmdagg

```python
from mmdagg import mmdagg
output = mmdagg(X, Y) # 0 or 1
```

# X shape (m, d)
# Y shape (n, d)
MMDAgg parameter-free user-friendly implementation

Radial basis function (RBF) kernel: \( k_\lambda(x, y) := K\left(\frac{|x - y|}{\lambda}\right) \)

Collection of bandwidths \( \Lambda \): discretisation of the interval \([\lambda_{\text{min}}, \lambda_{\text{max}}]\)
where \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \) are the (robust) minimum and maximum of
\[
\left\{ ||x - y|| : x \in X_m, y \in Y_n \right\}
\]
Possible to aggregate several kernels each with multiple bandwidths

Uniform weights: \( w_\lambda := 1 / |\Lambda| \)

Number of permutations / wild bootstraps: \( B_1 = B_2 = 2000 \)

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

- JAX GPU runs 100 times faster than Numpy CPU

mmdagg package: [github.com/antoninschrab/mmdagg](https://github.com/antoninschrab/mmdagg)

```python
from mmdagg import mmdagg
output = mmdagg(X, Y) # 0 or 1
```

# X shape (m, d)
# Y shape (n, d)
MMDAgg parameter-free user-friendly implementation

Radial basis function (RBF) kernel: \( k_\lambda(x, y) := K \left( \frac{x - y}{\lambda} \right) \)

Collection of bandwidths \( \Lambda \): discretisation of the interval \([\lambda_{\min}, \lambda_{\max}]\)
where \( \lambda_{\min} \) and \( \lambda_{\max} \) are the (robust) minimum and maximum of \( \{||x - y|| : x \in X_m, y \in Y_n\} \)

Possible to aggregate several kernels each with multiple bandwidths

Uniform weights: \( w_\lambda := 1 / |\Lambda| \)

Number of permutations / wild bootstraps: \( B_1 = B_2 = 2000 \)

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

- JAX GPU runs 100 times faster than Numpy CPU

mmdagg package: github.com/antoninschrab/mmdagg

```python
from mmdagg import mmdagg  # X shape (m, d)
output = mmdagg(X, Y)  # 0 or 1  # Y shape (n, d)
```
MMDAgg parameter-free user-friendly implementation

Radial basis function (RBF) kernel: \( k_\lambda(x, y) := K\left(\frac{x - y}{\lambda}\right) \)

Collection of bandwidths \( \Lambda \): discretisation of the interval \([\lambda_{\text{min}}, \lambda_{\text{max}}]\)
where \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \) are the (robust) minimum and maximum of
\[ \{\|x - y\| : x \in X_m, y \in Y_n\} \]

Possible to aggregate several kernels each with multiple bandwidths

Uniform weights: \( w_\lambda := 1 / |\Lambda| \)

Number of permutations / wild bootstraps: \( B_1 = B_2 = 2000 \)

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

JAX GPU runs 100 times faster than Numpy CPU

mmdagg package: github.com/antoninschrab/mmdagg

```python
from mmdagg import mmdagg
output = mmdagg(X, Y) # 0 or 1
```

# X shape (m, d)
# Y shape (n, d)
Experiment on perturbed uniform \( d = 1 \)
Experiment on perturbed uniform $d = 1$

Number of perturbations

Power

- MMDAgg
- MMD median
- MMD split
- MMD extra data
- AutoML

Number of perturbations

1 2 3 4
Experiment on perturbed uniform $d = 2$
Experiment on perturbed uniform $d = 2$
Experiment on MNIST digits
Experiment on MNIST digits
Experiment on MNIST digits
Experiment on MNIST digits
Experiment on MNIST digits
Experiment on MNIST digits
## Experiment on MNIST digits

<table>
<thead>
<tr>
<th>Number of digits missing from one sample</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.94</td>
</tr>
<tr>
<td>0.50</td>
<td>0.88</td>
</tr>
<tr>
<td>0.75</td>
<td>0.78</td>
</tr>
<tr>
<td>1.00</td>
<td>0.67</td>
</tr>
</tbody>
</table>

### Graph

- **Red star**: MMDAgg
- **Orange dash**: MMD split
- **Blue dash**: MMD median
- **Purple dash-dot**: AutoML
- **Green dash-dot**: MMD extra data

**Power** decreases as the number of digits missing from one sample increases, reflecting the performance of different methods in handling missing data.
Experiment on image shifts on MNIST & CIFAR-10

Failing Loudly Benchmark: Rabanser et al., 2019
Experiment on image shifts on MNIST & CIFAR-10

Failing Loudly Benchmark: Rabanser et al., 2019
Experiment on image shifts on MNIST & CIFAR-10

Number of samples

0.00
0.25
0.50
0.75
1.00

Power

MMDAgg AutoML

97/100
MMD kernel choice without data splitting

MMD Aggregated Two-Sample Test (JMLR 2023):

Code:

https://github.com/antoninschrab/mmdagg-paper
Work supported by:

The Gatsby Charitable Foundation

Deepmind
Questions?