

Representing and comparing probabilities with kernels: Part 1

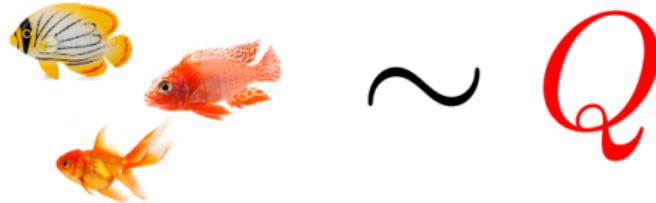
Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

MLSS Madrid, 2018

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

- Have: Two collections of samples X , Y from unknown distributions P and Q .
- Goal: do P and Q differ?



MNIST samples

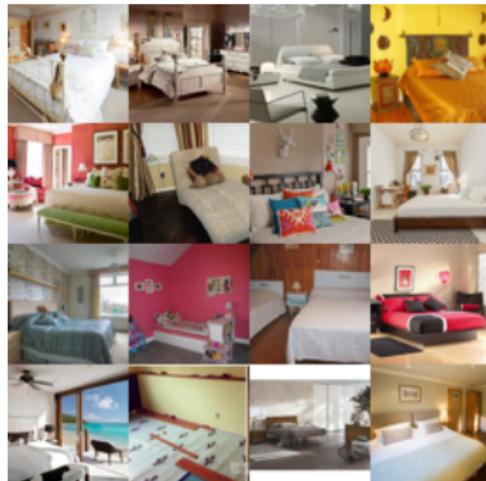


Samples from a GAN

Significant difference in GAN and MNIST?

Training generative models

- Have: One collection of samples X from unknown distribution P .
- Goal: generate samples Q that look like P



LSUN bedroom samples P



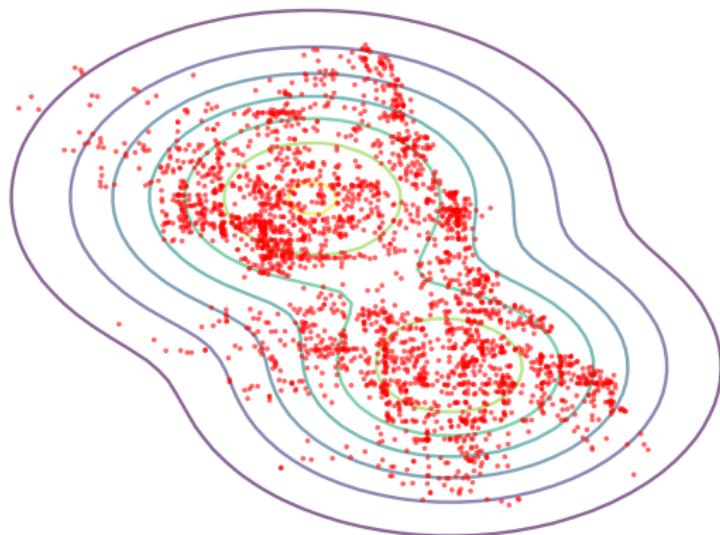
Generated Q , MMD GAN

Using MMD to train a GAN

(Binkowski, Sutherland, Arbel, G., ICLR 2018)
(Arbel, Sutherland, Binkowski, G., arXiv 2018)

Testing goodness of fit

- Given: A model P and samples and Q .
- Goal: is P a good fit for Q ?



Chicago crime data

Model is Gaussian mixture with **two** components.

Testing independence

- Given: Samples from a distribution P_{XY}
- Goal: Are X and Y independent?

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

Text from dogtime.com and petfinder.com

Outline: part 1

What is a reproducing kernel Hilbert space?

- 1 Hilbert space
- 2 Kernel (lots of examples: e.g. you can build kernels from simpler kernels)
- 3 Reproducing property
- 4 Using kernels to enforce smoothness

Classical results

- 1 Representer theorem
- 2 Kernel ridge regression

Outline: part 2

The maximum mean discrepancy (MMD)

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

Statistical testing with the MMD

- How to choose the best kernel

Training GANs with MMD

- Learning kernel features with gradient regularisation

Characteristic kernels: “is my feature space rich enough?”

Outline: part 3

Goodness of fit testing

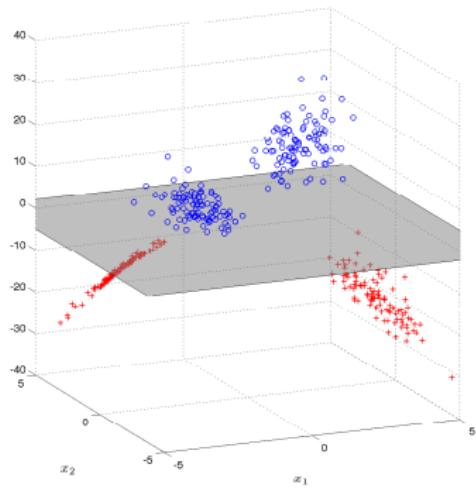
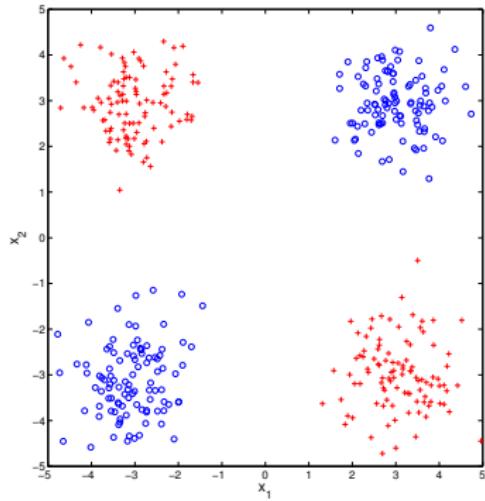
- The kernel Stein discrepancy

Dependence testing

- Dependence using the MMD
- Dependence using feature covariances
- Statistical testing

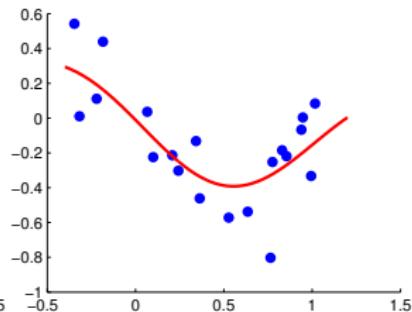
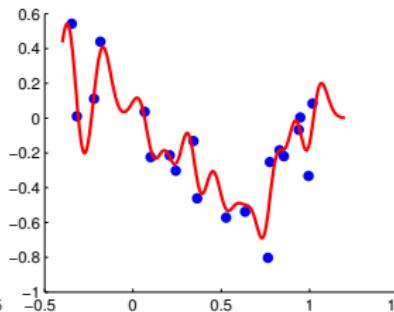
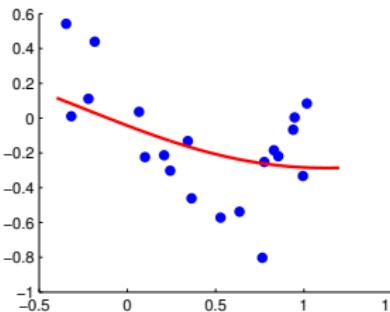
Reproducing Kernel Hilbert Spaces

Kernels and feature space (1): XOR example



- No linear classifier separates red from blue
- Map points to higher dimensional feature space:
$$\phi(x) = \begin{bmatrix} x_1 & x_2 & x_1x_2 \end{bmatrix} \in \mathbb{R}^3$$

Kernels and feature space (2): smoothing



Kernel methods can control smoothness and avoid overfitting/underfitting.

Hilbert space

Definition (Inner product)

Let \mathcal{H} be a vector space over \mathbb{R} . A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ is an **inner product** on \mathcal{H} if

- 1 Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
- 2 Symmetric: $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
- 3 $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

Norm induced by the inner product: $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

Definition (Hilbert space)

Inner product space containing Cauchy sequence limits.

Hilbert space

Definition (Inner product)

Let \mathcal{H} be a vector space over \mathbb{R} . A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ is an **inner product** on \mathcal{H} if

- 1 Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
- 2 Symmetric: $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
- 3 $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

Norm induced by the inner product: $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

Definition (Hilbert space)

Inner product space containing Cauchy sequence limits.

Hilbert space

Definition (Inner product)

Let \mathcal{H} be a vector space over \mathbb{R} . A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ is an **inner product** on \mathcal{H} if

- 1 Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
- 2 Symmetric: $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
- 3 $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

Norm induced by the inner product: $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

Definition (Hilbert space)

Inner product space containing Cauchy sequence limits.

Kernel

Definition

Let \mathcal{X} be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a **kernel** if there exists an \mathbb{R} -Hilbert space and a map $\phi : \mathcal{X} \rightarrow \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}.$$

- Almost no conditions on \mathcal{X} (eg, \mathcal{X} itself doesn't need an inner product, eg. documents).
- A single kernel can correspond to several possible features. A trivial example for $\mathcal{X} := \mathbb{R}$:

$$\phi_1(x) = x \quad \text{and} \quad \phi_2(x) = \begin{bmatrix} x/\sqrt{2} \\ x/\sqrt{2} \end{bmatrix}$$

New kernels from old: sums, transformations

Theorem (Sums of kernels are kernels)

Given $\alpha > 0$ and k, k_1 and k_2 all kernels on \mathcal{X} , then αk and $k_1 + k_2$ are kernels on \mathcal{X} .

(Proof via positive definiteness: **later!**) A difference of kernels may not be a kernel (**why?**)

New kernels from old: products

Theorem (Products of kernels are kernels)

Given k_1 on \mathcal{X}_1 and k_2 on \mathcal{X}_2 , then $k_1 \times k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$.

If $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$, then $k := k_1 \times k_2$ is a kernel on \mathcal{X} .

Proof: Main idea only!

\mathcal{H}_1 space of kernels between shapes,

$$\phi_1(x) = \begin{bmatrix} \mathbb{I}_{\square} \\ \mathbb{I}_{\triangle} \end{bmatrix} \quad \phi_1(\square) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad k_1(\square, \triangle) = 0.$$

\mathcal{H}_2 space of kernels between colors,

$$\phi_2(x) = \begin{bmatrix} \mathbb{I}_{\bullet} \\ \mathbb{I}_{\bullet} \end{bmatrix} \quad \phi_2(\bullet) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad k_2(\bullet, \bullet) = 1.$$

New kernels from old: products

“Natural” feature space for colored shapes:

$$\Phi(x) = \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \\ \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{\bullet} \\ \mathbb{I}_{\bullet} \end{bmatrix} \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \phi_2(x)\phi_1^\top(x)$$

Kernel is:

$$\begin{aligned} k(x, x') &= \sum_{i \in \{\bullet, \circ\}} \sum_{j \in \{\square, \triangle\}} \Phi_{ij}(x) \Phi_{ij}(x') = \text{tr} \left(\underbrace{\phi_1(x) \phi_2^\top(x)}_{k_2(x, x')} \underbrace{\phi_2(x') \phi_1^\top(x')}_{k_2(x', x)} \right) \\ &= \text{tr} \left(\underbrace{\phi_1^\top(x') \phi_1(x)}_{k_1(x, x')} \right) k_2(x, x') = k_1(x, x') k_2(x, x') \end{aligned}$$

New kernels from old: products

“Natural” feature space for colored shapes:

$$\Phi(x) = \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \\ \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{\bullet} \\ \mathbb{I}_{\bullet} \end{bmatrix} \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \phi_2(x)\phi_1^\top(x)$$

Kernel is:

$$\begin{aligned} k(x, x') &= \sum_{i \in \{\bullet, \bullet\}} \sum_{j \in \{\square, \triangle\}} \Phi_{ij}(x)\Phi_{ij}(x') = \text{tr} \left(\underbrace{\phi_1(x)\phi_2^\top(x)\phi_2(x')\phi_1^\top(x')}_{k_2(x, x')} \right) \\ &= \text{tr} \left(\underbrace{\phi_1^\top(x')\phi_1(x)}_{k_1(x, x')} \right) k_2(x, x') = k_1(x, x')k_2(x, x') \end{aligned}$$

Sums and products \implies polynomials

Theorem (Polynomial kernels)

Let $x, x' \in \mathbb{R}^d$ for $d \geq 1$, and let $m \geq 1$ be an integer and $c \geq 0$ be a positive real. Then

$$k(x, x') := (\langle x, x' \rangle + c)^m$$

is a valid kernel.

To prove: expand into a sum (with non-negative scalars) of kernels $\langle x, x' \rangle$ raised to integer powers. These individual terms are valid kernels by the product rule.

Infinite sequences

The kernels we've seen so far are dot products between **finitely** many features. E.g.

$$k(x, y) = \begin{bmatrix} \sin(x) & x^3 & \log x \end{bmatrix}^\top \begin{bmatrix} \sin(y) & y^3 & \log y \end{bmatrix}$$

$$\text{where } \phi(x) = \begin{bmatrix} \sin(x) & x^3 & \log x \end{bmatrix}$$

Can a kernel be a dot product between **infinitely many features**?

Infinite sequences

Definition

The space ℓ_2 (**square summable sequences**) comprises all sequences $a := (a_i)_{i \geq 1}$ for which

$$\|a\|_{\ell_2}^2 = \sum_{\ell=1}^{\infty} a_{\ell}^2 < \infty.$$

Definition

Given sequence of functions $(\phi_{\ell}(x))_{\ell \geq 1}$ in ℓ_2 where $\phi_{\ell} : \mathcal{X} \rightarrow \mathbb{R}$ is the i th coordinate of $\phi(x)$. Then

$$k(x, x') := \sum_{\ell=1}^{\infty} \phi_{\ell}(x)\phi_{\ell}(x') \tag{1}$$

Infinite sequences

Definition

The space ℓ_2 (**square summable sequences**) comprises all sequences $a := (a_i)_{i \geq 1}$ for which

$$\|a\|_{\ell_2}^2 = \sum_{\ell=1}^{\infty} a_{\ell}^2 < \infty.$$

Definition

Given sequence of functions $(\phi_{\ell}(x))_{\ell \geq 1}$ in ℓ_2 where $\phi_{\ell} : \mathcal{X} \rightarrow \mathbb{R}$ is the i th coordinate of $\phi(x)$. Then

$$k(x, x') := \sum_{\ell=1}^{\infty} \phi_{\ell}(x) \phi_{\ell}(x') \tag{1}$$

Infinite sequences (proof)

Why square summable? By Cauchy-Schwarz,

$$\left| \sum_{\ell=1}^{\infty} \phi_{\ell}(x) \phi_{\ell}(x') \right| \leq \| \phi(x) \|_{\ell_2} \| \phi(x') \|_{\ell_2},$$

so the sequence defining the inner product converges for all $x, x' \in \mathcal{X}$

A famous infinite feature space kernel

Exponentiated quadratic kernel,

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) = \sum_{\ell=1}^{\infty} \underbrace{(\sqrt{\lambda_\ell} e_\ell(x))}_{\phi_\ell(x)} \underbrace{(\sqrt{\lambda_\ell} e_\ell(x'))}_{\phi_\ell(x')}$$

$$\lambda_\ell e_\ell(x) = \int k(x, x') e_\ell(x') p(x') dx',$$

$$p(x) = \mathcal{N}(0, \sigma^2).$$

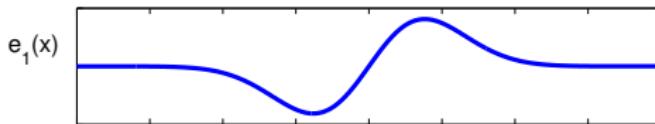
A famous infinite feature space kernel

Exponentiated quadratic kernel,

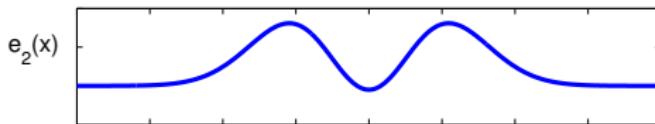
$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) = \sum_{\ell=1}^{\infty} \underbrace{(\sqrt{\lambda_\ell} e_\ell(x))}_{\phi_\ell(x)} \underbrace{(\sqrt{\lambda_\ell} e_\ell(x'))}_{\phi_\ell(x')}$$

$$\lambda_\ell e_\ell(x) = \int k(x, x') e_\ell(x') p(x') dx',$$

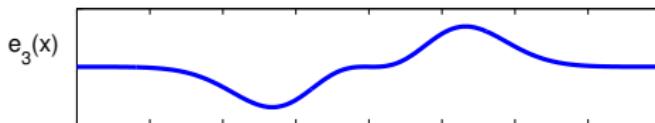
$$p(x) = \mathcal{N}(0, \sigma^2).$$



$$\lambda_\ell \propto b^\ell \quad b < 1$$



$$e_\ell(x) \propto \exp(-(c - a)x^2) H_\ell(x\sqrt{2c}),$$



a, b, c are functions of σ ,
and H_ℓ is ℓ th order Hermite polynomial.

Positive definite functions

If we are given a function of two arguments, $k(x, x')$, how can we determine if it is a valid kernel?

1 Find a feature map?

- 1 Sometimes this is not obvious (eg if the feature vector is infinite dimensional, e.g. the exponentiated quadratic kernel in the last slide)
- 2 The feature map is not unique.

2 A direct property of the function: **positive definiteness**.

Positive definite functions

Definition (Positive definite functions)

A symmetric function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is **positive definite** if
 $\forall n \geq 1, \forall (a_1, \dots, a_n) \in \mathbb{R}^n, \forall (x_1, \dots, x_n) \in \mathcal{X}^n,$

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \geq 0.$$

The function $k(\cdot, \cdot)$ is **strictly positive definite** if for mutually distinct x_i , the equality holds only when all the a_i are zero.

Kernels are positive definite

Theorem

Let \mathcal{H} be a Hilbert space, \mathcal{X} a non-empty set and $\phi : \mathcal{X} \rightarrow \mathcal{H}$. Then $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} =: k(x, y)$ is positive definite.

Proof.

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) &= \sum_{i=1}^n \sum_{j=1}^n \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^n a_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0.\end{aligned}$$

Reverse also holds: positive definite $k(x, x')$ is inner product in a unique \mathcal{H} (**Moore-Aronson**: coming later!). □

Sum of kernels is a kernel

Proof by positive definiteness:

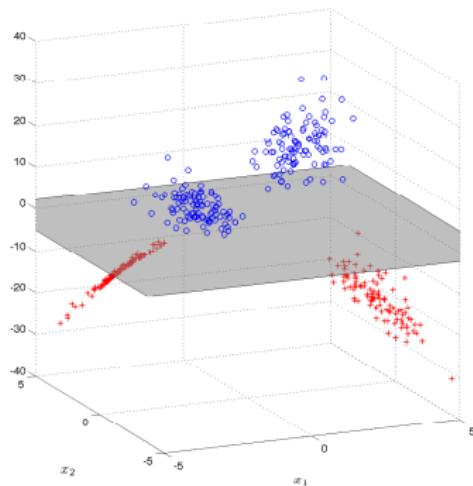
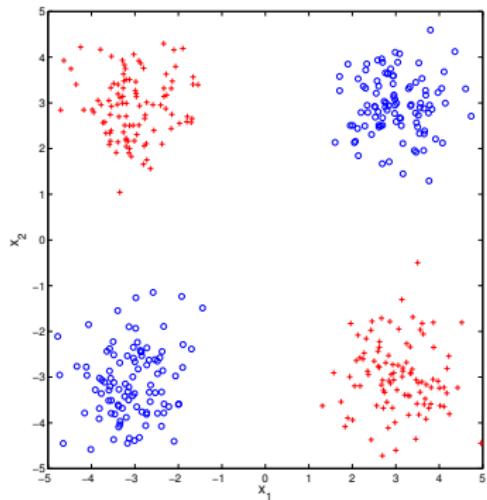
Consider two kernels $k_1(x, x')$ and $k_2(x, x')$. Then

$$\begin{aligned} & \sum_{i=1}^n \sum_{j=1}^n a_i a_j [k_1(x_i, x_j) + k_2(x_i, x_j)] \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j k_1(x_i, x_j) + \sum_{i=1}^n \sum_{j=1}^n a_i a_j k_2(x_i, x_j) \\ &\geq 0 \end{aligned}$$

The reproducing kernel Hilbert space

First example: finite space, polynomial features

Reminder: XOR example:



Example: finite space, polynomial features

Reminder: Feature space from XOR motivating example:

$$\begin{aligned}\phi : \mathbb{R}^2 &\rightarrow \mathbb{R}^3 \\ x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &\mapsto \phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix},\end{aligned}$$

with kernel

$$k(x, y) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}^\top \begin{bmatrix} y_1 \\ y_2 \\ y_1 y_2 \end{bmatrix}$$

(the standard inner product in \mathbb{R}^3 between features). Denote this feature space by \mathcal{H} .

Example: finite space, polynomial features

Define a **linear function** of the inputs x_1, x_2 , and their product $x_1 x_2$,

$$f(x) = f_1 x_1 + f_2 x_2 + f_3 x_1 x_2.$$

f in a space of functions mapping from $\mathcal{X} = \mathbb{R}^2$ to \mathbb{R} . Equivalent representation for f ,

$$f(\cdot) = \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix}^\top.$$

$f(\cdot)$ refers to the function as an object (here as a **vector** in \mathbb{R}^3)

$f(x) \in \mathbb{R}$ is function evaluated at a point (a **real number**).

$$f(x) = f(\cdot)^\top \phi(x) = \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}}$$

Evaluation of f at x is an inner product in feature space (here standard inner product in \mathbb{R}^3)

\mathcal{H} is a space of functions mapping \mathbb{R}^2 to \mathbb{R} .

Example: finite space, polynomial features

Define a **linear function** of the inputs x_1, x_2 , and their product $x_1 x_2$,

$$f(x) = f_1 x_1 + f_2 x_2 + f_3 x_1 x_2.$$

f in a space of functions mapping from $\mathcal{X} = \mathbb{R}^2$ to \mathbb{R} . Equivalent representation for f ,

$$f(\cdot) = \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix}^\top.$$

$f(\cdot)$ refers to the function as an object (here as a **vector** in \mathbb{R}^3)

$f(x) \in \mathbb{R}$ is function evaluated at a point (a **real number**).

$$f(x) = f(\cdot)^\top \phi(x) = \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}}$$

Evaluation of f at x is an **inner product in feature space** (here standard inner product in \mathbb{R}^3)

\mathcal{H} is a space of functions mapping \mathbb{R}^2 to \mathbb{R} .

Functions of infinitely many features

Functions are linear combinations of features:

$$f(x) = \langle \mathbf{f}, \phi(x) \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(x) = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{bmatrix}^T \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \\ \vdots \end{bmatrix}$$

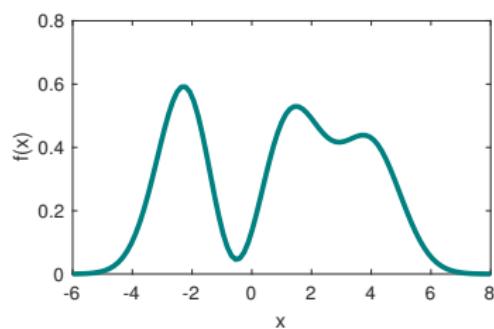
$$k(x, y) = \sum_{\ell=1}^{\infty} \phi_{\ell}(x) \phi_{\ell}(y)$$

$$\mathbf{f}(x) = \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(x) \quad \sum_{\ell=1}^{\infty} f_{\ell}^2 < \infty.$$

Expressing the functions with kernels

Function with **exponentiated quadratic kernel**:

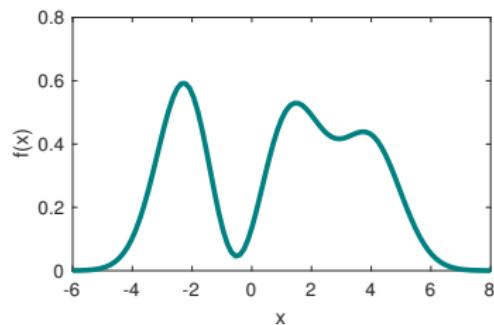
$$\begin{aligned}f(x) &= \sum_{\ell=1}^{\infty} \cancel{f_\ell} \phi_\ell(x) \\&= \sum_{\ell=1}^{\infty} \underbrace{\left(\sum_{i=1}^m \alpha_i \phi_\ell(x_i) \right)}_{\cancel{f_\ell}} \phi_\ell(x) \\&= \left\langle \sum_{i=1}^m \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}} \\&= \sum_{i=1}^m \alpha_i k(x_i, x)\end{aligned}$$



Expressing the functions with kernels

Function with **exponentiated quadratic kernel**:

$$\begin{aligned}f(x) &= \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(x) \\&= \sum_{\ell=1}^{\infty} \underbrace{\left(\sum_{i=1}^m \alpha_i \phi_{\ell}(x_i) \right)}_{f_{\ell}} \phi_{\ell}(x) \\&= \left\langle \sum_{i=1}^m \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}} \\&= \sum_{i=1}^m \alpha_i k(x_i, x)\end{aligned}$$

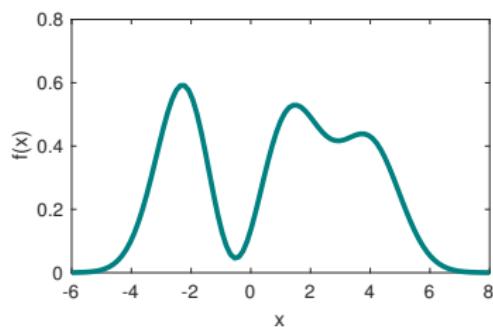


$$f_{\ell} := \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i)$$

Expressing the functions with kernels

Function with **exponentiated quadratic kernel**:

$$\begin{aligned}f(x) &= \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(x) \\&= \sum_{\ell=1}^{\infty} \underbrace{\left(\sum_{i=1}^m \alpha_i \phi_{\ell}(x_i) \right)}_{f_{\ell}} \phi_{\ell}(x) \\&= \left\langle \sum_{i=1}^m \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}} \\&= \sum_{i=1}^m \alpha_i k(x_i, x)\end{aligned}$$

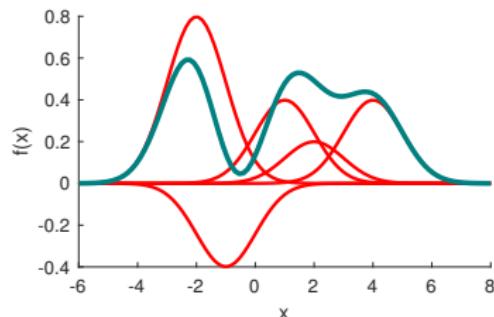


$$f_{\ell} := \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i)$$

Expressing the functions with kernels

Function with **exponentiated quadratic kernel**:

$$\begin{aligned} f(x) &= \sum_{\ell=1}^{\infty} \textcolor{teal}{f}_{\ell} \phi_{\ell}(x) \\ &= \sum_{\ell=1}^{\infty} \left(\underbrace{\sum_{i=1}^m \alpha_i \phi_{\ell}(x_i)}_{\textcolor{teal}{f}_{\ell}} \right) \phi_{\ell}(x) \\ &= \left\langle \sum_{i=1}^m \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^m \alpha_i k(x_i, x) \end{aligned}$$



$$\textcolor{teal}{f}_{\ell} := \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i)$$

Function of **infinitely many features** expressed using m coefficients.

The feature map is also a function

On previous page,

$$f(x) := \sum_{i=1}^m \alpha_i k(x_i, x) = \langle \mathbf{f}(\cdot), \phi(x) \rangle_{\mathcal{H}} \quad \text{where } \mathbf{f}_{\ell} = \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i).$$

What if $m = 1$ and $\alpha_1 = 1$?

Then

$$f(x) = k(x_1, x) = \left\langle \underbrace{k(x_1, \cdot)}_{f(\cdot)}, \phi(x) \right\rangle_{\mathcal{H}}$$

The feature map is also a function

On previous page,

$$f(x) := \sum_{i=1}^m \alpha_i k(x_i, x) = \langle \mathbf{f}(\cdot), \phi(x) \rangle_{\mathcal{H}} \quad \text{where} \quad \mathbf{f}_{\ell} = \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i).$$

What if $m = 1$ and $\alpha_1 = 1$?

Then

$$f(x) = k(\mathbf{x}_1, \mathbf{x}) = \left\langle \underbrace{k(\mathbf{x}_1, \cdot)}_{f(\cdot)}, \phi(\mathbf{x}) \right\rangle_{\mathcal{H}}$$

The feature map is also a function

On previous page,

$$f(\mathbf{x}) := \sum_{i=1}^m \alpha_i k(\mathbf{x}_i, \mathbf{x}) = \langle \mathbf{f}(\cdot), \phi(\mathbf{x}) \rangle_{\mathcal{H}} \quad \text{where } \mathbf{f}_{\ell} = \sum_{i=1}^m \alpha_i \phi_{\ell}(\mathbf{x}_i).$$

What if $m = 1$ and $\alpha_1 = 1$?

Then

$$\begin{aligned} f(\mathbf{x}) &= k(\mathbf{x}_1, \mathbf{x}) = \left\langle \underbrace{k(\mathbf{x}_1, \cdot)}_{\mathbf{f}(\cdot)}, \phi(\mathbf{x}) \right\rangle_{\mathcal{H}} \\ &= \langle k(\mathbf{x}, \cdot), \phi(\mathbf{x}_1) \rangle_{\mathcal{H}} \end{aligned}$$

....so the feature map is a (very simple) function!

We can write without ambiguity

$$k(\mathbf{x}, \mathbf{y}) = \langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{y}) \rangle_{\mathcal{H}}.$$

The feature map is also a function

On previous page,

$$f(\mathbf{x}) := \sum_{i=1}^m \alpha_i k(\mathbf{x}_i, \mathbf{x}) = \langle \mathbf{f}(\cdot), \phi(\mathbf{x}) \rangle_{\mathcal{H}} \quad \text{where } \mathbf{f}_{\ell} = \sum_{i=1}^m \alpha_i \phi_{\ell}(\mathbf{x}_i).$$

What if $m = 1$ and $\alpha_1 = 1$?

Then

$$\begin{aligned} f(\mathbf{x}) &= k(\mathbf{x}_1, \mathbf{x}) = \left\langle \underbrace{k(\mathbf{x}_1, \cdot)}_{\mathbf{f}(\cdot)}, \phi(\mathbf{x}) \right\rangle_{\mathcal{H}} \\ &= \langle k(\mathbf{x}, \cdot), \phi(\mathbf{x}_1) \rangle_{\mathcal{H}} \end{aligned}$$

....so the feature map is a (very simple) function!

We can write without ambiguity

$$k(\mathbf{x}, \mathbf{y}) = \langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{y}) \rangle_{\mathcal{H}}.$$

The reproducing property

This example illustrates the two defining features of an RKHS:

- **The reproducing property: (kernel trick)**

$$\forall x \in \mathcal{X}, \forall f(\cdot) \in \mathcal{H}, \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$$

...or use shorter notation $\langle f, \phi(x) \rangle_{\mathcal{H}}$.

- The feature map of every point is a function: $k(\cdot, x) = \phi(x) \in \mathcal{H}$ for any $x \in \mathcal{X}$, and

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}}.$$

Understanding smoothness in the RKHS

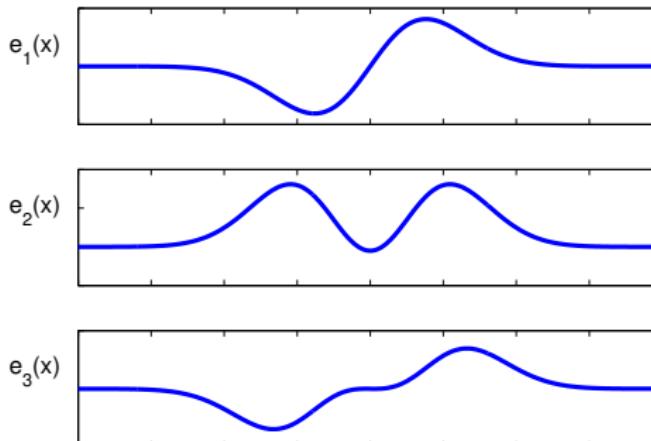
Smoothness in RKHS with exp. quad. kernel

Reminder, **exponentiated quadratic kernel**,

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) = \sum_{\ell=1}^{\infty} \underbrace{(\sqrt{\lambda_\ell} e_\ell(x))}_{\phi_\ell(x)} \underbrace{(\sqrt{\lambda_\ell} e_\ell(x'))}_{\phi_\ell(x')}$$

$$\lambda_\ell e_\ell(x) = \int k(x, x') e_\ell(x') p(x') dx',$$

$$p(x) = \mathcal{N}(0, \sigma^2).$$

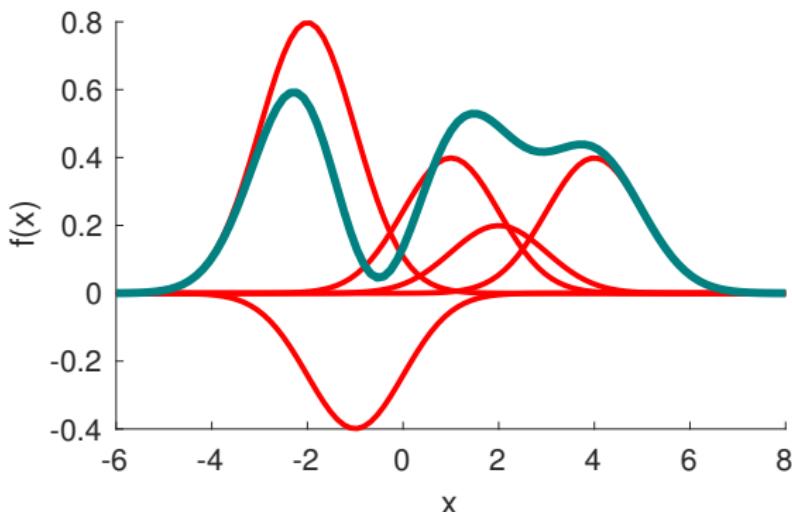


Smoothness in RKHS with exp. quad. kernel

RKHS function, exponentiated quadratic kernel:

$$f(x) := \sum_{i=1}^m \alpha_i k(x_i, x) = \sum_{\ell=1}^{\infty} f_\ell \underbrace{\left[\sqrt{\lambda_\ell} e_\ell(x) \right]}_{\phi_\ell(x)}$$

where $f_\ell = \sum_{i=1}^m \alpha_i \sqrt{\lambda_\ell} e_\ell(x_i)$.



NOTE that this enforces smoothing:
 λ_ℓ decay as e_ℓ become rougher,
 f_ℓ decay since $\sum_\ell f_\ell^2 < \infty$.

Second (infinite) example: fourier series

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

Fourier series:

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

using the orthonormal basis on $[-\pi, \pi]$,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(\imath \ell x) \overline{\exp(\imath m x)} dx = \begin{cases} 1 & \ell = m, \\ 0 & \ell \neq m. \end{cases}$$

Example: “top hat” function,

$$f(x) = \begin{cases} 1 & |x| < T, \\ 0 & T \leq |x| < \pi. \end{cases}$$

$$\hat{f}_{\ell} := \frac{\sin(\ell T)}{\ell \pi} \quad f(x) = \sum_{\ell=0}^{\infty} 2\hat{f}_{\ell} \cos(\ell x).$$

Second (infinite) example: fourier series

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

Fourier series:

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

using the orthonormal basis on $[-\pi, \pi]$,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(\imath \ell x) \overline{\exp(\imath m x)} dx = \begin{cases} 1 & \ell = m, \\ 0 & \ell \neq m. \end{cases}$$

Example: “top hat” function,

$$f(x) = \begin{cases} 1 & |x| < T, \\ 0 & T \leq |x| < \pi. \end{cases}$$

$$\hat{f}_{\ell} := \frac{\sin(\ell T)}{\ell \pi} \quad f(x) = \sum_{\ell=0}^{\infty} 2\hat{f}_{\ell} \cos(\ell x).$$

Second (infinite) example: fourier series

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

Fourier series:

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

using the orthonormal basis on $[-\pi, \pi]$,

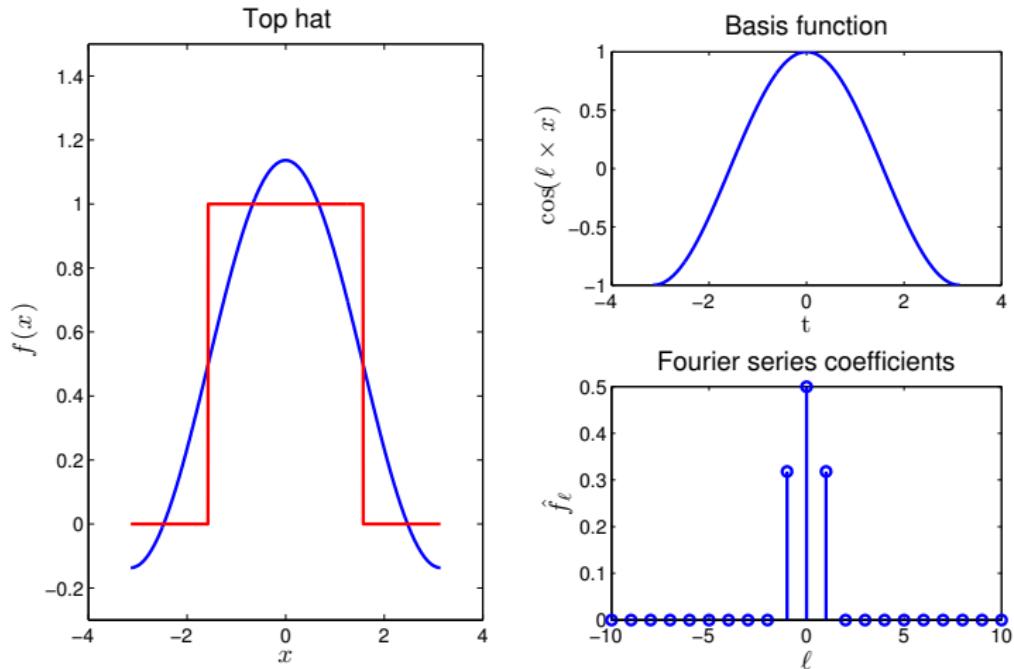
$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(\imath \ell x) \overline{\exp(\imath m x)} dx = \begin{cases} 1 & \ell = m, \\ 0 & \ell \neq m. \end{cases}$$

Example: “top hat” function,

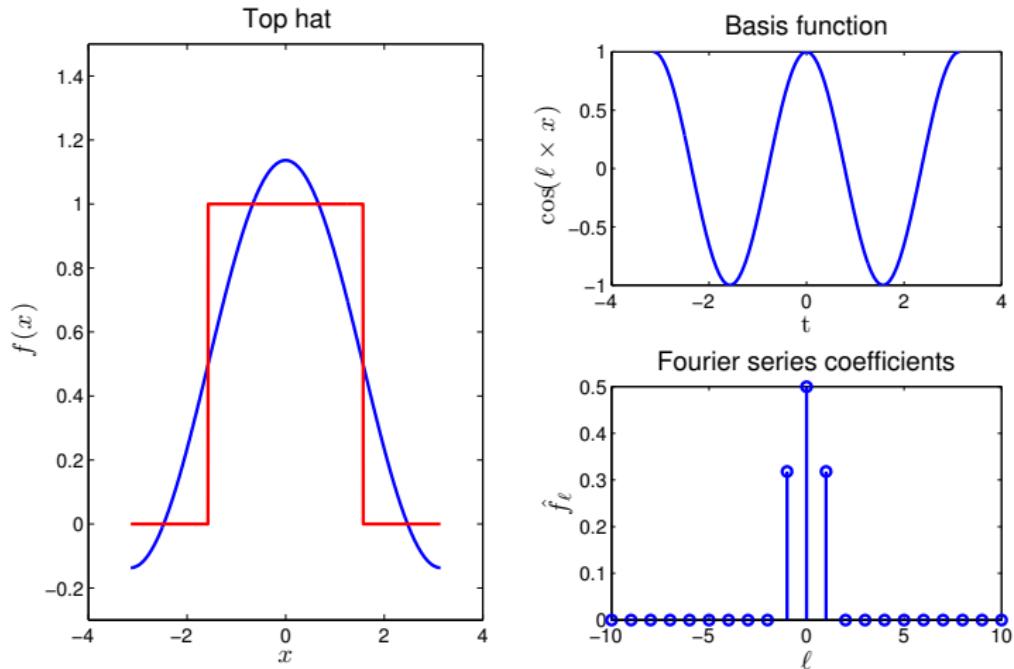
$$f(x) = \begin{cases} 1 & |x| < T, \\ 0 & T \leq |x| < \pi. \end{cases}$$

$$\hat{f}_{\ell} := \frac{\sin(\ell T)}{\ell \pi} \quad f(x) = \sum_{\ell=0}^{\infty} 2\hat{f}_{\ell} \cos(\ell x).$$

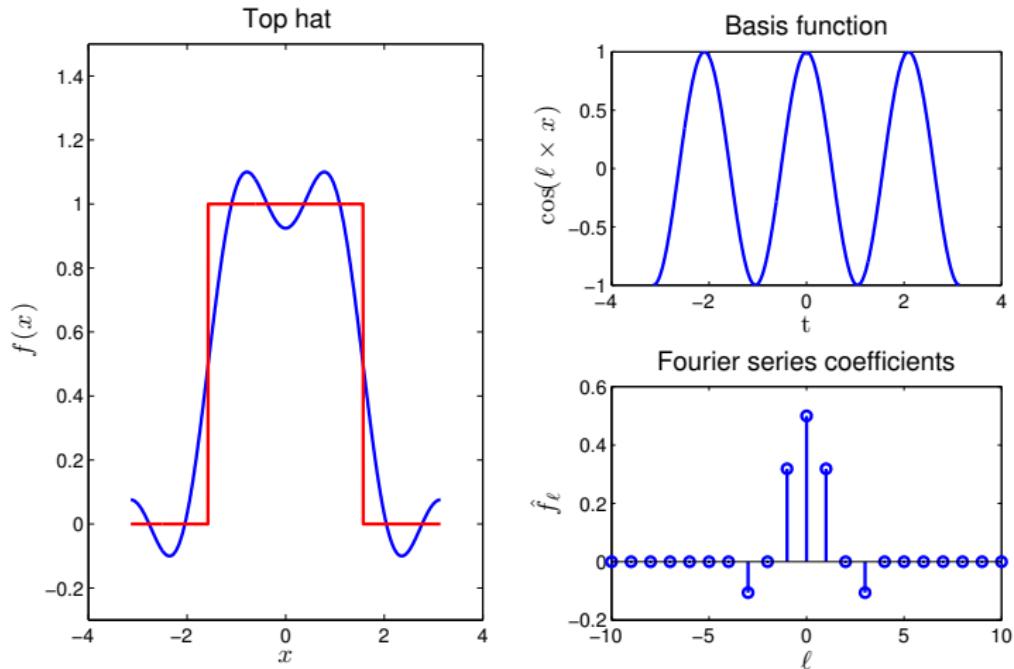
Fourier series for top hat function



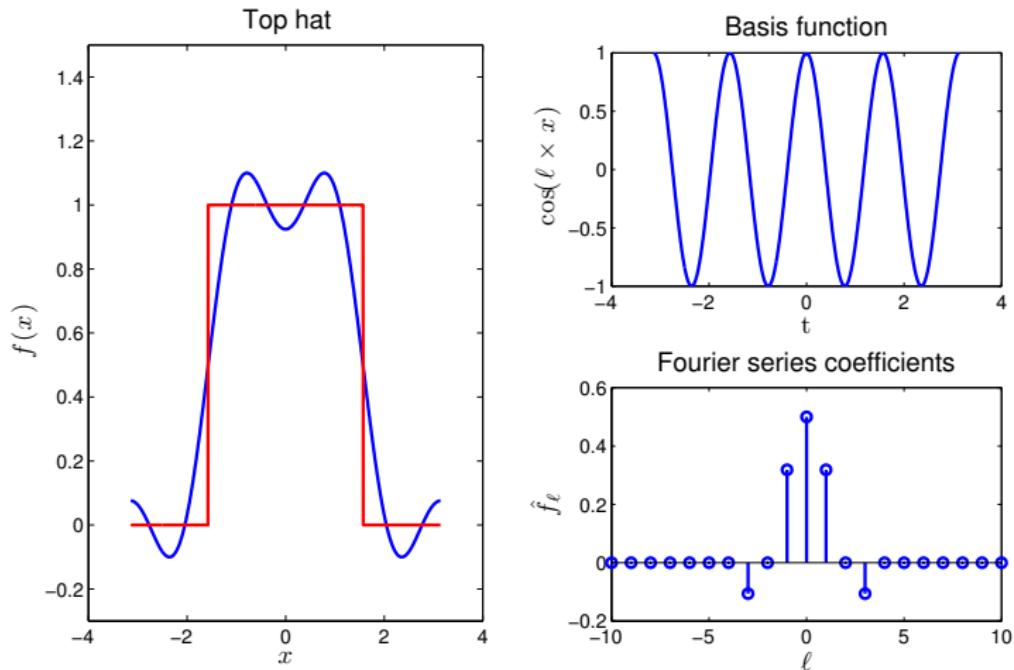
Fourier series for top hat function



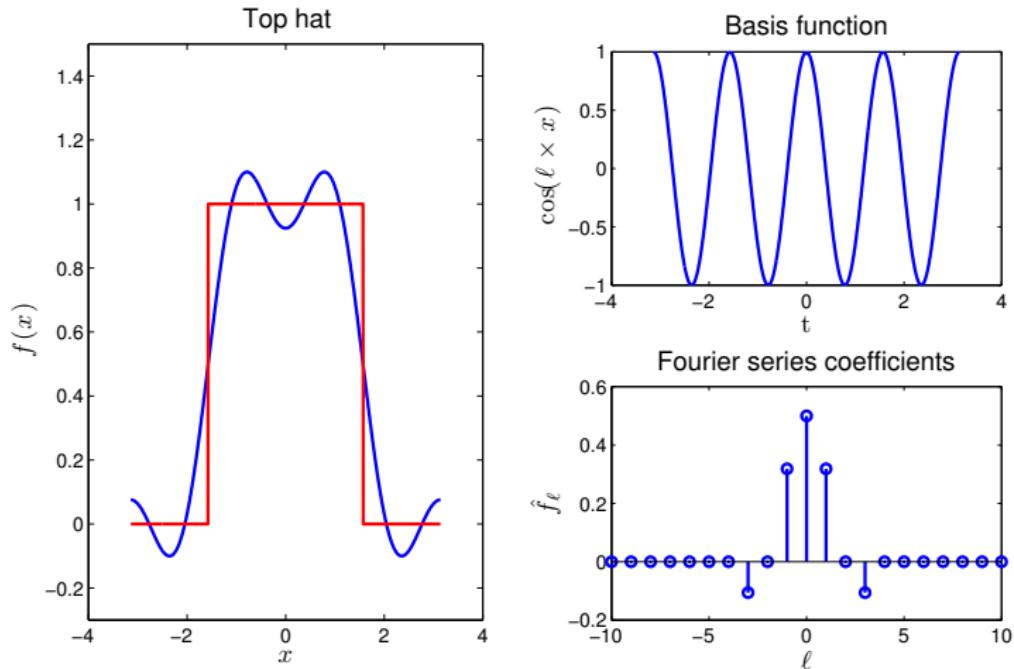
Fourier series for top hat function



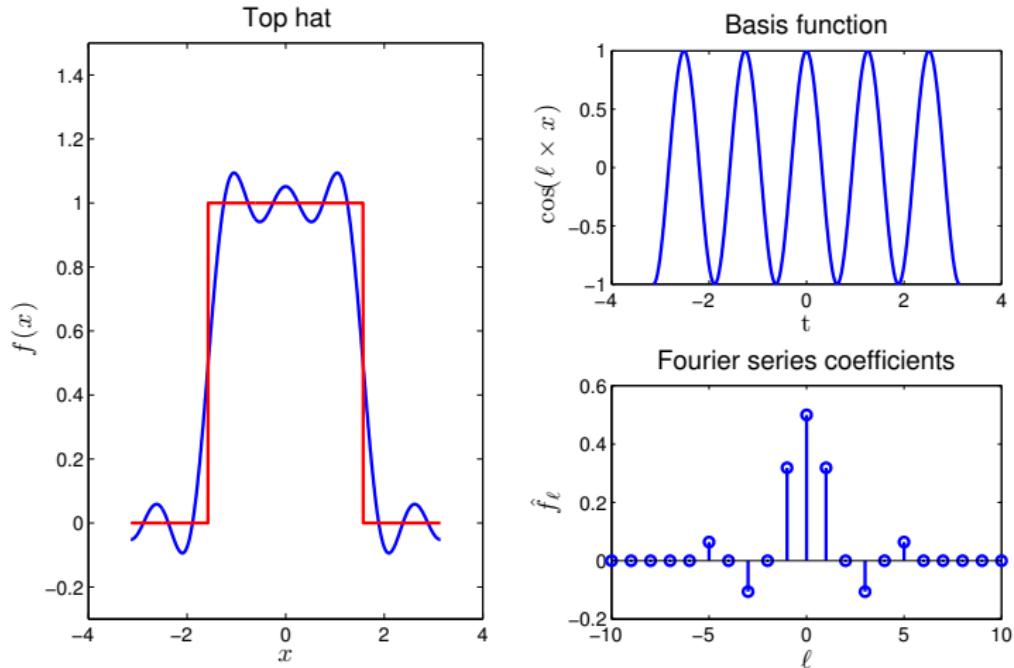
Fourier series for top hat function



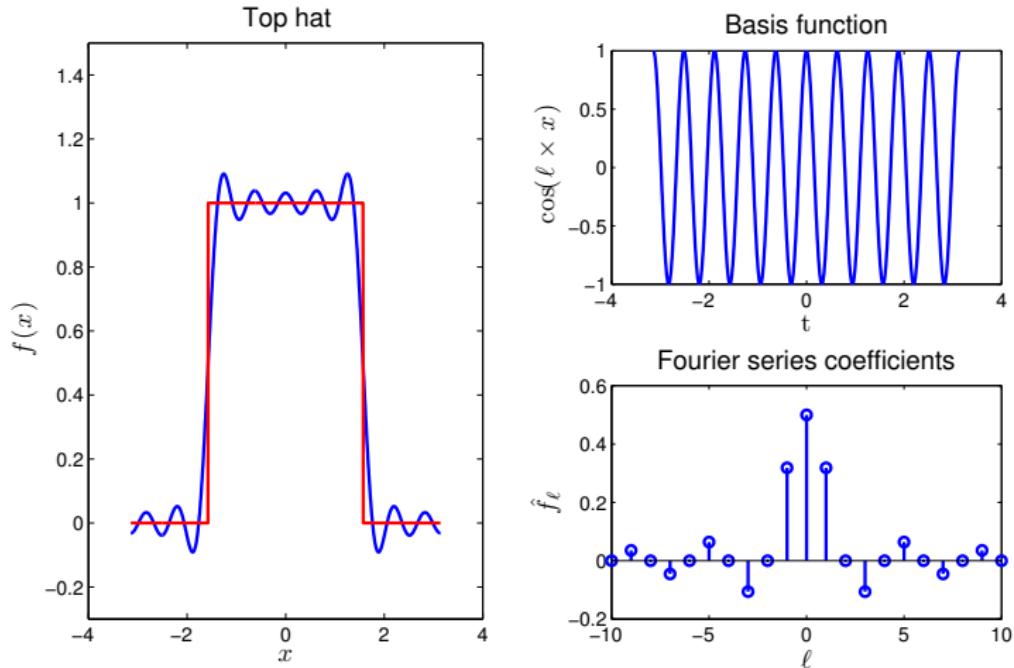
Fourier series for top hat function



Fourier series for top hat function



Fourier series for top hat function



Fourier series for kernel function

Assume kernel translation invariant,

$$k(x, y) = k(x - y),$$

Fourier series representation of k

$$\begin{aligned} k(x - y) &= \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp(\imath \ell(x - y)) \\ &= \sum_{\ell=-\infty}^{\infty} \left[\underbrace{\sqrt{\hat{k}_\ell} \exp(\imath \ell x)}_{e_\ell(x)} \right] \left[\overline{\sqrt{\hat{k}_\ell} \exp(-\imath \ell y)} \right]. \end{aligned}$$

Example: Jacobi theta kernel:

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

ϑ is Jacobi theta function, close to Gaussian when σ^2 much narrower than $[-\pi, \pi]$.

Fourier series for kernel function

Assume kernel translation invariant,

$$k(x, y) = k(x - y),$$

Fourier series representation of k

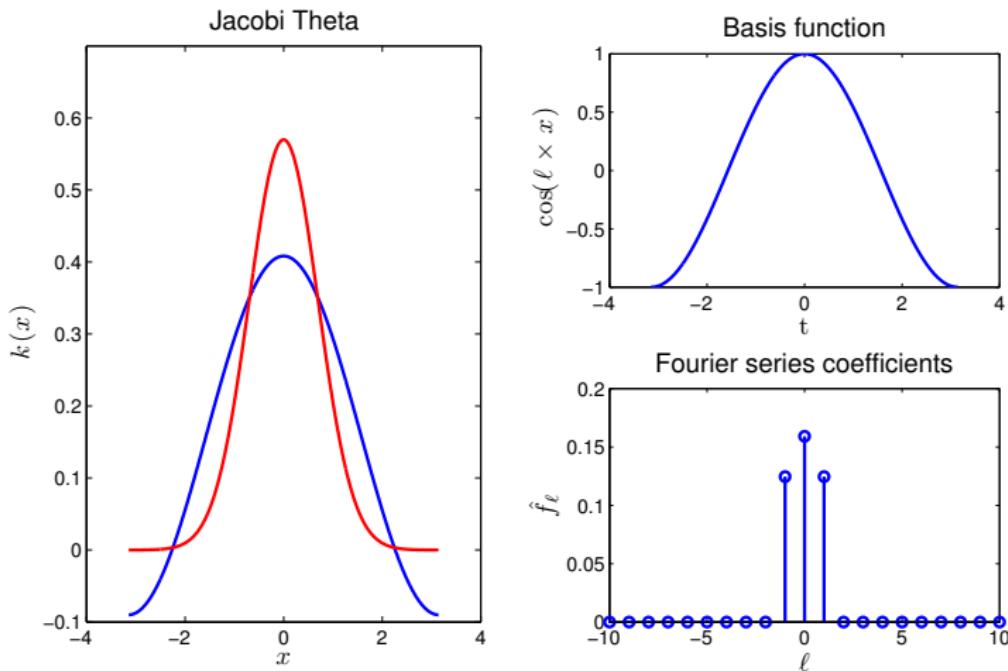
$$\begin{aligned} k(x - y) &= \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp(\imath \ell(x - y)) \\ &= \sum_{\ell=-\infty}^{\infty} \left[\underbrace{\sqrt{\hat{k}_\ell} \exp(\imath \ell x)}_{e_\ell(x)} \right] \left[\overbrace{\sqrt{\hat{k}_\ell} \exp(-\imath \ell y)}^{\overline{e_\ell(y)}} \right]. \end{aligned}$$

Example: **Jacobi theta kernel:**

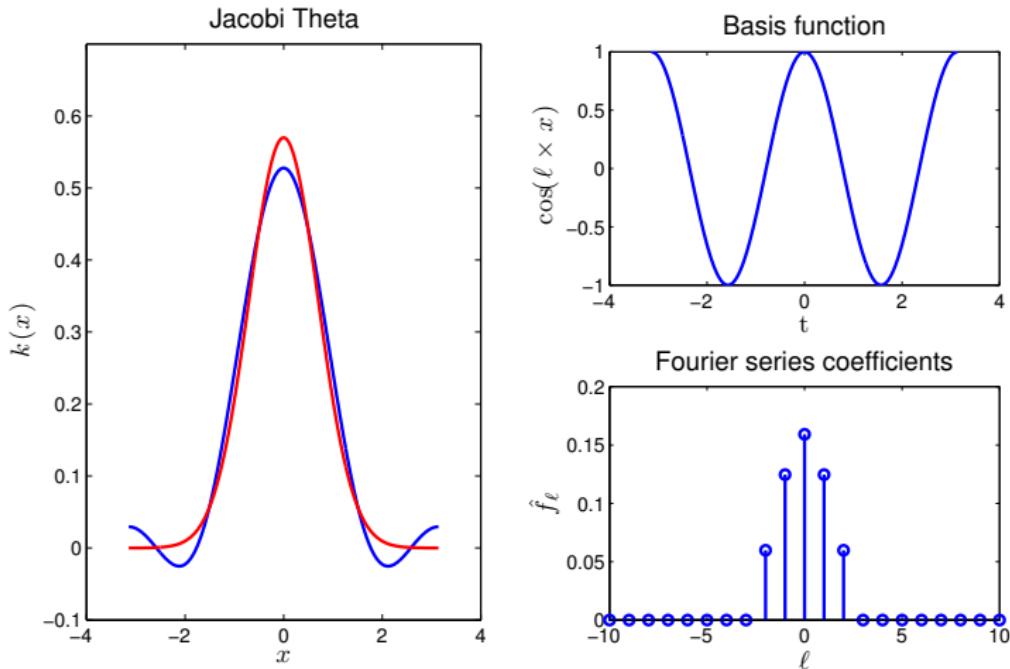
$$k(x - y) = \frac{1}{2\pi} \vartheta \left(\frac{(x - y)}{2\pi}, \frac{\imath \sigma^2}{2\pi} \right), \quad \hat{k}_\ell = \frac{1}{2\pi} \exp \left(\frac{-\sigma^2 \ell^2}{2} \right).$$

ϑ is Jacobi theta function, close to Gaussian when σ^2 much narrower than $[-\pi, \pi]$.

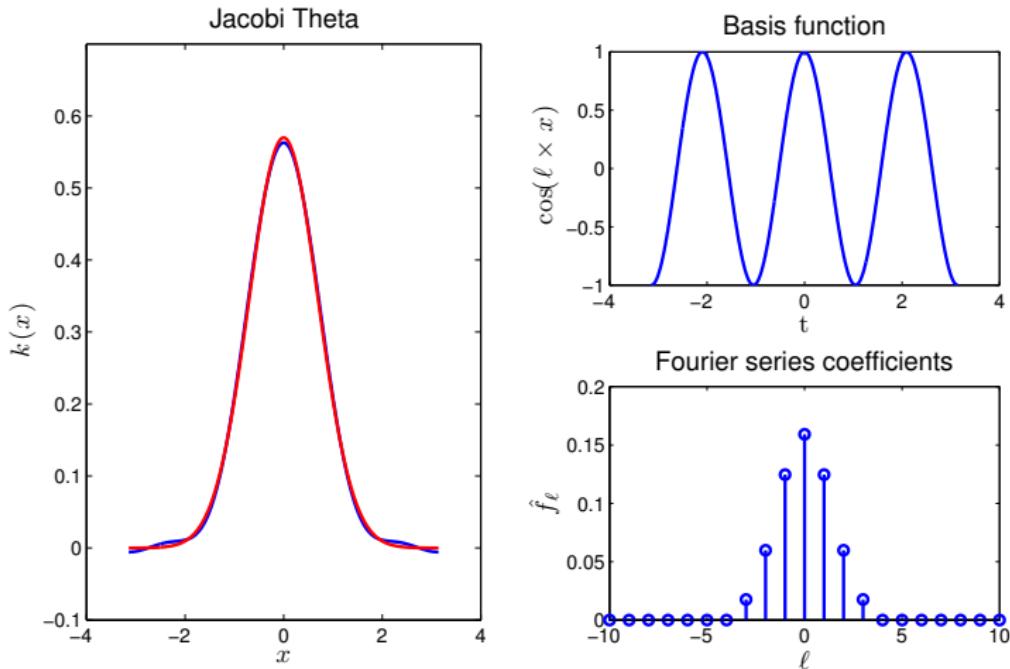
Fourier series for Gaussian-spectrum kernel



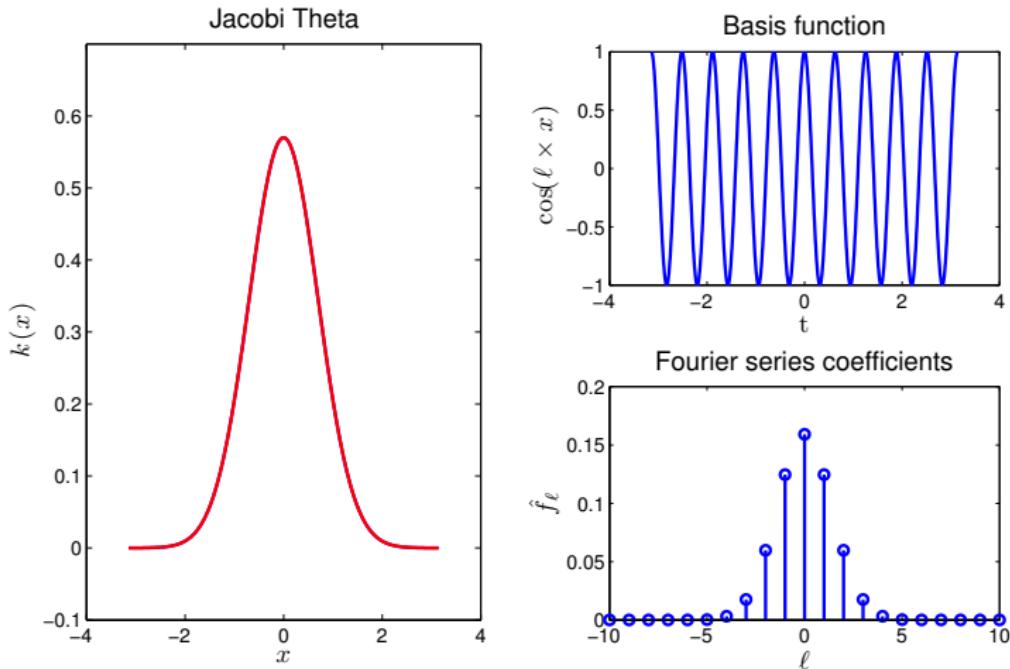
Fourier series for Gaussian-spectrum kernel



Fourier series for Gaussian-spectrum kernel



Fourier series for Gaussian-spectrum kernel



RKHS via fourier series

Recall standard dot product in L_2 :

$$\begin{aligned}\langle f, g \rangle_{L_2} &= \left\langle \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell x), \sum_{m=-\infty}^{\infty} \overline{\hat{g}_m \exp(\imath mx)} \right\rangle_{L_2} \\ &= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}_\ell} \langle \exp(\imath \ell x), \exp(-\imath mx) \rangle_{L_2} \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}_\ell}.\end{aligned}$$

Define the dot product in \mathcal{H} to have a *roughness penalty*,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \overline{\hat{g}_\ell}}{\hat{k}_\ell}.$$

RKHS via fourier series

Recall standard dot product in L_2 :

$$\begin{aligned}\langle f, g \rangle_{L_2} &= \left\langle \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell x), \sum_{m=-\infty}^{\infty} \overline{\hat{g}_m \exp(\imath mx)} \right\rangle_{L_2} \\ &= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}_\ell} \langle \exp(\imath \ell x), \exp(-\imath mx) \rangle_{L_2} \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}_\ell}.\end{aligned}$$

Define the **dot product** in \mathcal{H} to have a **roughness penalty**,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \overline{\hat{g}_\ell}}{\hat{k}_\ell}.$$

Roughness penalty explained

The squared norm of a function f in \mathcal{H} **enforces smoothness**:

$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \bar{\hat{f}}_l}{\hat{k}_l} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_l|^2}{\hat{k}_l}.$$

If \hat{k}_l decays fast, then so must \hat{f}_l if we want $\|f\|_{\mathcal{H}}^2 < \infty$.

Recall $f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l (\cos(\ell x) + i \sin(\ell x))$.

Question: is the top hat function in the “Gaussian spectrum” RKHS?

Warning: need stronger conditions on kernel than L_2 convergence: **Mercer’s theorem**.

Roughness penalty explained

The squared norm of a function f in \mathcal{H} **enforces smoothness**:

$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \bar{\hat{f}}_l}{\hat{k}_l} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_l|^2}{\hat{k}_l}.$$

If \hat{k}_l decays fast, then so must \hat{f}_l if we want $\|f\|_{\mathcal{H}}^2 < \infty$.

Recall $f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l (\cos(\ell x) + i \sin(\ell x))$.

Question: is the top hat function in the “Gaussian spectrum” RKHS?

Warning: need stronger conditions on kernel than L_2 convergence: **Mercer’s theorem**.

Roughness penalty explained

The squared norm of a function f in \mathcal{H} **enforces smoothness**:

$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \bar{\hat{f}}_l}{\hat{k}_l} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_l|^2}{\hat{k}_l}.$$

If \hat{k}_l decays fast, then so must \hat{f}_l if we want $\|f\|_{\mathcal{H}}^2 < \infty$.

Recall $f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l (\cos(\ell x) + i \sin(\ell x))$.

Question: is the top hat function in the “Gaussian spectrum” RKHS?

Warning: need stronger conditions on kernel than L_2 convergence: **Mercer’s theorem**.

Feature map and reproducing property

Reproducing property: define a function

$$g(x) := k(x - z) = \sum_{\ell=-\infty}^{\infty} \exp(\imath \ell x) \underbrace{\hat{k}_\ell \exp(-\imath \ell z)}_{\hat{g}_\ell}$$

Then for a function $f(\cdot) \in \mathcal{H}$,

$$\langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}}$$

$$\begin{aligned} & \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \underbrace{\frac{\hat{k}_\ell \exp(\imath \ell z)}{\hat{k}_\ell}}_{\hat{g}_\ell} \\ & \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell z) = f(z). \end{aligned}$$

Feature map and reproducing property

Reproducing property: define a function

$$g(x) := k(x - z) = \sum_{\ell=-\infty}^{\infty} \exp(\imath \ell x) \underbrace{\hat{k}_\ell \exp(-\imath \ell z)}_{\hat{g}_\ell}$$

Then for a function $f(\cdot) \in \mathcal{H}$,

$$\begin{aligned} \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} &= \langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}} \\ &= \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell}{\hat{k}_\ell} \overbrace{\frac{\hat{k}_\ell \exp(\imath \ell z)}{\hat{k}_\ell}}^{\hat{g}_\ell} \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell z) = f(z). \end{aligned}$$

Feature map and reproducing property

Reproducing property: define a function

$$g(x) := k(x - z) = \sum_{\ell=-\infty}^{\infty} \exp(\imath \ell x) \underbrace{\hat{k}_\ell \exp(-\imath \ell z)}_{\hat{g}_\ell}$$

Then for a function $f(\cdot) \in \mathcal{H}$,

$$\begin{aligned} \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} &= \langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}} \\ &= \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell}{\hat{k}_\ell} \overbrace{\hat{k}_\ell \exp(\imath \ell z)}^{\hat{g}_\ell} \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell z) = f(z). \end{aligned}$$

Feature map and reproducing property

Reproducing property for the kernel:

You can also show

$$\langle k(\cdot, y), k(\cdot, z) \rangle_{\mathcal{H}} = k(y - z)$$

This is an exercise!

Hint: define a second function

$$f(x) := k(x - y) = \sum_{\ell=-\infty}^{\infty} \exp(\imath \ell x) \underbrace{\hat{k}_\ell}_{\hat{f}_\ell} \exp(-\imath \ell y)$$

Feature map and reproducing property

Reproducing property for the kernel:

You can also show

$$\langle k(\cdot, y), k(\cdot, z) \rangle_{\mathcal{H}} = k(y - z)$$

This is an exercise!

Hint: define a second function

$$f(x) := k(x - y) = \sum_{\ell=-\infty}^{\infty} \exp(\imath \ell x) \underbrace{\hat{k}_\ell}_{\hat{f}_\ell} \exp(-\imath \ell y)$$

Link back to original RKHS function definition

Original form of a function in the RKHS was

(detail: sum now from $-\infty$ to ∞ , complex conjugate)

$$f(x) = \sum_{\ell=-\infty}^{\infty} \textcolor{blue}{f}_\ell \overline{\phi_\ell(x)} = \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}}.$$

We've defined the RKHS dot product as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \overline{\hat{g}_l}}{\hat{k}_l} \quad \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \left(\overline{\hat{k}_\ell \exp(-\imath \ell z)} \right)}{\hat{k}_\ell}$$

Link back to original RKHS function definition

Original form of a function in the RKHS was

(detail: sum now from $-\infty$ to ∞ , complex conjugate)

$$f(x) = \sum_{\ell=-\infty}^{\infty} \textcolor{blue}{f}_\ell \overline{\phi_\ell(x)} = \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}}.$$

We've defined the RKHS dot product as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \overline{\hat{g}_l}}{\hat{k}_l}$$

$$\langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \left(\hat{k}_\ell \exp(-\imath \ell z) \right)}{\left(\sqrt{\hat{k}_\ell} \right)^2}$$

Link back to original RKHS function definition

Original form of a function in the RKHS was

(detail: sum now from $-\infty$ to ∞ , complex conjugate)

$$f(x) = \sum_{\ell=-\infty}^{\infty} \textcolor{blue}{f}_\ell \overline{\phi_\ell(x)} = \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}}.$$

We've defined the RKHS dot product as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \overline{\hat{g}_l}}{\hat{k}_l} \quad \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \left(\hat{k}_\ell \exp(-\imath \ell z) \right)}{\left(\sqrt{\hat{k}_\ell} \right)^2}$$

By inspection

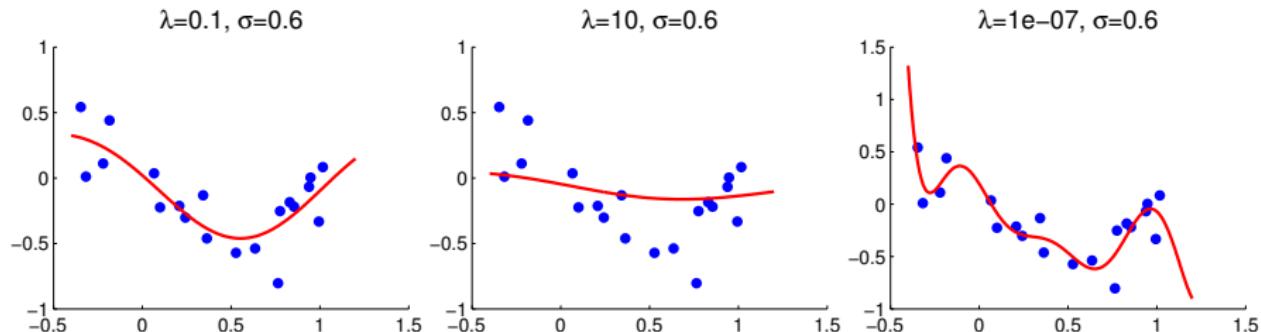
$$\textcolor{blue}{f}_\ell = \hat{f}_\ell / \sqrt{\hat{k}_\ell} \quad \phi_\ell(x) = \sqrt{\hat{k}_\ell} \exp(-\imath \ell x).$$

Main message

Small RKHS norm results in **smooth functions**.

E.g. kernel ridge regression with **exponentiated quadratic** kernel:

$$f^* = \arg \min_{f \in \mathcal{H}} \left(\sum_{i=1}^n (y_i - \langle f, \phi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$



Some reproducing kernel Hilbert space theory

Reproducing kernel Hilbert space (1)

Definition

\mathcal{H} a Hilbert space of \mathbb{R} -valued functions on non-empty set \mathcal{X} . A function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a **reproducing kernel** of \mathcal{H} , and \mathcal{H} is a **reproducing kernel Hilbert space**, if

- $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (the reproducing property).

In particular, for any $x, y \in \mathcal{X}$,

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}. \quad (2)$$

Original definition: kernel an inner product between feature maps.
Then $\phi(x) = k(\cdot, x)$ a valid feature map.

Reproducing kernel Hilbert space (2)

Another RKHS definition:

Define δ_x to be the operator of evaluation at x , i.e.

$$\delta_x f = f(x) \quad \forall f \in \mathcal{H}, x \in \mathcal{X}.$$

Definition (Reproducing kernel Hilbert space)

\mathcal{H} is an RKHS if the evaluation operator δ_x is **bounded**: $\forall x \in \mathcal{X}$ there exists $\lambda_x \geq 0$ such that for all $f \in \mathcal{H}$,

$$|f(x)| = |\delta_x f| \leq \lambda_x \|f\|_{\mathcal{H}}$$

\implies two functions identical in RHKS norm agree at every point:

$$|f(x) - g(x)| = |\delta_x(f - g)| \leq \lambda_x \|f - g\|_{\mathcal{H}} \quad \forall f, g \in \mathcal{H}.$$

RKHS definitions equivalent

Theorem (Reproducing kernel equivalent to bounded δ_x)

\mathcal{H} is a reproducing kernel Hilbert space (i.e., its evaluation operators δ_x are bounded linear operators), if and only if \mathcal{H} has a reproducing kernel.

Proof: If \mathcal{H} has a reproducing kernel $\implies \delta_x$ bounded

$$\begin{aligned} |\delta_x[f]| &= |f(x)| \\ &= |\langle f, k(\cdot, x) \rangle_{\mathcal{H}}| \\ &\leq \|k(\cdot, x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \langle k(\cdot, x), k(\cdot, x) \rangle_{\mathcal{H}}^{1/2} \|f\|_{\mathcal{H}} \\ &= k(x, x)^{1/2} \|f\|_{\mathcal{H}} \end{aligned}$$

Cauchy-Schwarz in 3rd line . Consequently, $\delta_x : \mathcal{F} \rightarrow \mathbb{R}$ bounded with $\lambda_x = k(x, x)^{1/2}$.

RKHS definitions equivalent

Proof: δ_x bounded $\implies \mathcal{H}$ has a reproducing kernel

We use...

Theorem

(Riesz representation) In a Hilbert space \mathcal{H} , all bounded linear functionals are of the form $\langle \cdot, g \rangle_{\mathcal{H}}$, for some $g \in \mathcal{H}$.

If $\delta_x : \mathcal{F} \rightarrow \mathbb{R}$ is a bounded linear functional, by Riesz $\exists f_{\delta_x} \in \mathcal{H}$ such that

$$\delta_x f = \langle f, f_{\delta_x} \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}.$$

Define $k(\cdot, x) = f_{\delta_x}(\cdot)$, $\forall x, x' \in \mathcal{X}$. By its definition, both $k(\cdot, x) = f_{\delta_x}(\cdot) \in \mathcal{H}$ and $\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = \delta_x f = f(x)$. Thus, k is the reproducing kernel.

Moore-Aronszajn Theorem

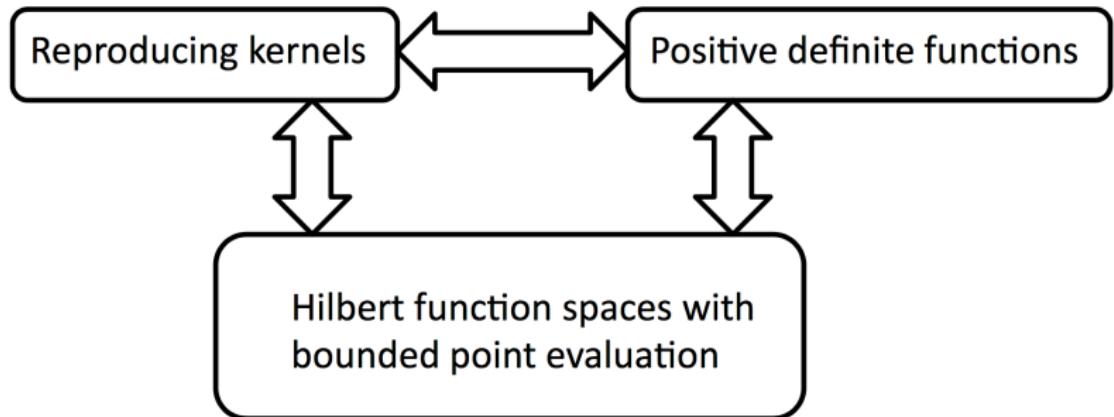
Theorem (Moore-Aronszajn)

Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be positive definite. There is a unique RKHS $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ with reproducing kernel k .

Recall feature map is *not unique* (as we saw earlier):

only kernel is unique.

Main message



Representing and comparing probabilities with kernels: Part 2

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

MLSS Madrid, 2018

Comparing two samples

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



$\sim P$



$\sim Q$

Outline

Two sample testing

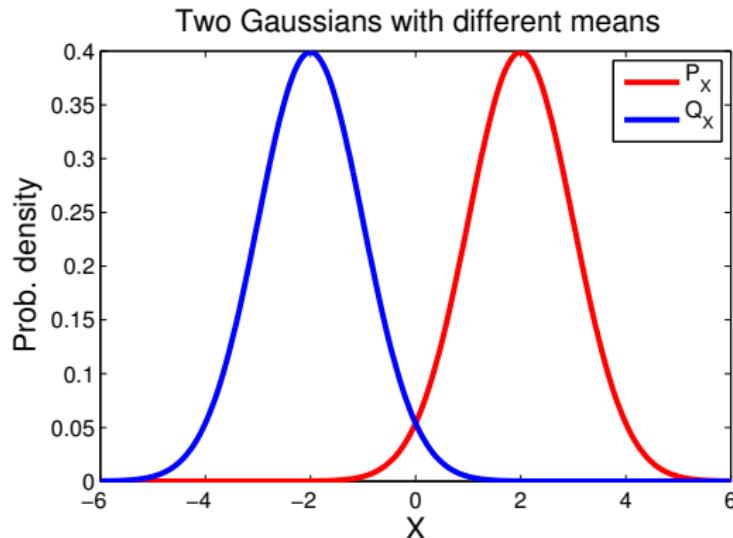
- Test statistic: Maximum Mean Discrepancy (MMD)...
 - ...as a difference in feature means
 - ...as an integral probability metric (*not just a technicality!*)
- Statistical testing with the MMD
- “How to choose the best kernel”

Training GANs with MMD

Maximum Mean Discrepancy

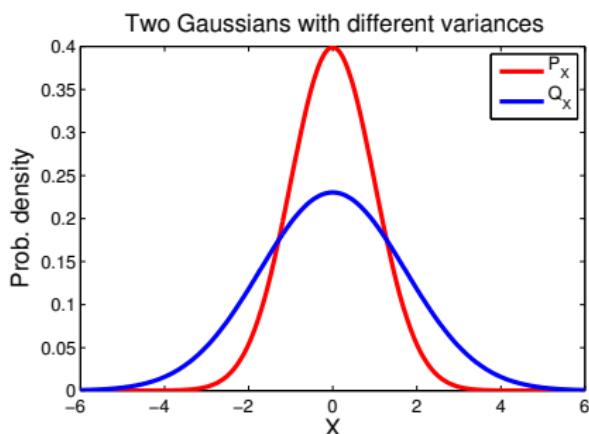
Feature mean difference

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



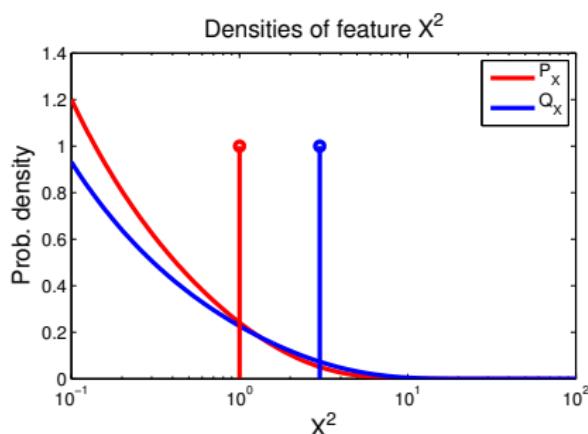
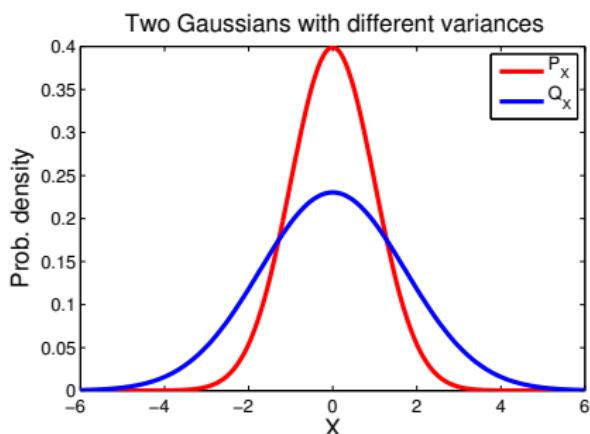
Feature mean difference

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



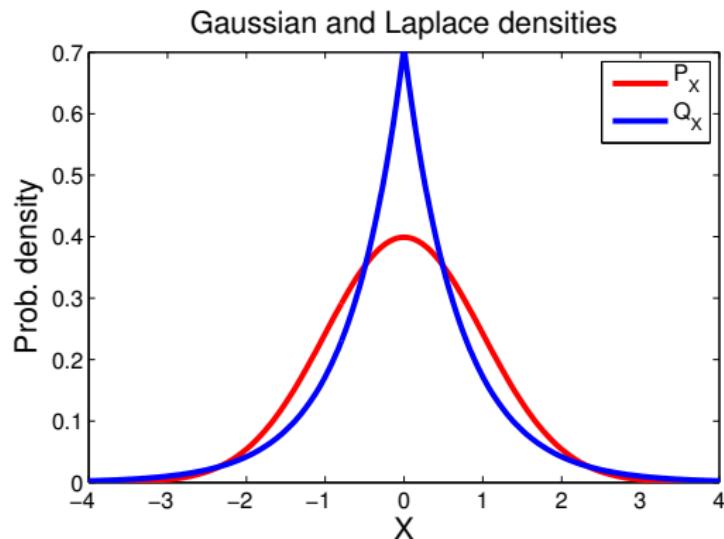
Feature mean difference

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



Feature mean difference

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



Infinitely many features using kernels

Kernels: dot products
of features

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

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

For positive definite k ,

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

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

Infinitely many features using kernels

Kernels: dot products
of features

Exponentiated quadratic kernel

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

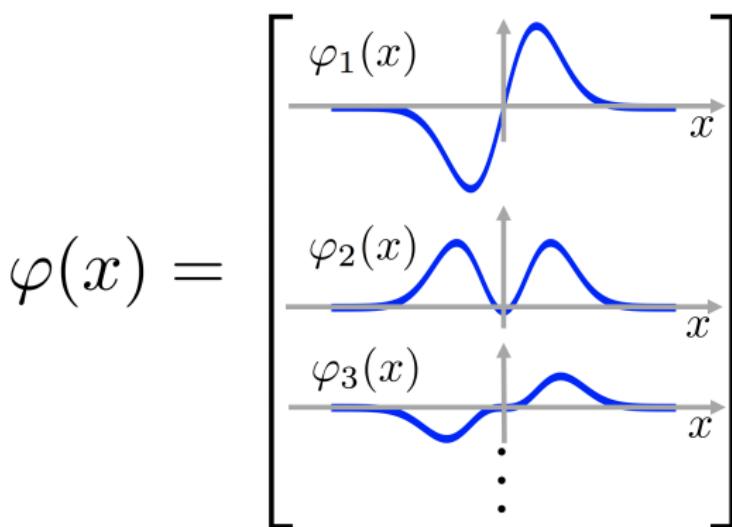
Feature map $\varphi(x) \in \mathcal{F}$,

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

For positive definite k ,

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

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



Infinitely many features of *distributions*

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

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

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

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

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

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

Infinitely many features of *distributions*

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

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

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

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

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

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

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

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

Illustration of MMD

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

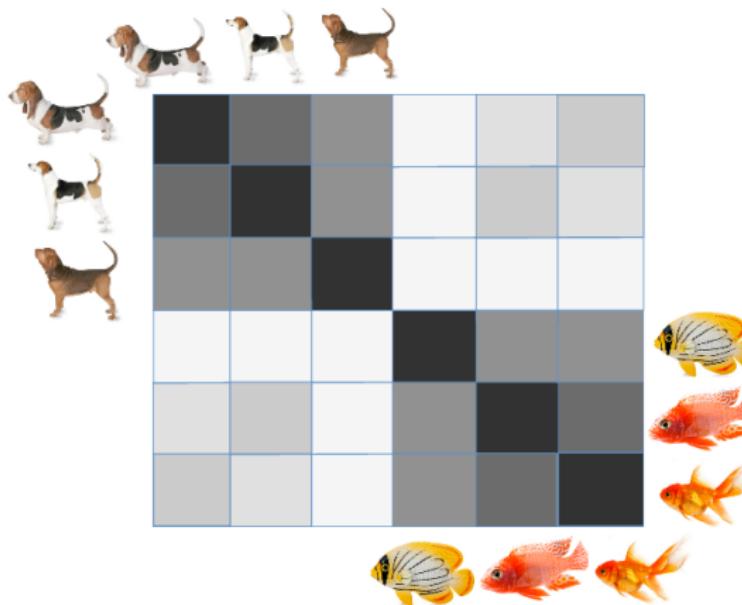
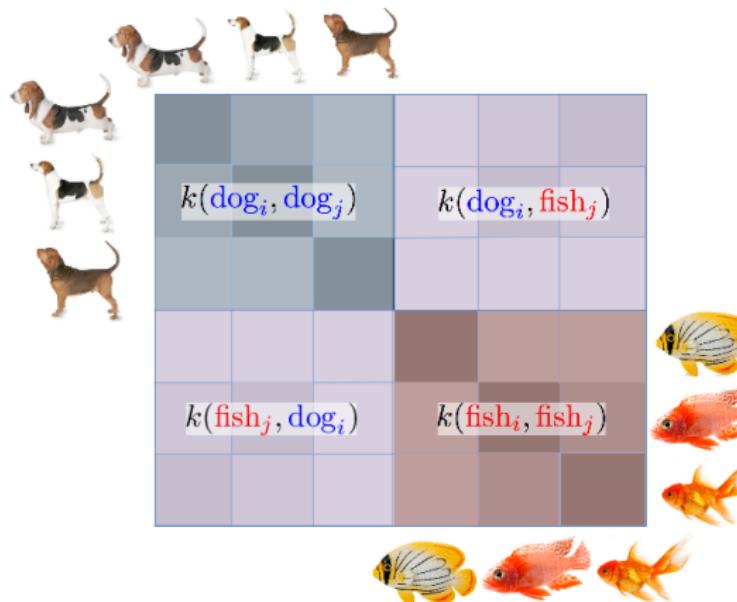


Illustration of MMD

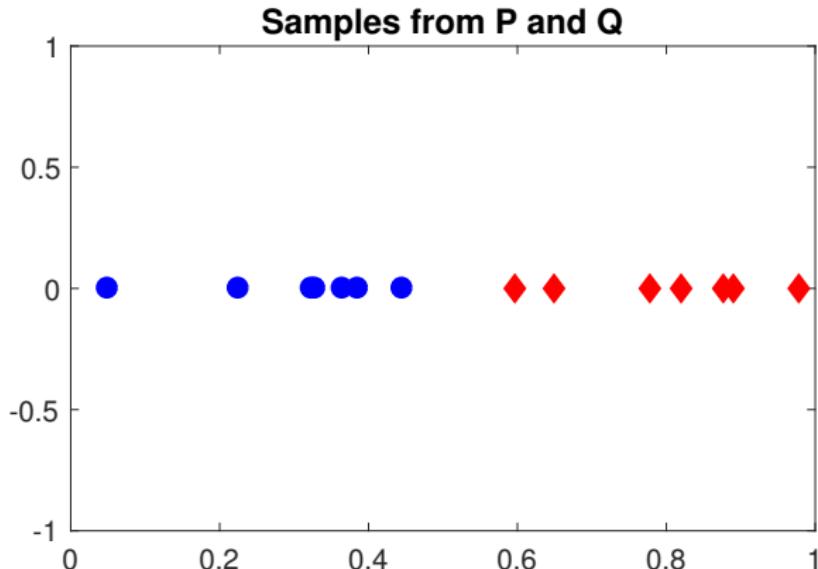
The maximum mean discrepancy:

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



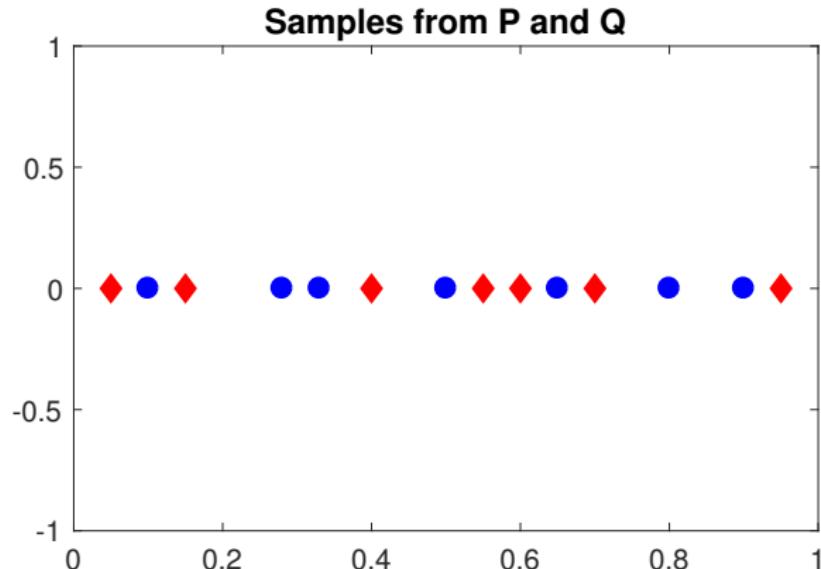
MMD as an integral probability metric

Are P and Q different?



MMD as an integral probability metric

Are P and Q different?

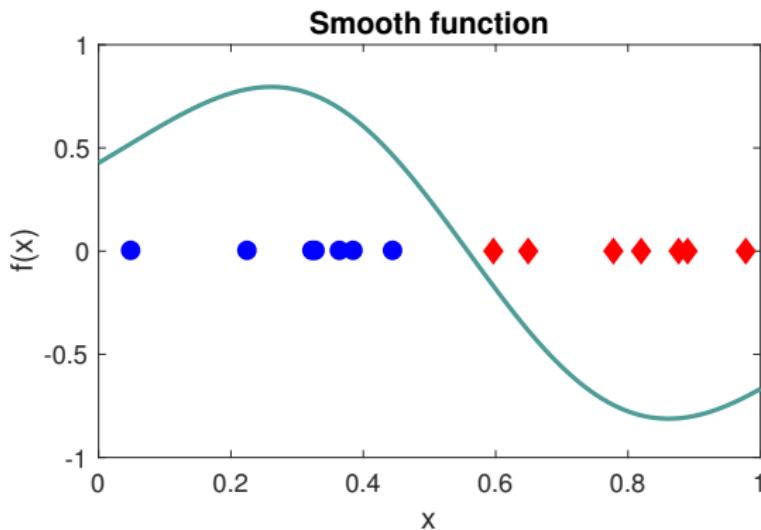


MMD as an integral probability metric

Integral probability metric:

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

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

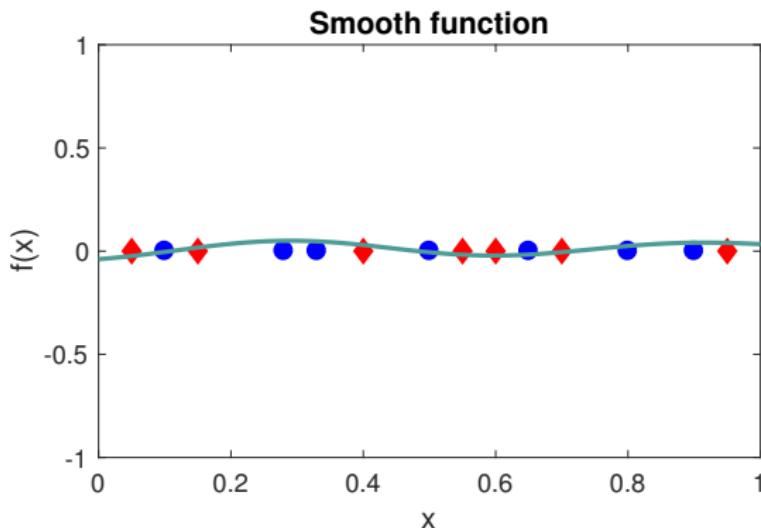


MMD as an integral probability metric

Integral probability metric:

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

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



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

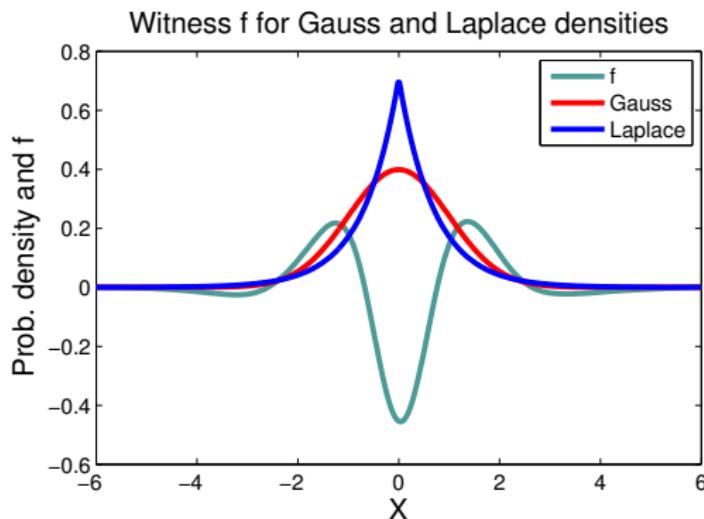
$(F = \text{unit ball in RKHS } \mathcal{F})$

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

Functions are linear combinations of features:

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

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

Expectations of functions are linear combinations
of expected features

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

(always true if kernel is bounded)

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

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

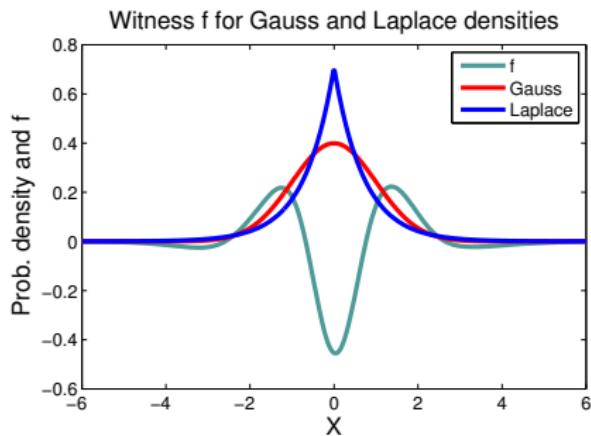
Other choices for witness function class:

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

Integral prob. metric vs feature difference

The MMD:

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



Integral prob. metric vs feature difference

The MMD:

use

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

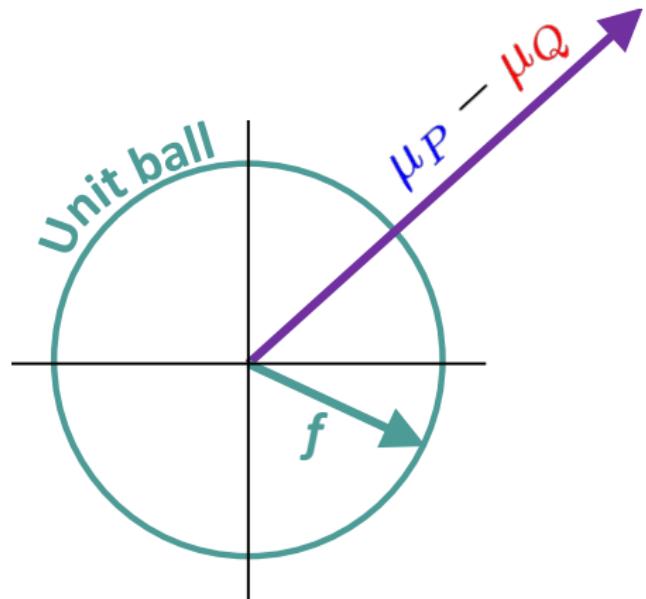
Integral prob. metric vs feature difference

The MMD:

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

$$= \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)]$$

$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



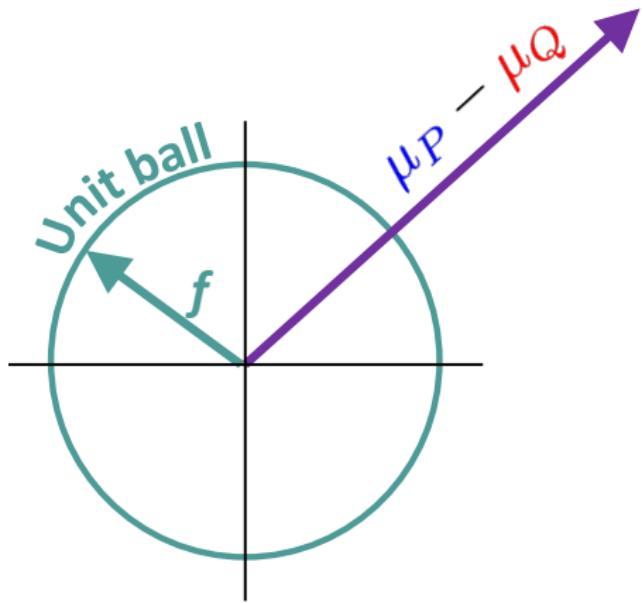
Integral prob. metric vs feature difference

The MMD:

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

$$= \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)]$$

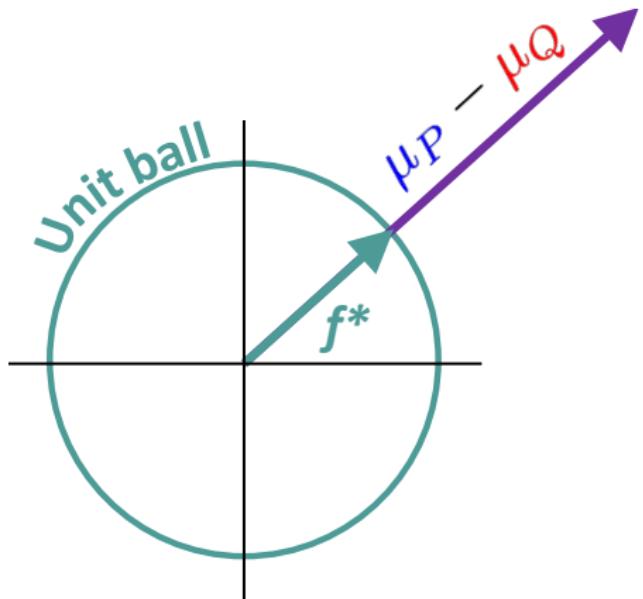
$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature difference

The MMD:

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



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

Integral prob. metric vs feature difference

The MMD:

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

Function view and feature view equivalent

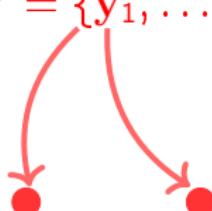
Construction of MMD witness

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

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

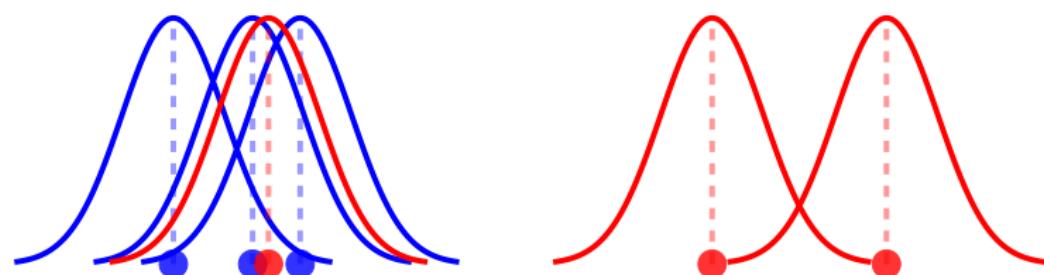


Observe $Y = \{y_1, \dots, y_n\} \sim Q$



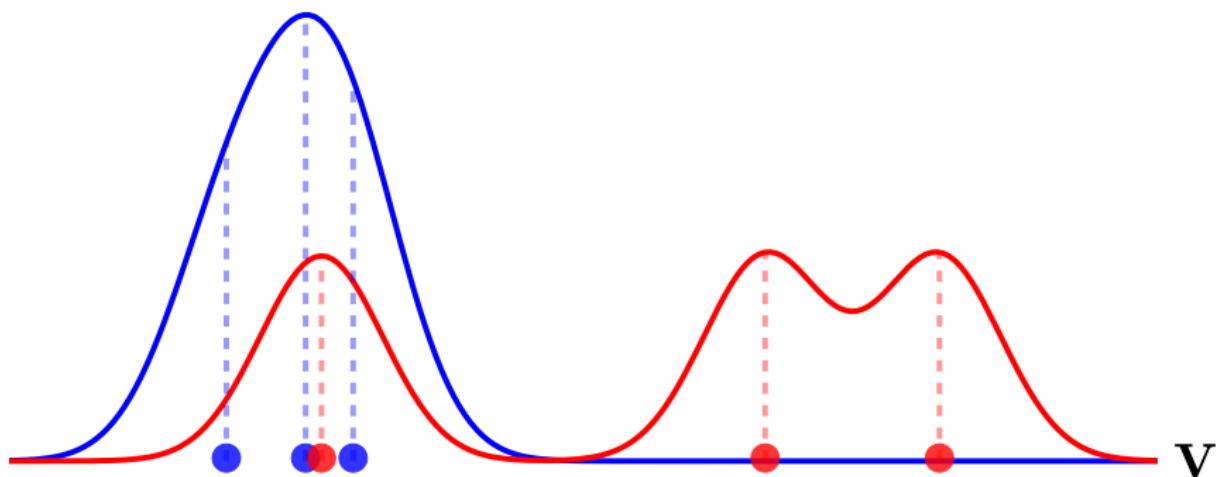
Construction of MMD witness

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



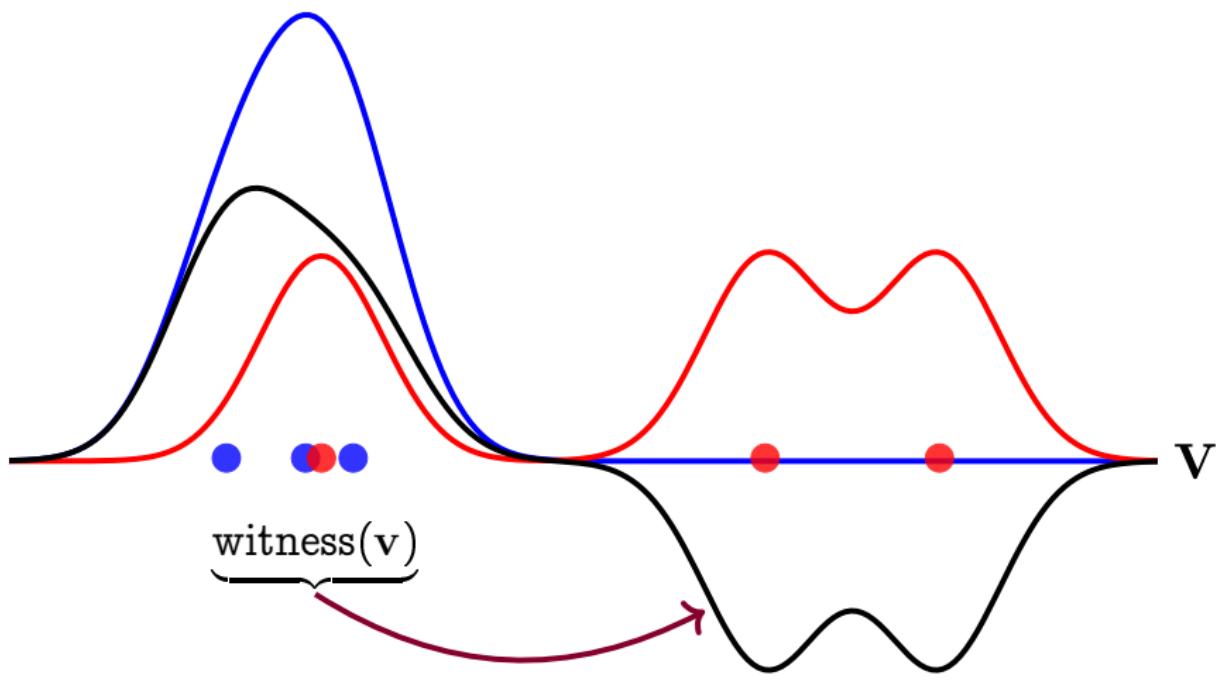
Construction of MMD witness

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



Construction of MMD witness

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



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

$$\textcolor{teal}{f}^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

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

The empirical witness function at v

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

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

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

Derivation of empirical witness function

Recall the witness function expression

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

The empirical feature mean for P

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

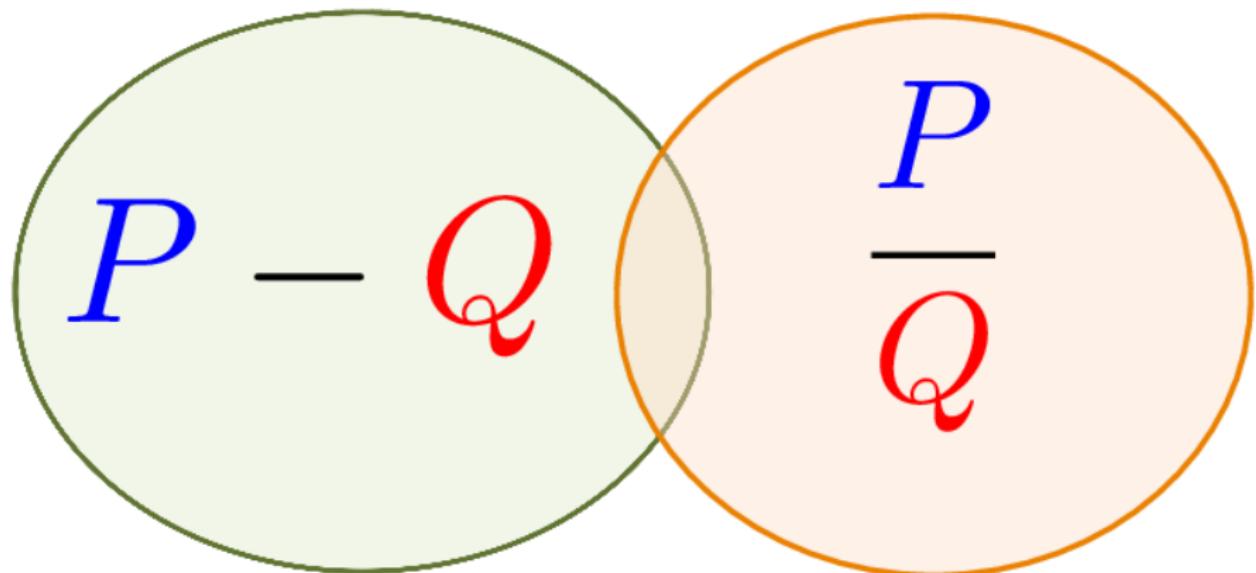
The empirical witness function at v

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

Don't need explicit feature coefficients $f^* := [f_1^* \ f_2^* \ \dots]$

Interlude: divergence measures

Divergences



Divergences

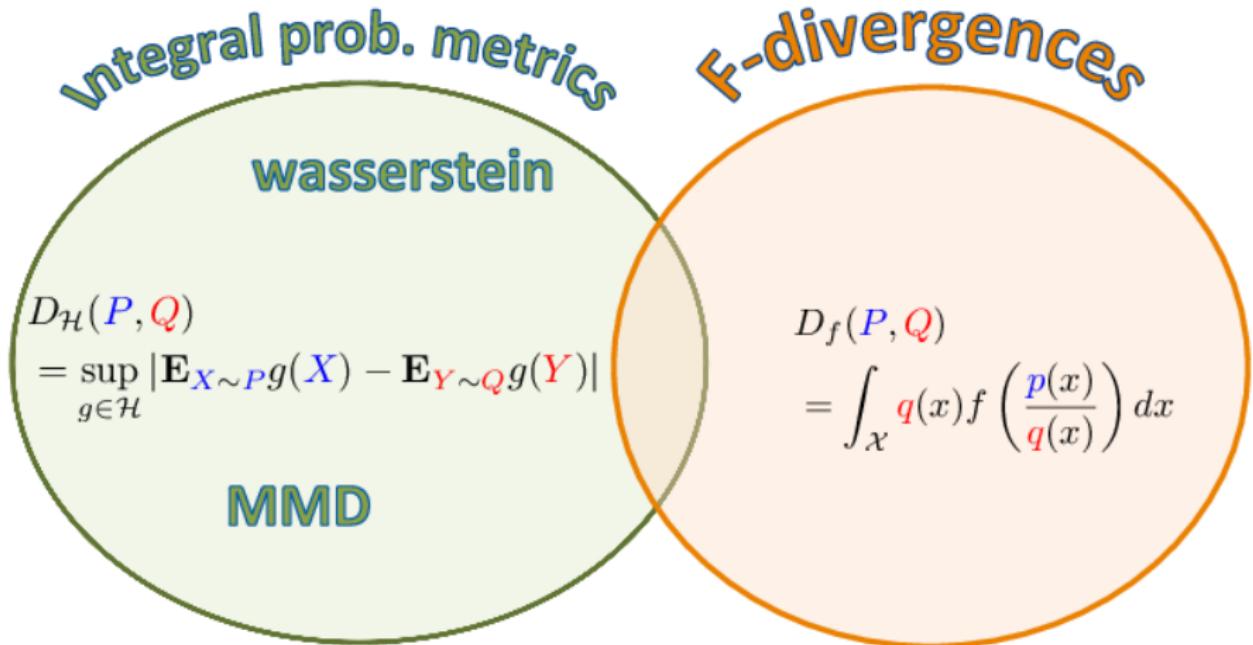
Integral prob. metrics

F-divergences

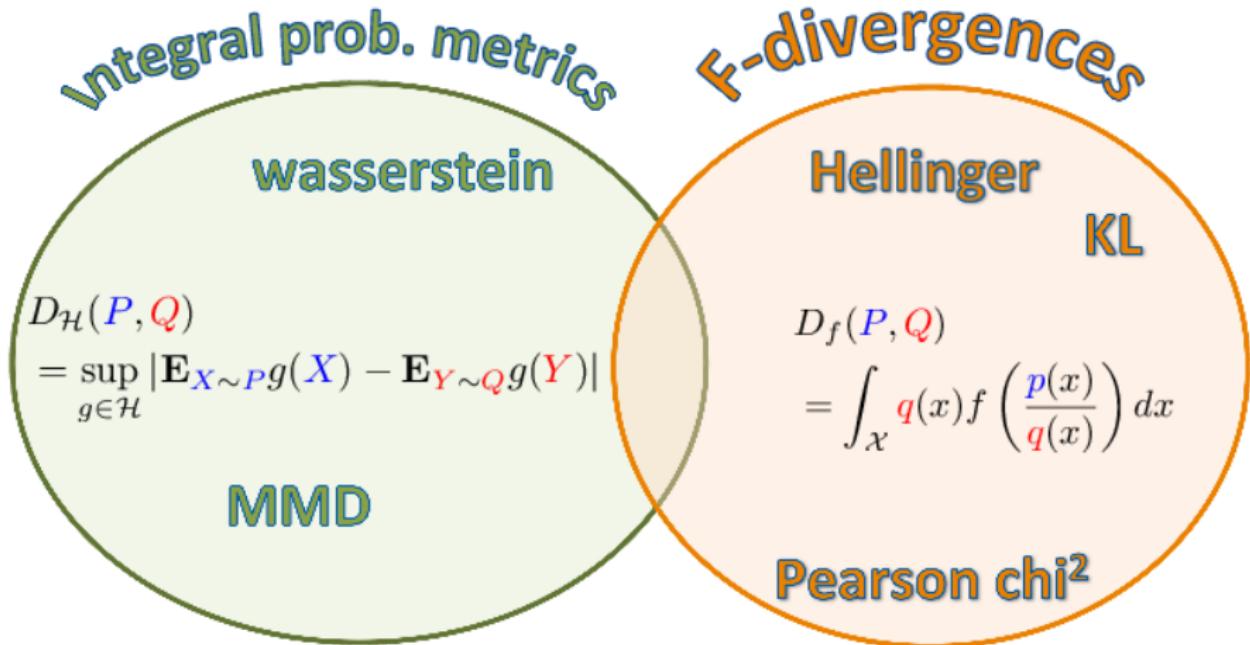
$$D_{\mathcal{H}}(\mathbf{P}, \mathbf{Q}) = \sup_{g \in \mathcal{H}} |\mathbf{E}_{X \sim \mathbf{P}} g(X) - \mathbf{E}_{Y \sim \mathbf{Q}} g(Y)|$$

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

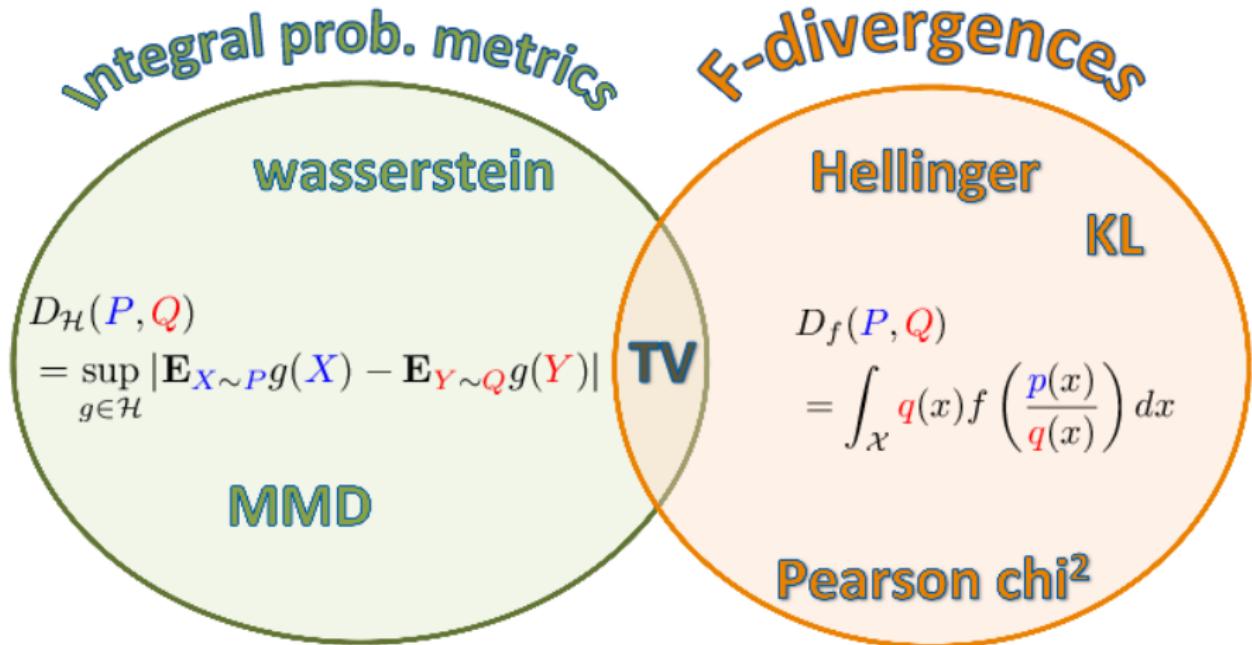
Divergences



Divergences



Divergences



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

Two-Sample Testing with MMD

A statistical test using MMD

The empirical MMD:

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

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

A statistical test using MMD

The empirical MMD:

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

Perspective from [statistical hypothesis testing](#):

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

A statistical test using MMD

The empirical MMD:

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

Perspective from [statistical hypothesis testing](#):

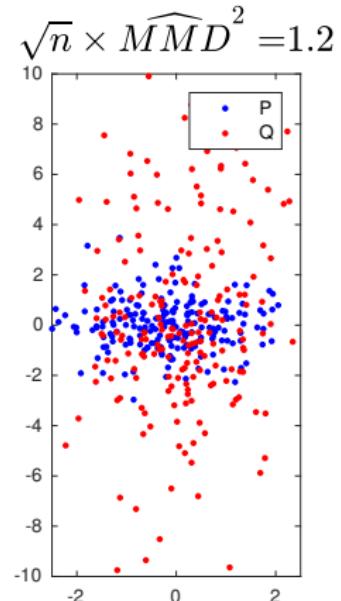
- Null hypothesis \mathcal{H}_0 when $P = Q$
 - should see \widehat{MMD}^2 “close to zero”.
- Alternative hypothesis \mathcal{H}_1 when $P \neq Q$
 - should see \widehat{MMD}^2 “far from zero”

Want [Threshold](#) c_α for \widehat{MMD}^2 to get [false positive rate](#) α

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

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

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

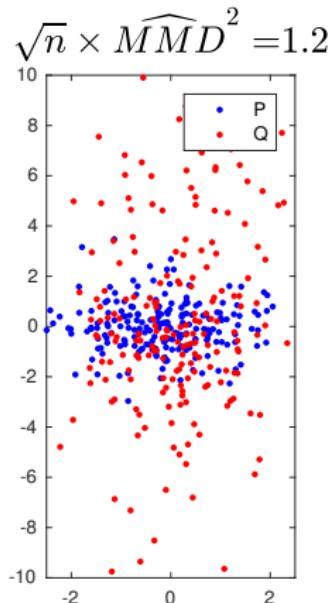
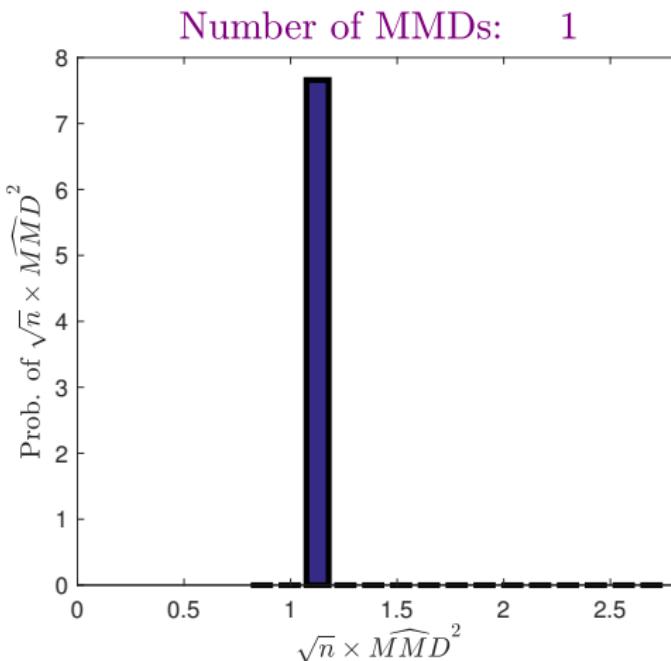


Behaviour of \widehat{MMD}^2 when $P \neq Q$

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

- Laplace with different y-variance.

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

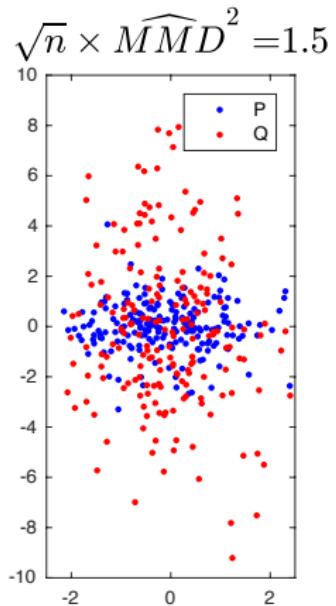
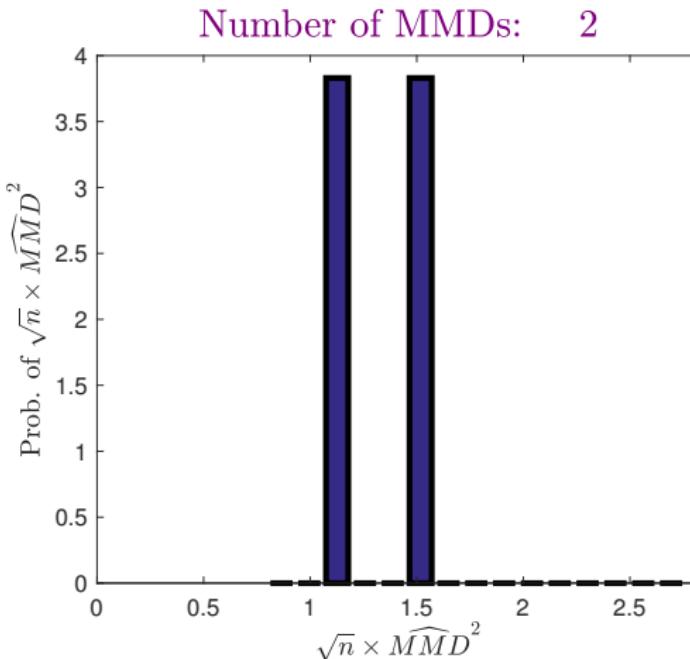


Behaviour of \widehat{MMD}^2 when $P \neq Q$

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

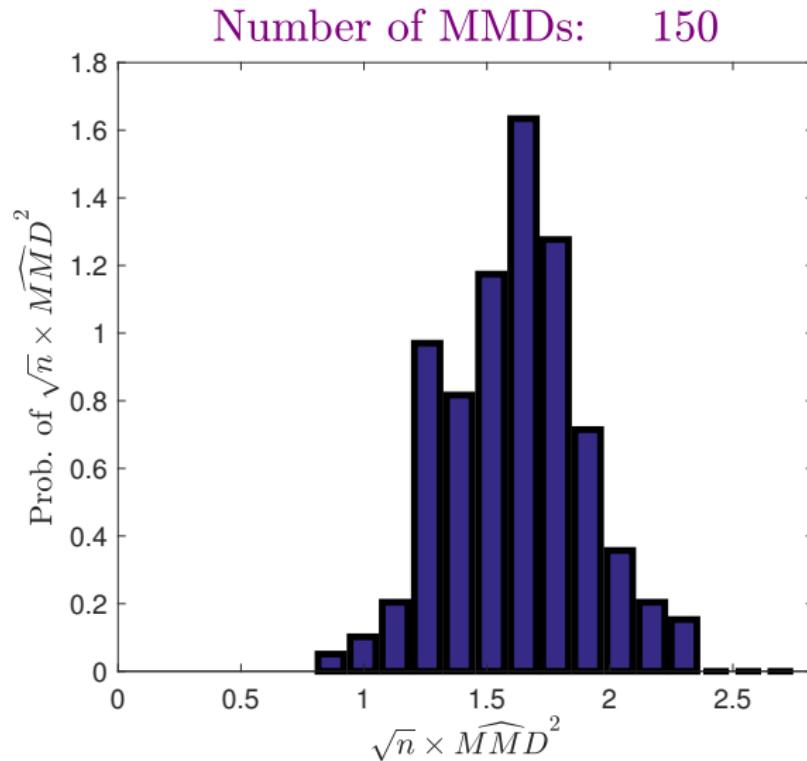
- Laplace with different y-variance.

- $\sqrt{n} \times \widehat{MMD}^2 = 1.5$



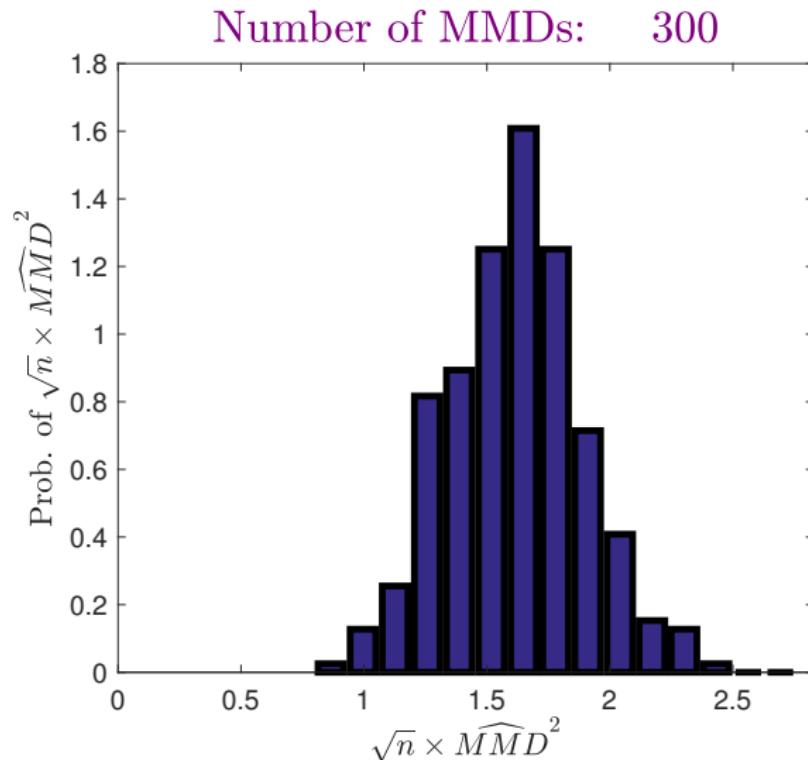
Behaviour of \widehat{MMD}^2 when $P \neq Q$

Repeat this 150 times ...



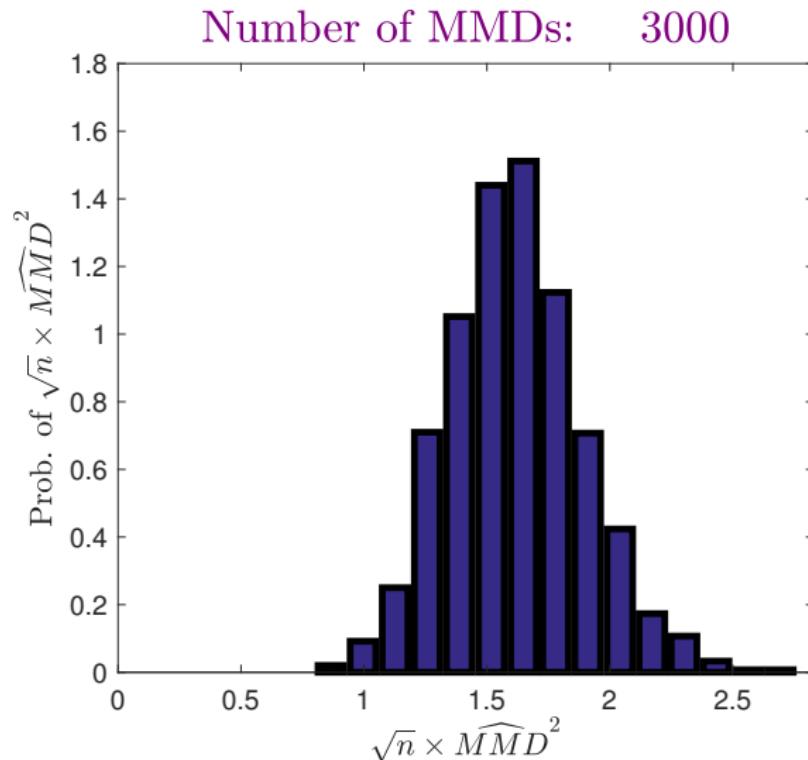
Behaviour of \widehat{MMD}^2 when $P \neq Q$

Repeat this 300 times ...



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

Repeat this 3000 times . . .

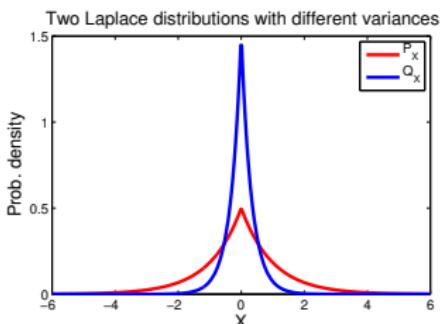
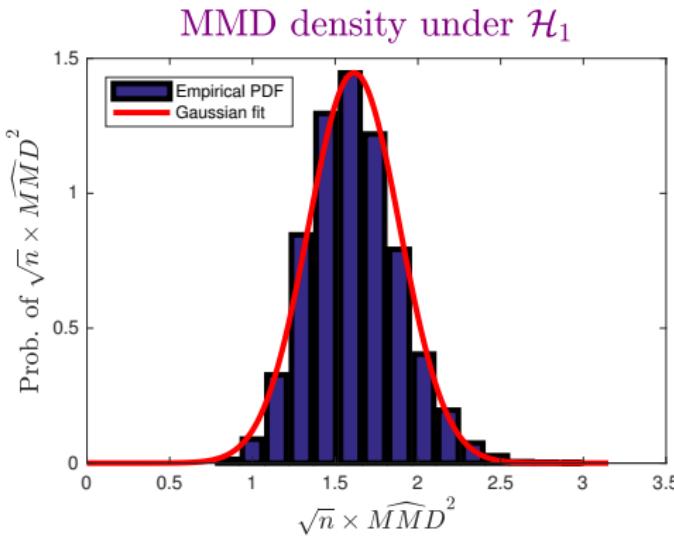


Asymptotics of \widehat{MMD}^2 when $P \neq Q$

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

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

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

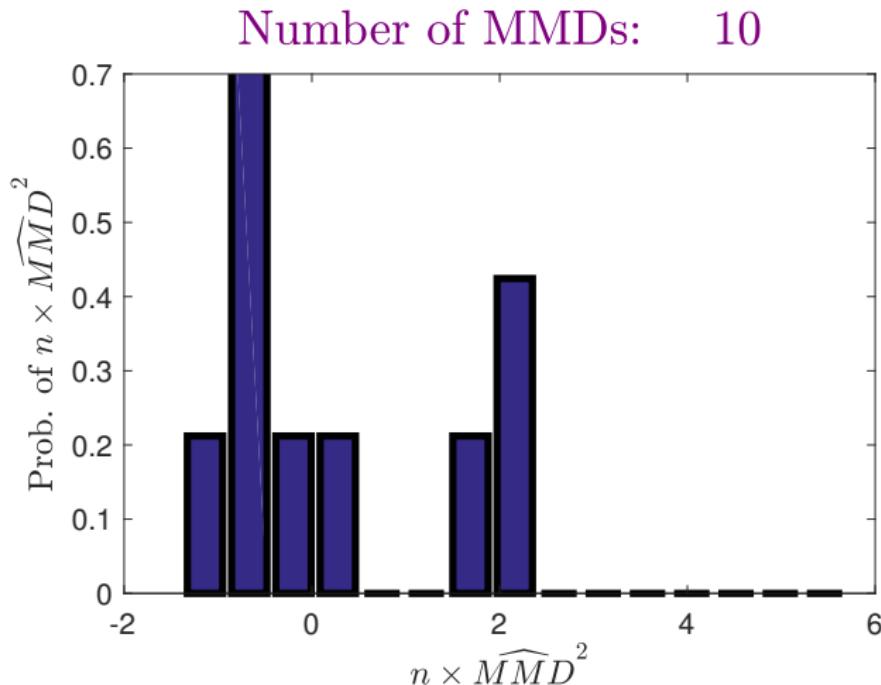


Behaviour of \widehat{MMD}^2 when $P = Q$

What happens when P and Q are the same?

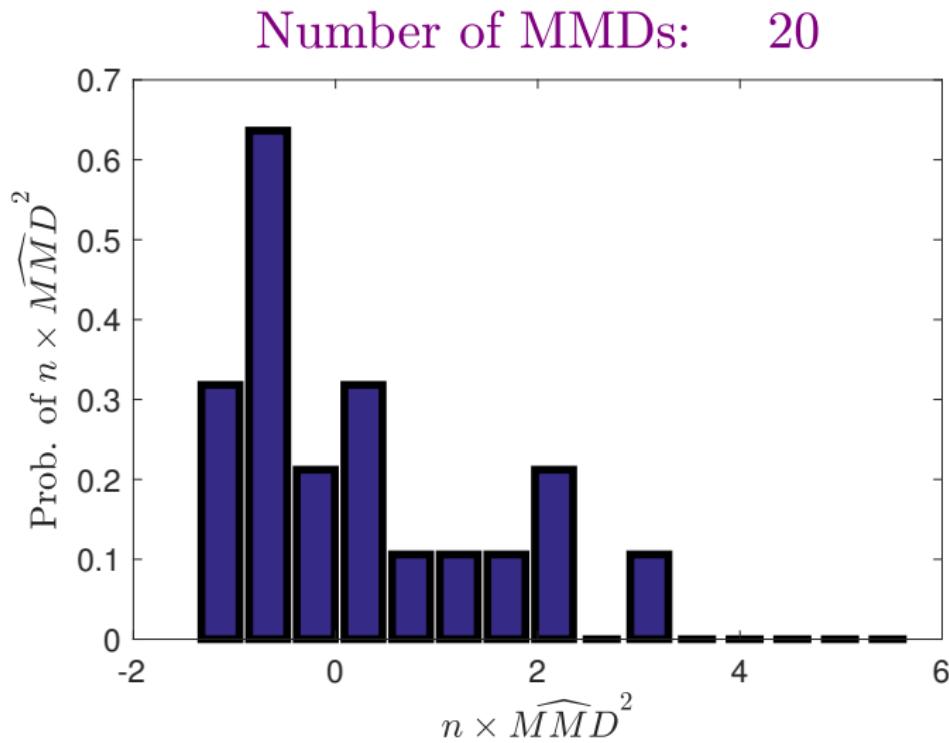
Behaviour of \widehat{MMD}^2 when $P = Q$

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



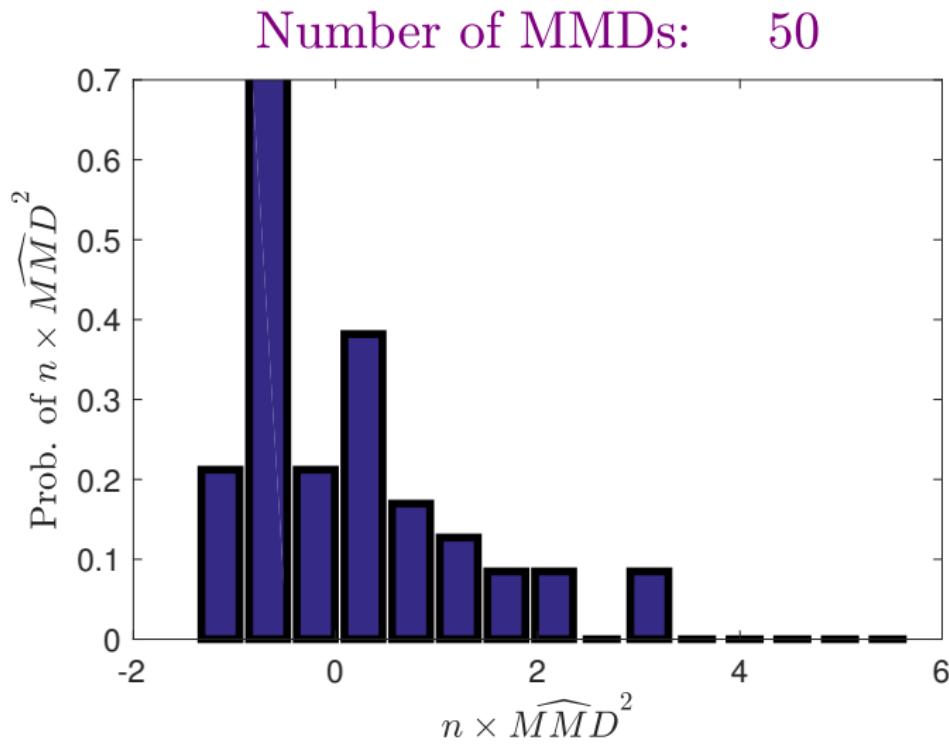
Behaviour of \widehat{MMD}^2 when $P = Q$

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



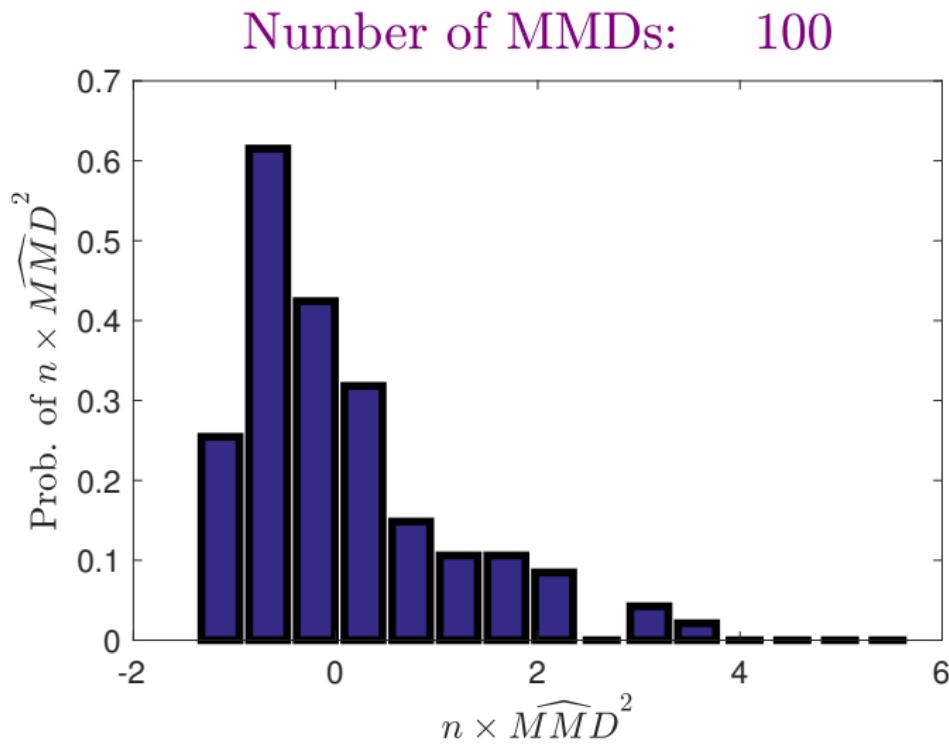
Behaviour of \widehat{MMD}^2 when $P = Q$

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



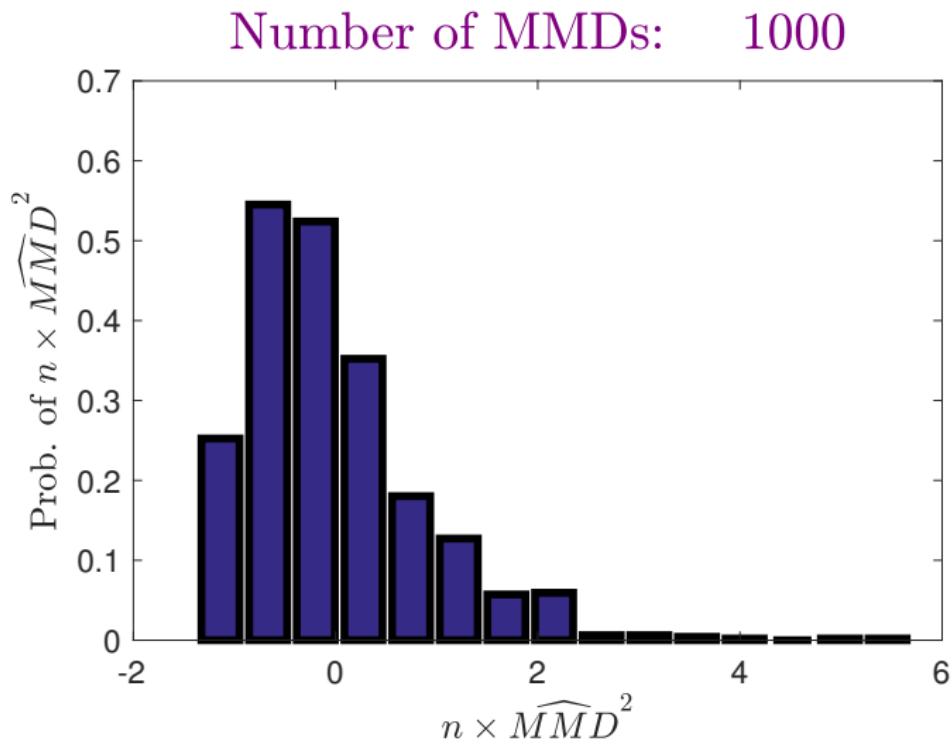
Behaviour of \widehat{MMD}^2 when $P = Q$

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



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

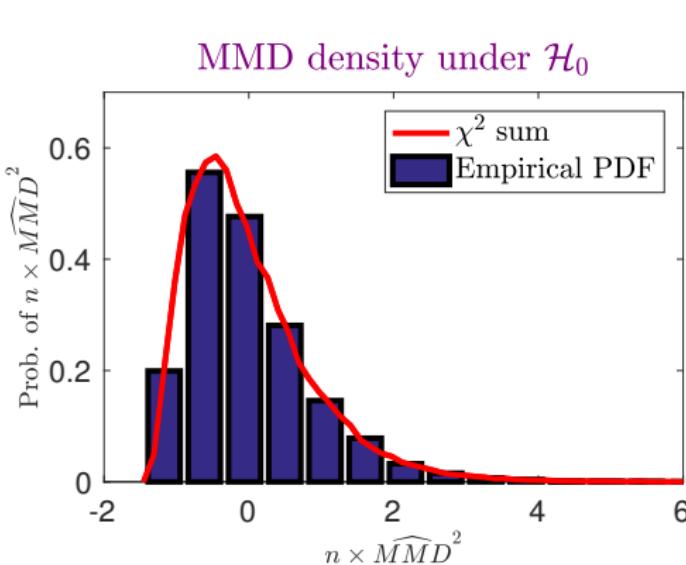
- Case of $P = Q = \mathcal{N}(0, 1)$



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

Where $P = Q$, statistic has asymptotic distribution

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



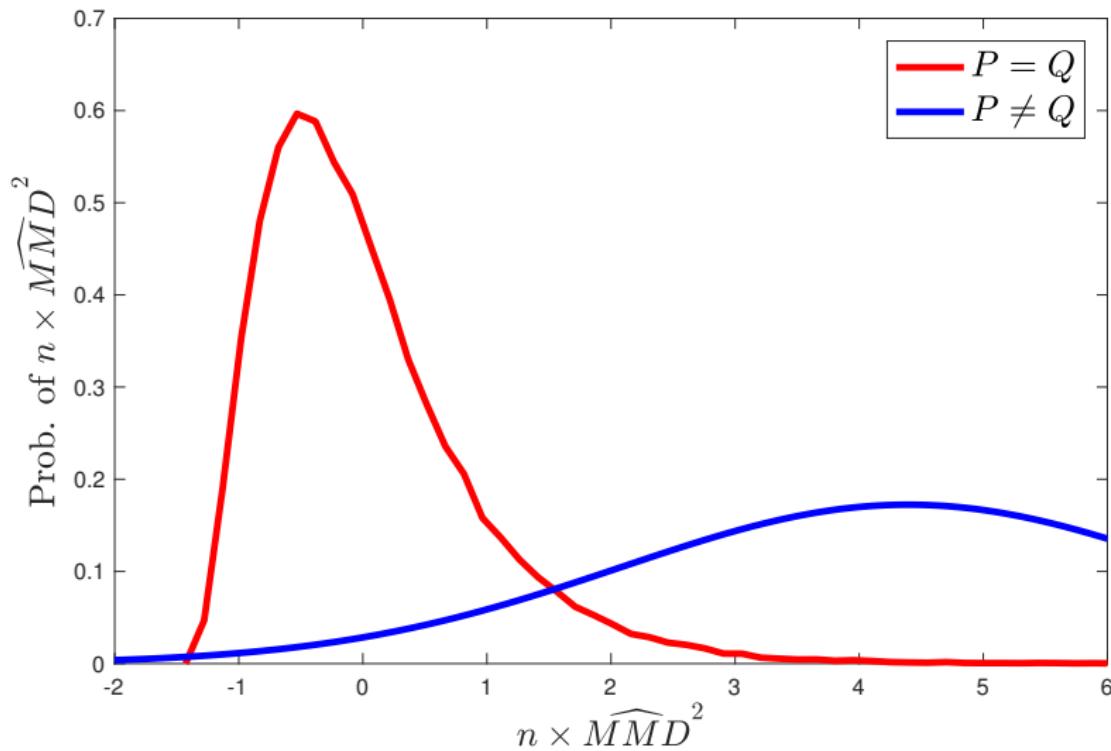
where

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

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

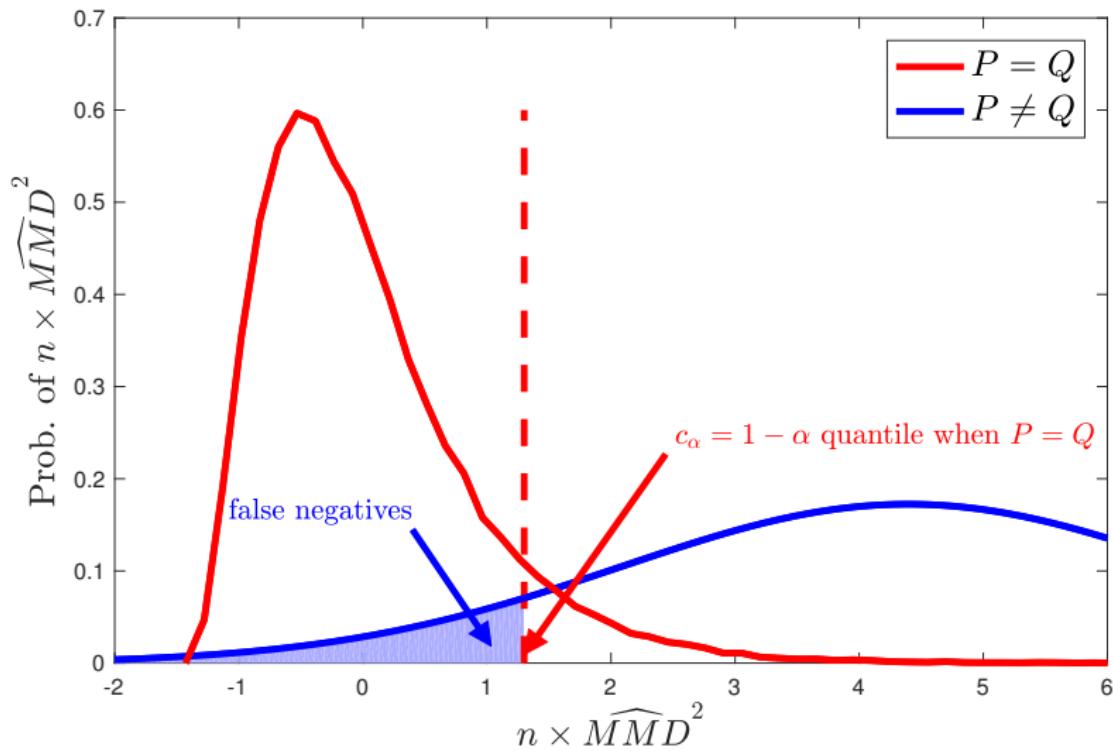
A statistical test

A summary of the asymptotics:



A statistical test

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



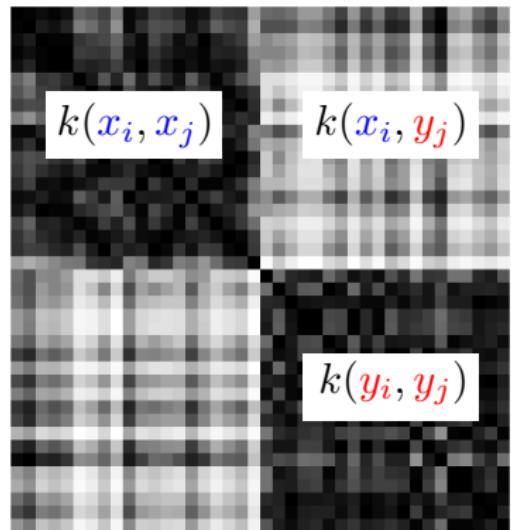
How do we get test threshold c_α ?

Original empirical MMD for dogs and fish:

$$X = \begin{bmatrix} \text{Basset Hound} & \text{Beagle} & \text{Basset Hound} & \dots \end{bmatrix}$$

$$Y = \begin{bmatrix} \text{Butterfly Fish} & \text{Coral Fish} & \text{Goldfish} & \dots \end{bmatrix}$$

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



How do we get test threshold c_α ?

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

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

$$\tilde{Y} = \begin{bmatrix} \text{dog emoji} & \text{fish emoji} & \text{dog emoji} & \dots \end{bmatrix}$$

How do we get test threshold c_α ?

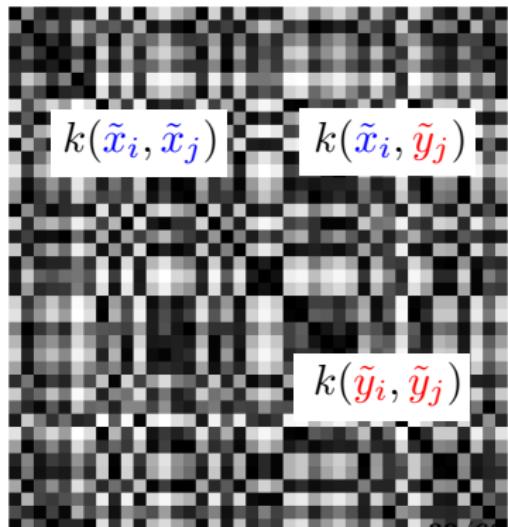
Permuted dog and fish samples (**merdogs**):

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

$$\tilde{Y} = [\text{dog emoji} \quad \text{fish emoji} \quad \text{dog emoji} \quad \dots]$$

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

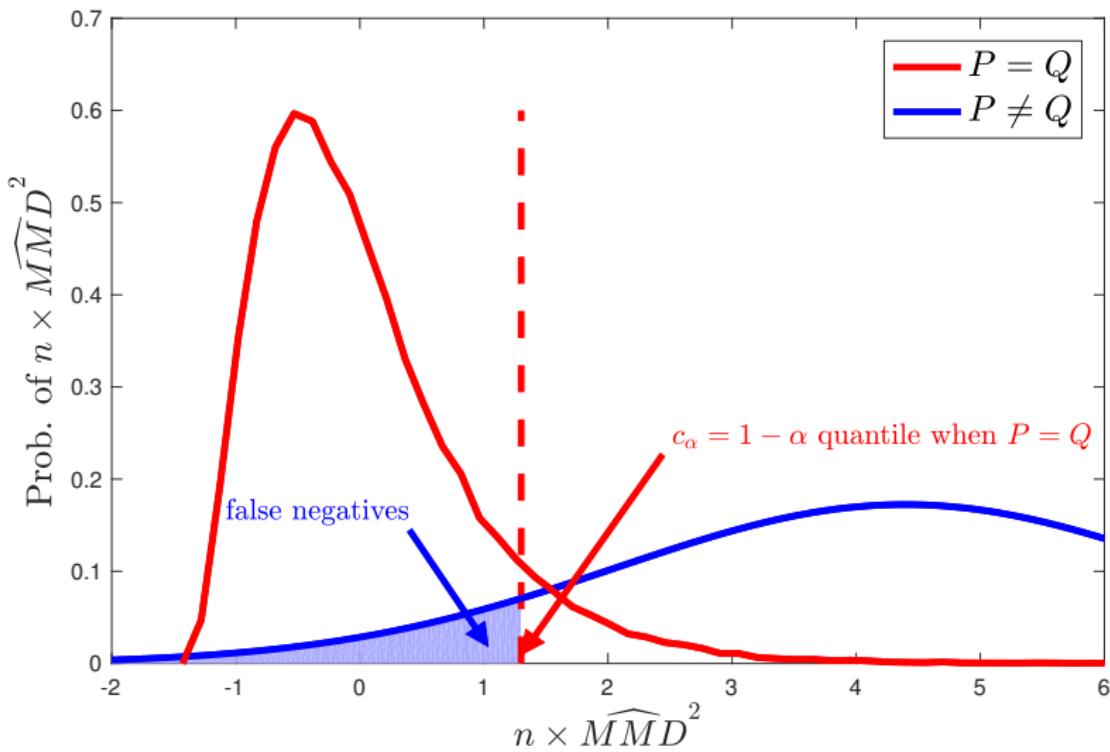
Permutation simulates
 $P = Q$



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

Graphical illustration

- Maximising test power same as minimizing false negatives



Optimizing kernel for test power

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

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

Optimizing kernel for test power

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

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

where

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

Optimizing kernel for test power

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

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

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

For large n , second term negligible!

Optimizing kernel for test power

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

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

To maximize test power, maximize

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

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

Code: github.com/dougal-sutherland/opt-mmd

Troubleshooting for generative adversarial networks



MNIST samples



Samples from a GAN

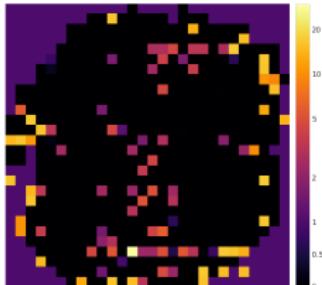
Troubleshooting for generative adversarial networks



MNIST samples



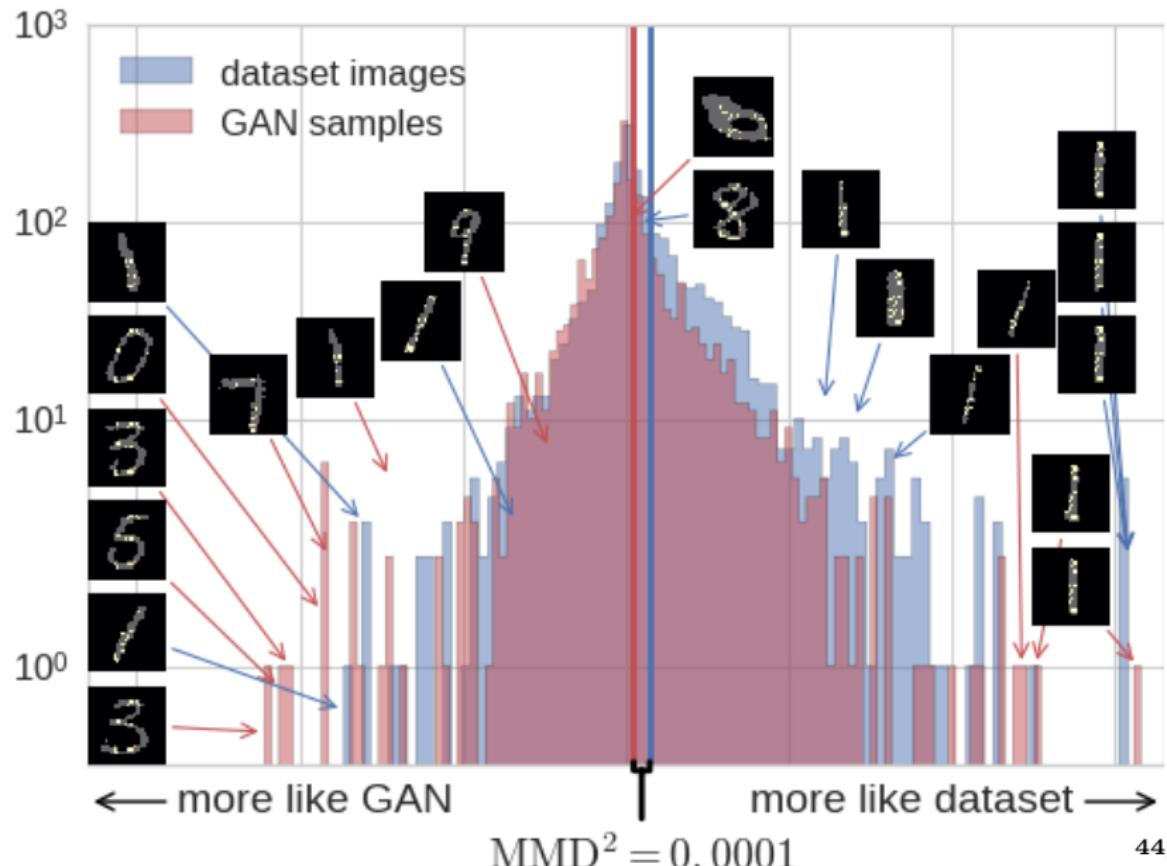
Samples from a GAN



ARD map

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

Troubleshooting generative adversarial networks



How to choose the best kernel (2) characteristic kernels

Characteristic kernels

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

[NIPS07b, JMLR10]

In the next slides:

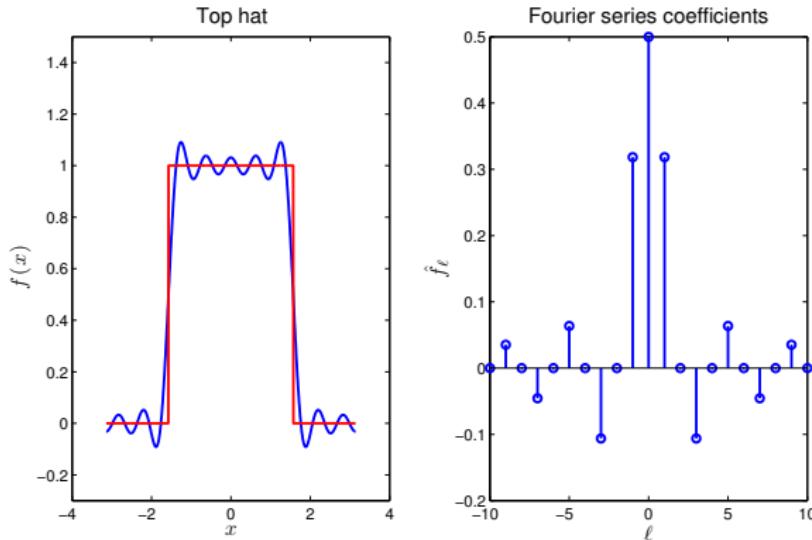
- Characteristic property on $[-\pi, \pi]$ with periodic boundary
- Characteristic property on \mathbb{R}^d

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

Reminder: **Fourier series**

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

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

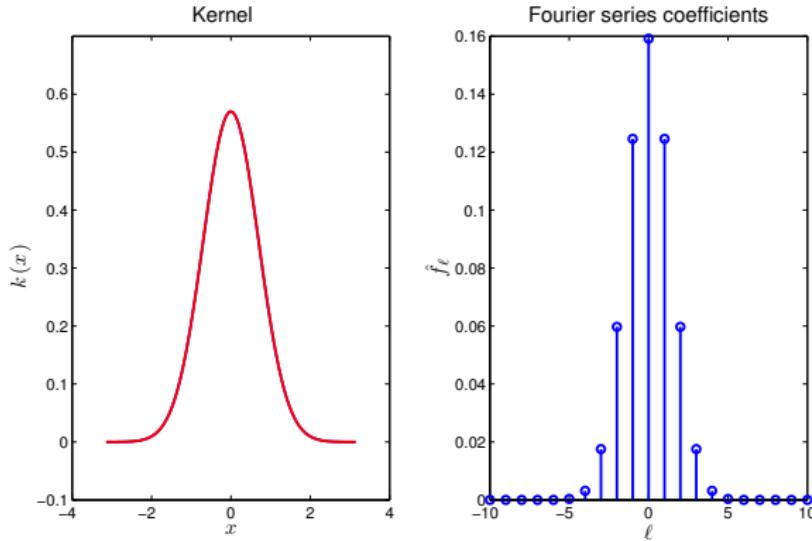


Characteristic kernels on $[-\pi, \pi]$

Jacobi theta kernel (close to exponentiated quadratic):

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

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



The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

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

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

The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

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

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

The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

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

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

The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

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

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

The MMD in a Fourier representation

Maximum mean embedding via Fourier series:

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

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

MMD can be written in terms of Fourier series:

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

A simpler Fourier representation for MMD

From previous slide,

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

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:

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

A simpler Fourier representation for MMD

From previous slide,

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

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

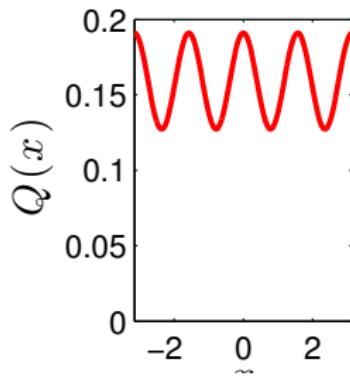
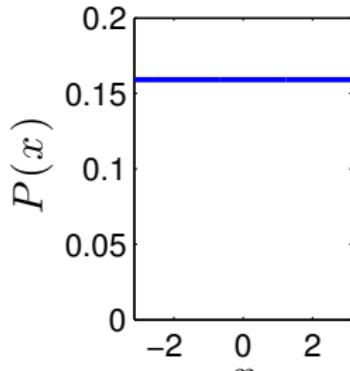
$$\|f\|_{\mathcal{F}}^2 = \sum_{\ell=-\infty}^{\infty} \frac{|\hat{f}_\ell|^2}{\hat{k}_\ell}.$$

Simple, interpretable expression for squared MMD:

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

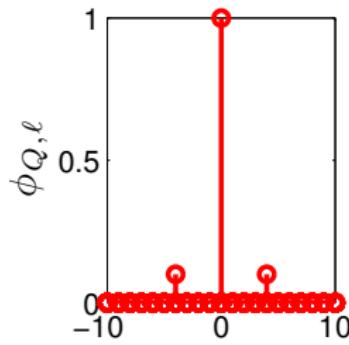
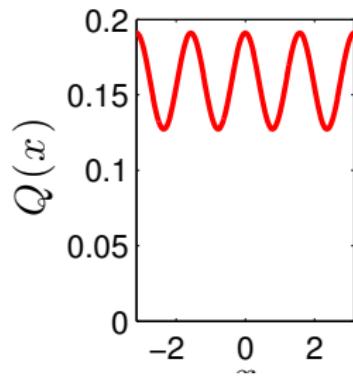
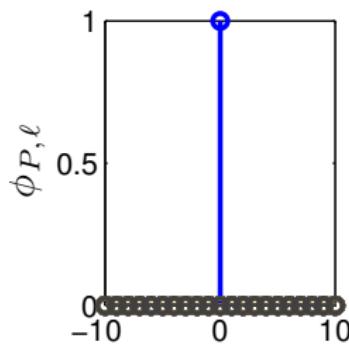
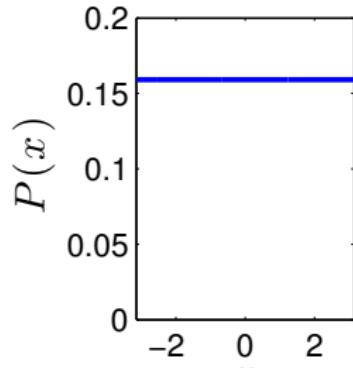
Characteristic kernels on $[-\pi, \pi]$

Example: P differs from Q at one frequency:



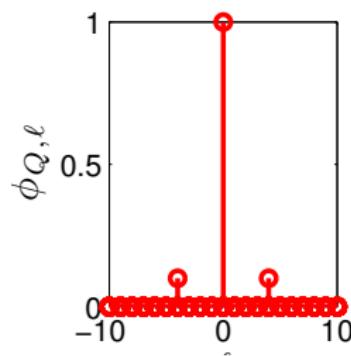
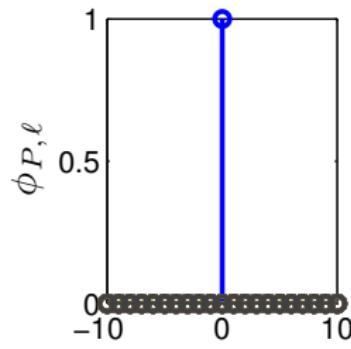
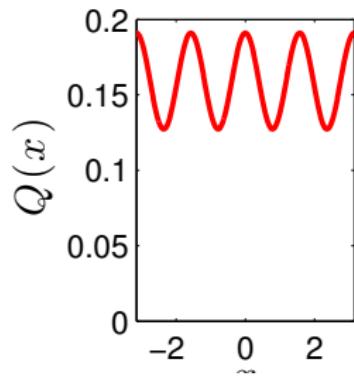
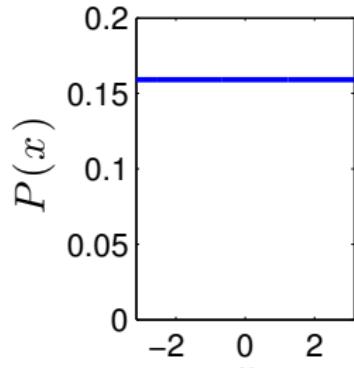
Characteristic kernels on $[-\pi, \pi]$

Example: P differs from Q at one frequency:

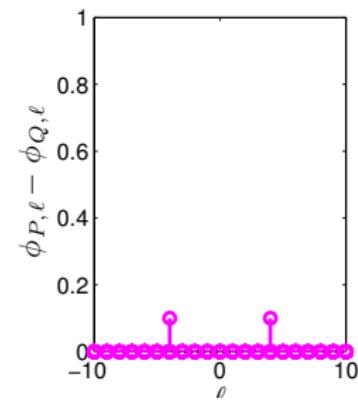


Characteristic kernels on $[-\pi, \pi]$

Example: P differs from Q at one frequency:

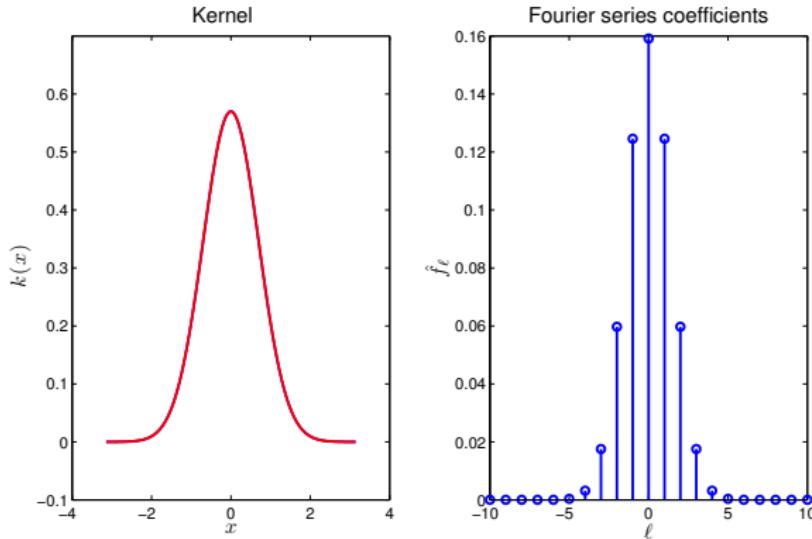


Characteristic function difference



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

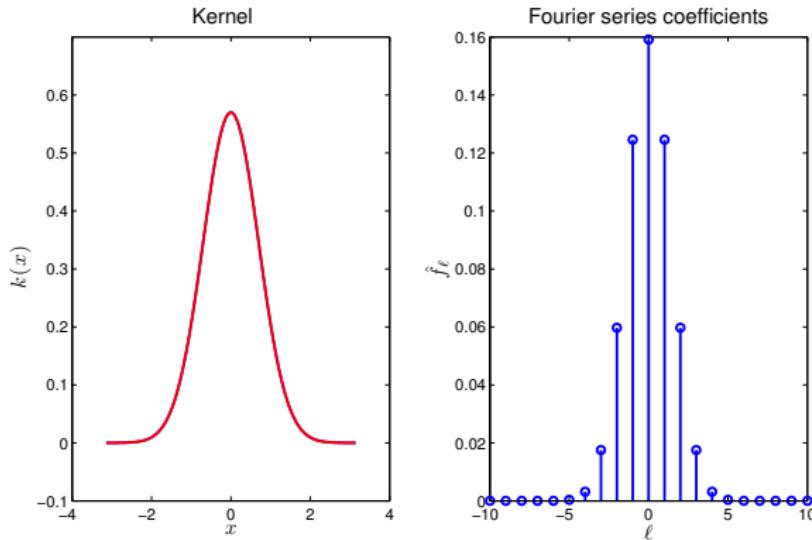
Is the Gaussian spectrum kernel characteristic?



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

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

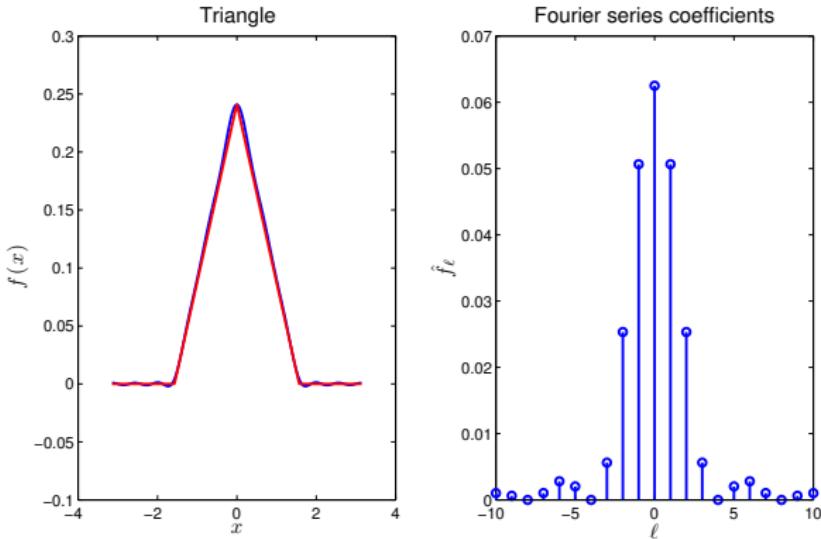
Is the Gaussian spectrum kernel characteristic? YES



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

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

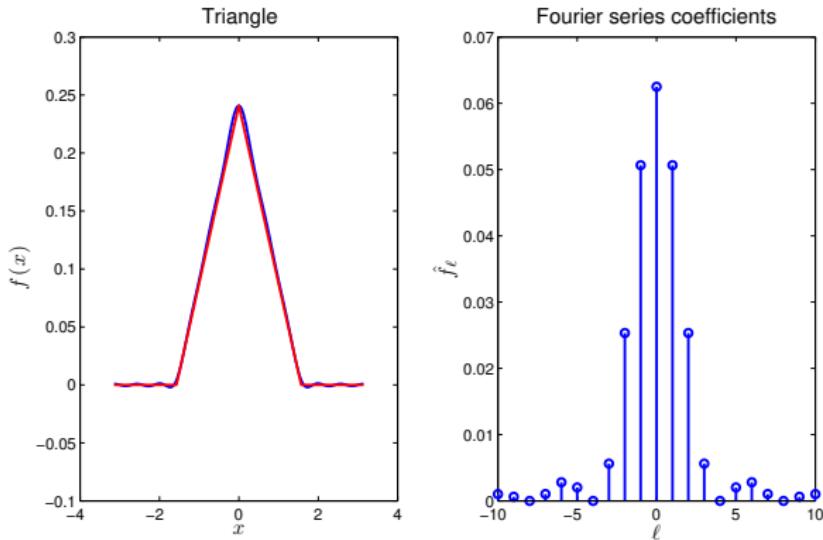
Is the triangle kernel characteristic?



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

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

Is the triangle kernel characteristic? NO



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

Characteristic kernels on \mathbb{R}^d

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

Characteristic function of P via Fourier transform

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

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

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

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

Characteristic kernels on \mathbb{R}^d

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

Characteristic function of P via Fourier transform

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

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

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

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

Characteristic kernels on \mathbb{R}^d

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

$$MMD^2(\textcolor{blue}{P}, \textcolor{red}{Q}; F) = \int |\varphi_{\textcolor{blue}{P}}(\omega) - \varphi_{\textcolor{red}{Q}}(\omega)|^2 d\Lambda(\omega)$$

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

$$MMD^2(\textcolor{blue}{P}, \textcolor{red}{Q}; F) = \sum_{l=-\infty}^{\infty} |\varphi_{\textcolor{blue}{P},l} - \varphi_{\textcolor{red}{Q},l}|^2 \hat{k}_l$$

Characteristic kernels on \mathbb{R}^d

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

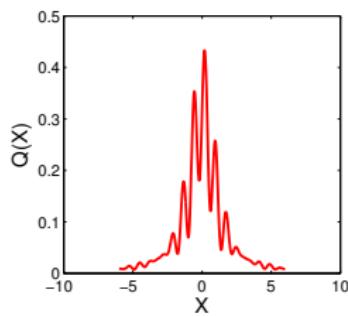
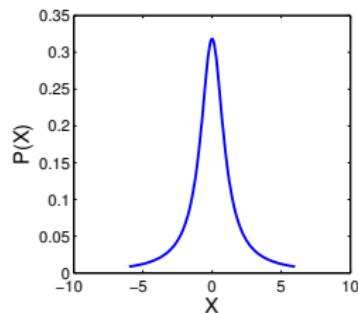
$$MMD^2(\mathcal{P}, \mathcal{Q}; F) = \int |\varphi_{\mathcal{P}}(\omega) - \varphi_{\mathcal{Q}}(\omega)|^2 d\Lambda(\omega)$$

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

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

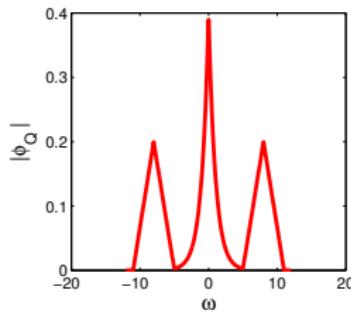
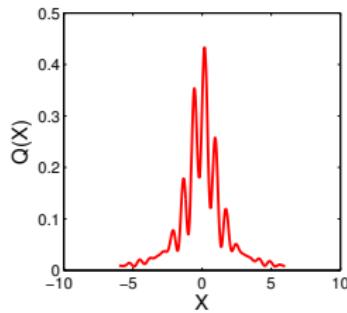
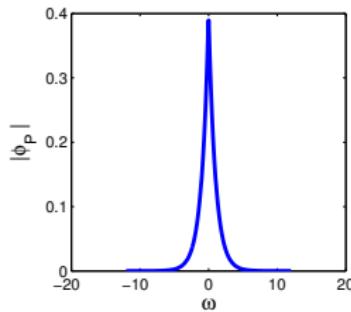
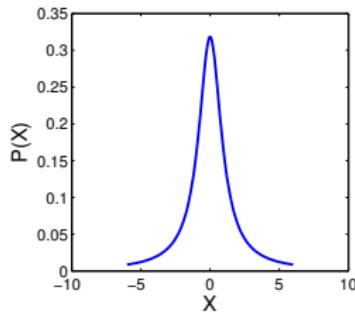
Characteristic kernels on \mathbb{R}^d

Example: P differs from Q at **roughly** one frequency:



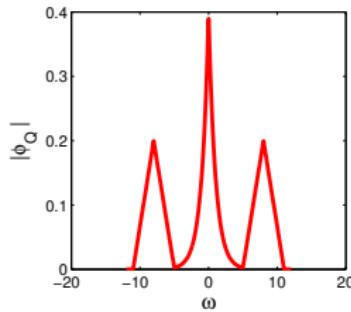
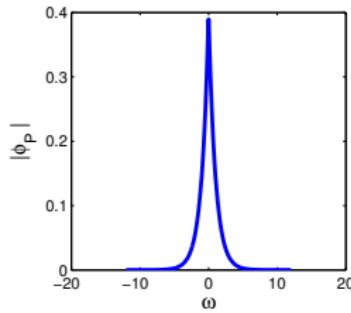
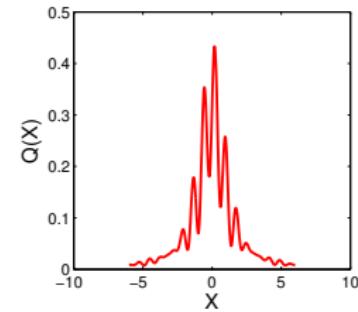
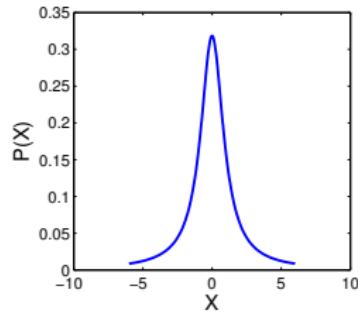
Characteristic kernels on \mathbb{R}^d

Example: P differs from Q at roughly one frequency:

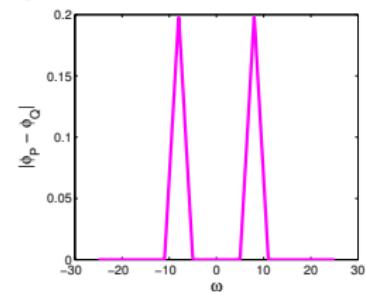


Characteristic kernels on \mathbb{R}^d

Example: P differs from Q at roughly one frequency:



Characteristic function difference

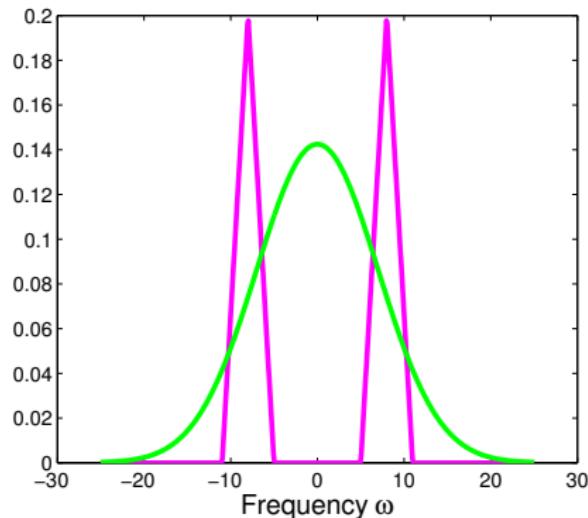


Characteristic kernels on \mathbb{R}^d

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

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

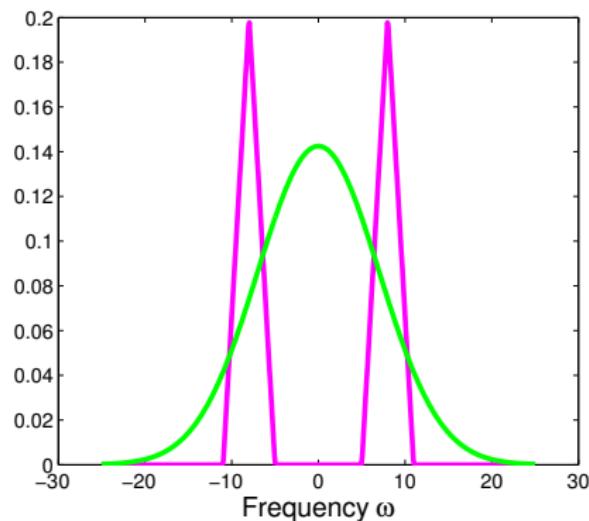
Difference $|\varphi_P - \varphi_Q|$



Characteristic kernels on \mathbb{R}^d

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

Characteristic

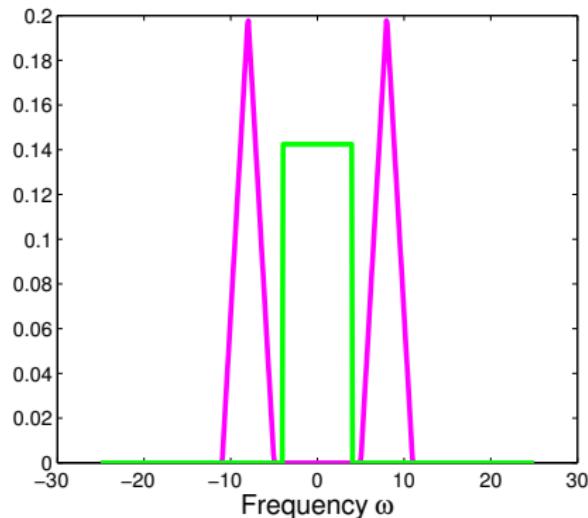


Characteristic kernels on \mathbb{R}^d

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

Sinc kernel spectrum $\Lambda(\omega)$

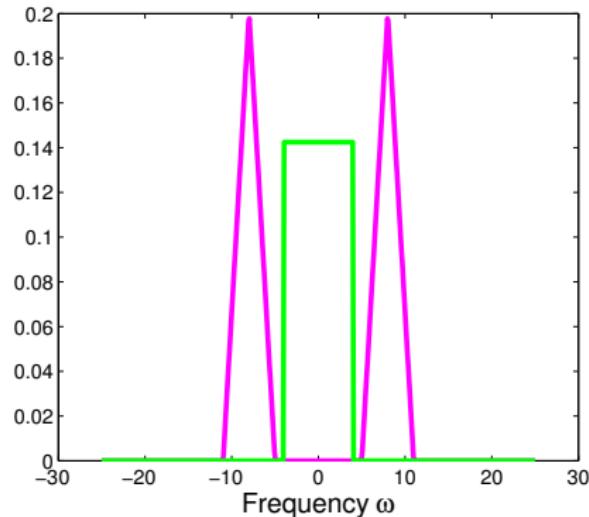
Difference $|\varphi_P - \varphi_Q|$



Characteristic kernels on \mathbb{R}^d

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

Not characteristic

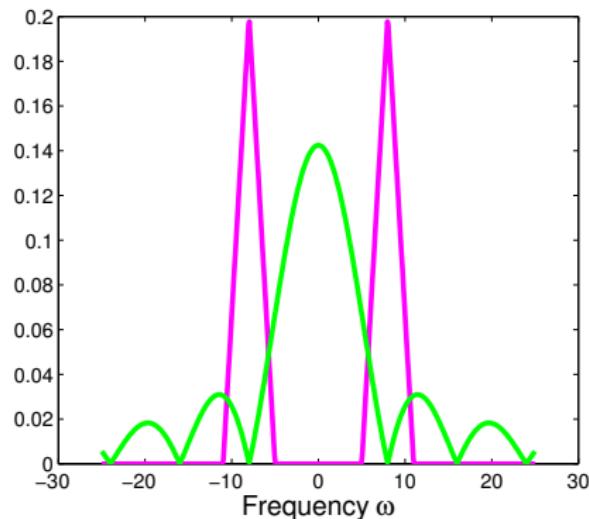


Characteristic kernels on \mathbb{R}^d

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

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

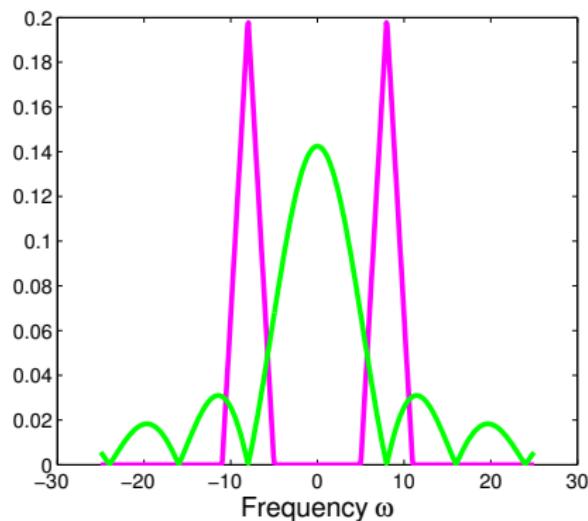
Difference $|\phi_P - \phi_Q|$



Characteristic kernels on \mathbb{R}^d

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

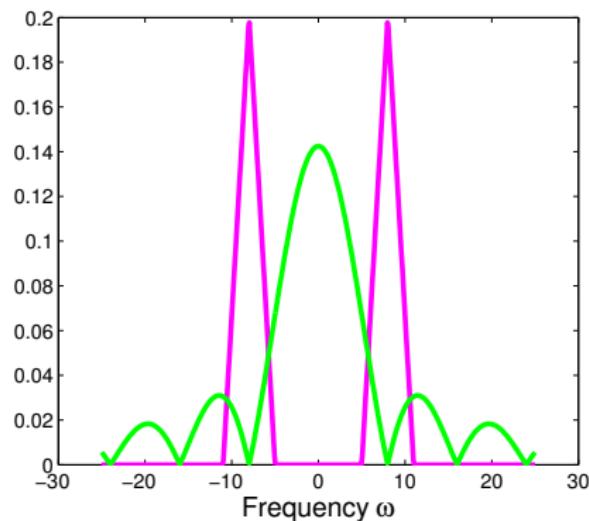
???



Characteristic kernels on \mathbb{R}^d

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

Characteristic



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

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

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

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

(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]

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]

Representing and comparing probabilities with kernels: Part 3

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

MLSS Madrid, 2018

Training GANs with MMD

What is a Generative Adversarial Network (GAN)?

- **Generator** (student)



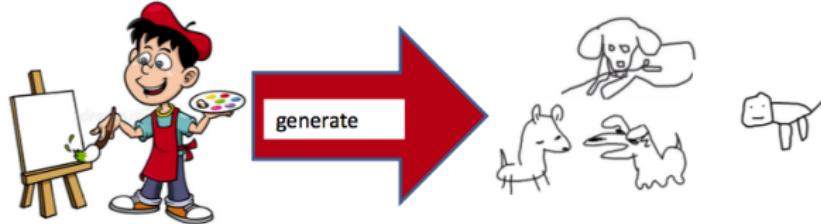
- Task: **critic** must teach **generator** to draw images (here dogs)



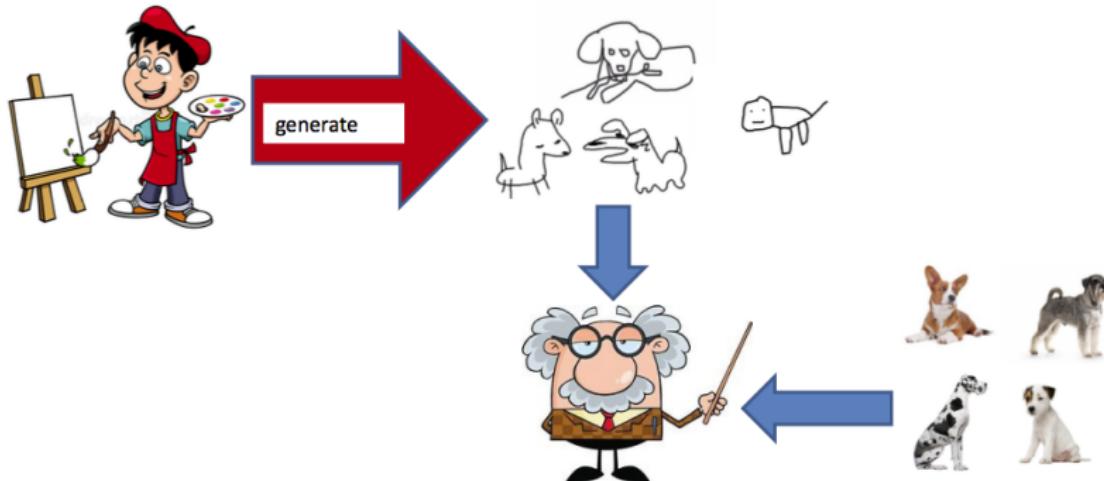
- **Critic** (teacher)



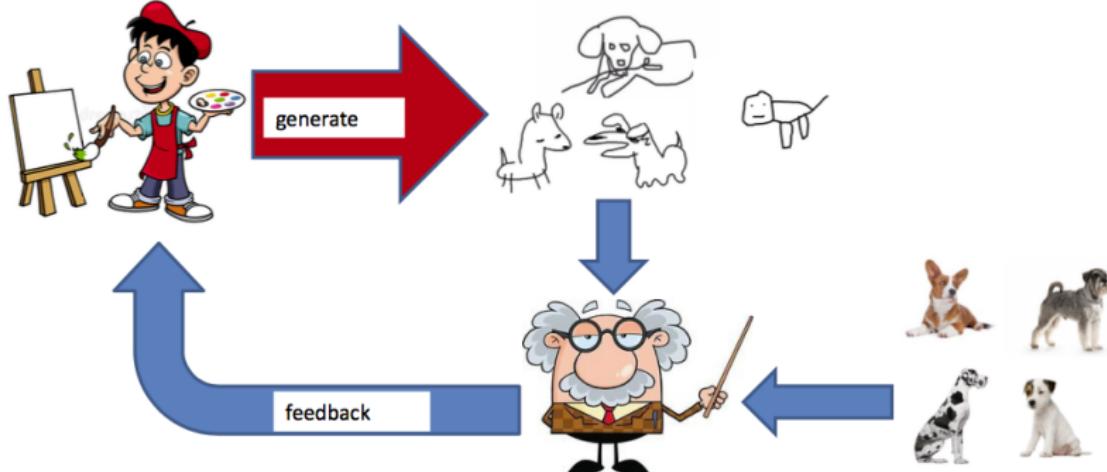
What is a Generative Adversarial Network (GAN)?



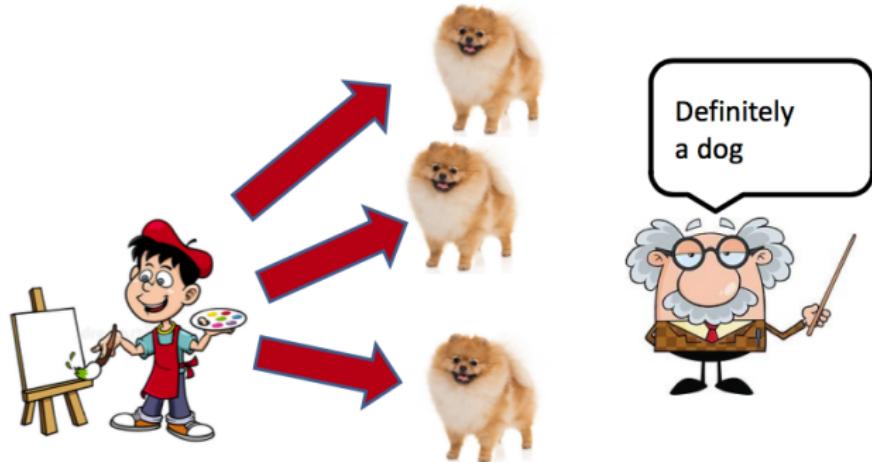
What is a Generative Adversarial Network (GAN)?



What is a Generative Adversarial Network (GAN)?



Why is classification not enough?



Classification **not** enough!
Need to compare **sets**

(otherwise student can just produce the **same** dog over and over)

MMD for GAN critic

Can you use MMD as a critic to train GANs?

From ICML 2015:

Generative Moment Matching Networks

Yujia Li¹

Kevin Swersky¹

Richard Zemel^{1,2}

YUJIALI@CS.TORONTO.EDU

KSWERSKY@CS.TORONTO.EDU

ZEMEL@CS.TORONTO.EDU

¹Department of Computer Science, University of Toronto, Toronto, ON, CANADA

²Canadian Institute for Advanced Research, Toronto, ON, CANADA

From UAI 2015:

Training generative neural networks via Maximum Mean Discrepancy optimization

Gintare Karolina Dziugaite
University of Cambridge

Daniel M. Roy
University of Toronto

Zoubin Ghahramani
University of Cambridge

MMD for GAN critic

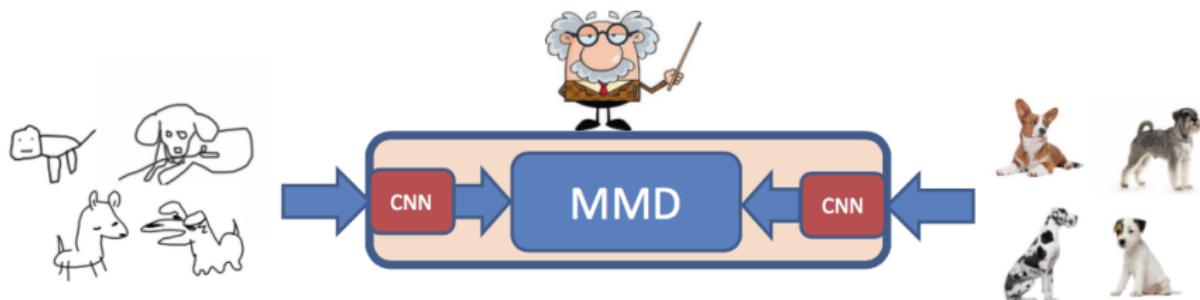
Can you use MMD as a critic to train GANs?



Need better image features.

How to improve the critic witness

- Add convolutional features!
- The **critic** (teacher) also needs to be trained.
- How to regularise?



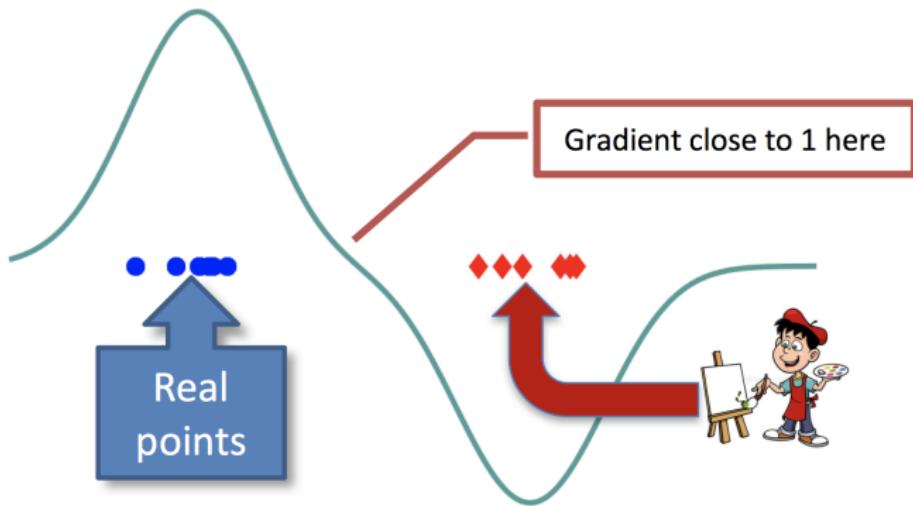
MMD GAN Li et al., [NIPS 2017]

Coulomb GAN Unterthiner et al., [ICLR 2018]

WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]

WGAN-GP Gukrajani et al. [NIPS 2017]



WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]

WGAN-GP Gukrajani et al. [NIPS 2017]



- Given a generator G_θ with parameters θ to be trained.
Samples $Y \sim G_\theta(Z)$ where $Z \sim R$



- Given critic features h_ψ with parameters ψ to be trained. f_ψ a linear function of h_ψ .

WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]

WGAN-GP Gukrajani et al. [NIPS 2017]



- Given a generator G_θ with parameters θ to be trained.

Samples $\textcolor{red}{Y} \sim G_\theta(\textcolor{red}{Z})$ where $\textcolor{red}{Z} \sim \textcolor{red}{R}$



- Given critic features h_ψ with parameters ψ to be trained. f_ψ a linear function of h_ψ .

WGAN-GP gradient penalty:

$$\max_{\psi} \mathbf{E}_{X \sim \textcolor{blue}{P}} f_\psi(\textcolor{blue}{X}) - \mathbf{E}_{Z \sim \textcolor{red}{R}} f_\psi(G_\theta(\textcolor{red}{Z})) + \lambda \mathbf{E}_{\widetilde{X}} \left(\left\| \nabla_{\widetilde{X}} f_\theta(\widetilde{X}) \right\| - 1 \right)^2$$

where

$$\widetilde{X} = \gamma \textcolor{blue}{x}_i + (1 - \gamma) G_\psi(\textcolor{red}{z}_j)$$

$$\gamma \sim \mathcal{U}([0, 1]) \quad x_i \in \{x_\ell\}_{\ell=1}^m \quad z_j \in \{z_\ell\}_{\ell=1}^n$$

The (W)MMD

Train MMD critic features with the witness function gradient penalty

Binkowski, Sutherland, Arbel, G. [ICLR 2018], Bellemare et al. [2017] for energy distance:

$$\max_{\psi} \text{MMD}^2(h_{\psi}(\mathbf{X}), h_{\psi}(G_{\theta}(\mathbf{Z}))) + \lambda \mathbf{E}_{\widetilde{\mathbf{X}}} \left(\|\nabla_{\widetilde{\mathbf{X}}} f_{\psi}(\widetilde{\mathbf{X}})\| - 1 \right)^2$$

where

$$f_{\psi}(\cdot) = \frac{1}{m} \sum_{i=1}^m k(h_{\psi}(\mathbf{x}_i), \cdot) - \frac{1}{n} \sum_{j=1}^n k(h_{\psi}(G_{\theta}(\mathbf{z}_j)), \cdot)$$


$$\widetilde{\mathbf{X}} = \gamma \mathbf{x}_i + (1 - \gamma) G_{\psi}(\mathbf{z}_j)$$

$$\gamma \sim \mathcal{U}([0, 1]) \quad \mathbf{x}_i \in \{\mathbf{x}_{\ell}\}_{\ell=1}^m \quad \mathbf{z}_j \in \{\mathbf{z}_{\ell}\}_{\ell=1}^n$$

Remark by Bottou et al. (2017): gradient penalty modifies the function class. So critic is not an MMD in RKHS \mathcal{F} . 8/71

MMD for GAN critic: revisited

From ICLR 2018:

DEMYSTIFYING MMD GANs

Mikołaj Bińkowski*

Department of Mathematics

Imperial College London

mikbinkowski@gmail.com

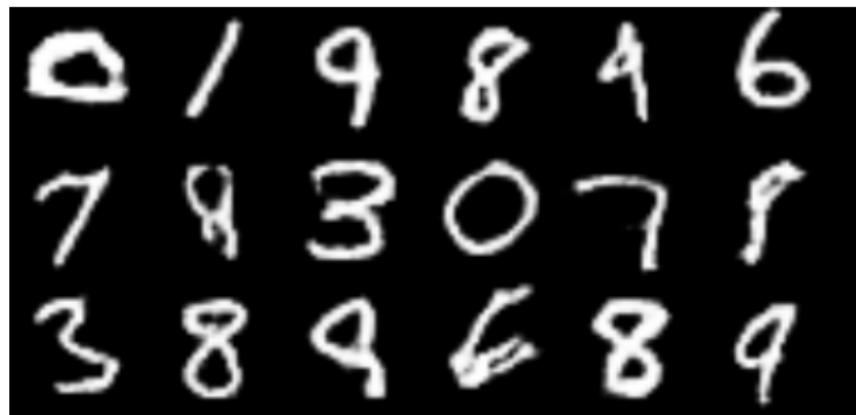
Dougal J. Sutherland,* Michael Arbel & Arthur Gretton

Gatsby Computational Neuroscience Unit

University College London

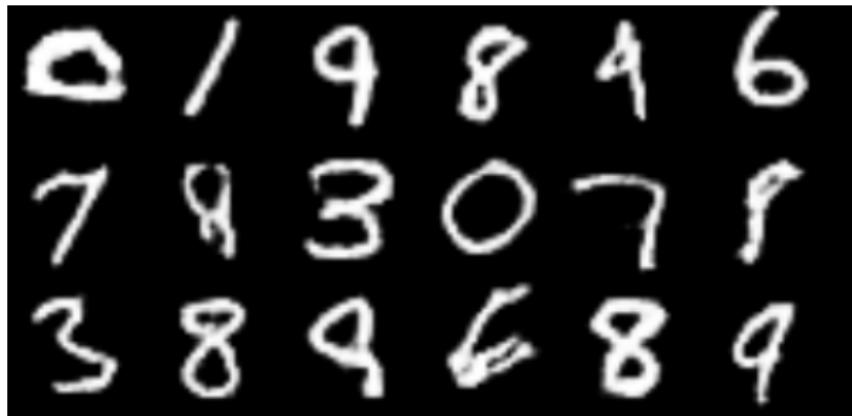
{dougal,michael.n.arbel,arthur.gretton}@gmail.com

MMD for GAN critic: revisited



Samples are better!

MMD for GAN critic: revisited



Samples are better!

Can we do better still?

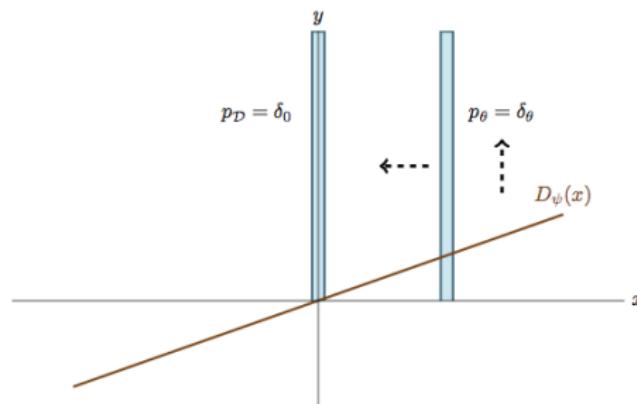
Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution

Nagarajan and Kolter [NIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

The Dirac-GAN

$$P = \delta_0 \quad Q = \delta_\theta \quad f_\psi(x) = \psi \cdot x$$



Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution

Nagarajan and Kolter [NIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

The Dirac-GAN

$$P = \delta_0 \quad Q = \delta_\theta \quad f_\psi(x) = \psi \cdot x$$

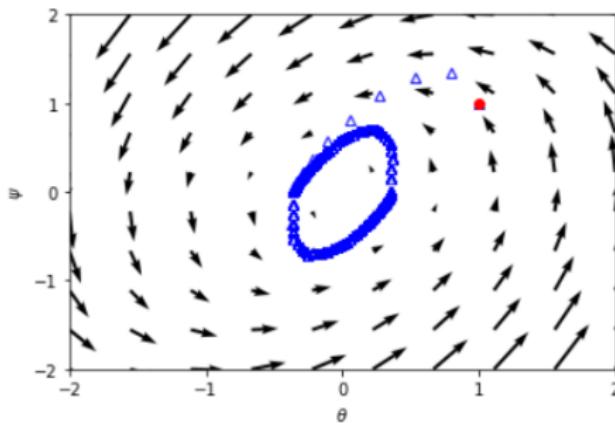


Figure from Mescheder et al. [ICML 2018]

A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)

Arbel, Sutherland, Binkowski, G. [NIPS 2018]

- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

arXiv.org > stat > arXiv:1805.11565

Statistics > Machine Learning

On gradient regularizers for MMD GANs

Michael Arbel, Dougal J. Sutherland, Mikołaj Bińkowski, Arthur Gretton

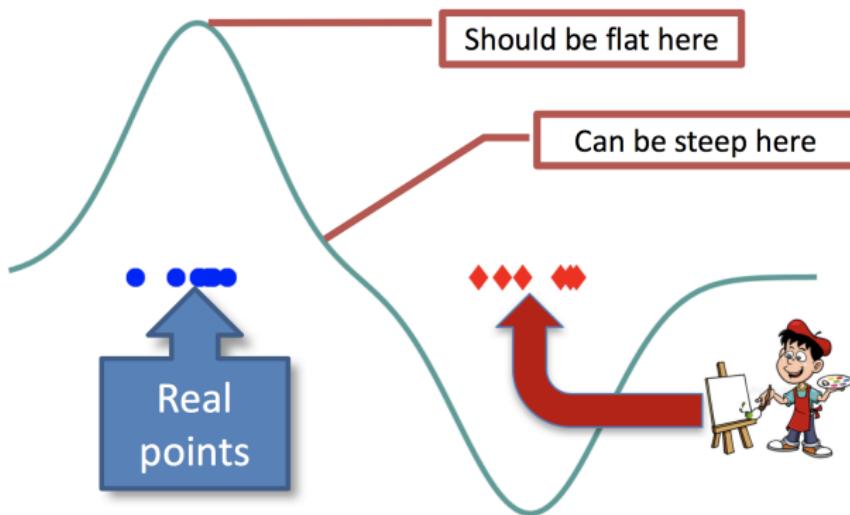
(Submitted on 29 May 2018)

A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)

Arbel, Sutherland, Binkowski, G. [NIPS 2018]

- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]



A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)

Arbel, Sutherland, Binkowski, G. [NIPS 2018]

- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness function:

$$\widetilde{MMD} := \sup_{\|\mathbf{f}\|_S \leq 1} [\mathbb{E}_{\mathbf{P}} \mathbf{f}(\mathbf{X}) - \mathbb{E}_{\mathbf{Q}} \mathbf{f}(\mathbf{Y})]$$

where

$$\|\mathbf{f}\|_S^2 = \|\mathbf{f}\|_{L_2(\mathbf{P})}^2 + \|\nabla \mathbf{f}\|_{L_2(\mathbf{P})}^2 + \lambda \|\mathbf{f}\|_k^2$$

The equation shows the squared norm of \mathbf{f} as the sum of three terms: the squared L_2 norm of \mathbf{f} over the domain \mathbf{P} , the squared gradient norm of \mathbf{f} over the same domain, and a regularization term involving the k -norm of \mathbf{f} . Three orange arrows point upwards from boxes labeled 'L₂ norm control', 'Gradient control', and 'RKHS smoothness' to the corresponding terms in the equation.

A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)

Arbel, Sutherland, Binkowski, G. [NIPS 2018]

- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness function:

$$\widetilde{MMD} := \sup_{\|\mathbf{f}\|_{\mathcal{S}} \leq 1} [\mathbb{E}_{\mathbf{P}} \mathbf{f}(\mathbf{X}) - \mathbb{E}_{\mathbf{Q}} \mathbf{f}(\mathbf{Y})]$$

where

$$\|\mathbf{f}\|_{\mathcal{S}}^2 = \|\mathbf{f}\|_{L_2(\mathbf{P})}^2 + \|\nabla \mathbf{f}\|_{L_2(\mathbf{P})}^2 + \lambda \|\mathbf{f}\|_k^2$$

The equation shows the squared norm of \mathbf{f} in a space \mathcal{S} as the sum of three squared terms. Below the equation, three orange arrows point upwards from three boxes to their respective terms:

- The first arrow points to $\|\mathbf{f}\|_{L_2(\mathbf{P})}^2$ and is labeled "L₂ norm control".
- The second arrow points to $\|\nabla \mathbf{f}\|_{L_2(\mathbf{P})}^2$ and is labeled "Gradient control".
- The third arrow points to $\lambda \|\mathbf{f}\|_k^2$ and is labeled "RKHS smoothness".

Problem: not computationally feasible: $O(n^3)$ per iteration.

A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)

Arbel, Sutherland, Binkowski, G. [NIPS 2018]

- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

The scaled MMD:

$$SMMD = \sigma_{k, P, \lambda} MMD$$

where

$$\sigma_{k, P, \lambda} = \left(\lambda + \int k(x, x) dP(x) + \sum_{i=1}^d \int \partial_i \partial_{i+d} k(x, x) dP(x) \right)^{-1/2}$$

Replace expensive constraint with cheap upper bound:

$$\|\mathbf{f}\|_S^2 \leq \sigma_{k, P, \lambda}^{-1} \|\mathbf{f}\|_k^2$$

A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)
Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

The scaled MMD:

$$SMMD = \sigma_{k, P, \lambda} MMD$$

where

$$\sigma_{k, P, \lambda} = \left(\lambda + \int k(x, x) dP(x) + \sum_{i=1}^d \int \partial_i \partial_{i+d} k(x, x) dP(x) \right)^{-1/2}$$

Replace expensive constraint with cheap upper bound:

$$\|\mathbf{f}\|_S^2 \leq \sigma_{k, P, \lambda}^{-1} \|\mathbf{f}\|_k^2$$

Idea: rather than regularise the critic or witness function, regularise features directly

Evaluation and experiments

Evaluation of GANs

The inception score? Salimans et al. [NIPS 2016]

Based on the classification output $p(y|x)$ of the inception model Szegedy et al. [ICLR 2014],

$$E_X \exp KL(P(y|X) \| P(y)).$$

High when:

- predictive label distribution $P(y|x)$ has low entropy (good quality images)
- label entropy $P(y)$ is high (good variety).

Evaluation of GANs

The inception score? Salimans et al. [NIPS 2016]

Based on the classification output $p(y|x)$ of the inception model Szegedy et al. [ICLR 2014],

$$E_X \exp KL(P(y|X) \| P(y)).$$

High when:

- predictive label distribution $P(y|x)$ has low entropy (good quality images)
- label entropy $P(y)$ is high (good variety).

Problem: relies on a trained classifier! Can't be used on new categories (celeb, bedroom...)

Evaluation of GANs

The Frechet inception distance? Heusel et al. [NIPS 2017]

Fits Gaussians to features in the inception architecture (pool3 layer):

$$FID(\mathcal{P}, \mathcal{Q}) = \|\mu_{\mathcal{P}} - \mu_{\mathcal{Q}}\|^2 + \text{tr}(\Sigma_{\mathcal{P}}) + \text{tr}(\Sigma_{\mathcal{Q}}) - 2\text{tr}\left((\Sigma_{\mathcal{P}}\Sigma_{\mathcal{Q}})^{\frac{1}{2}}\right)$$

where $\mu_{\mathcal{P}}$ and $\Sigma_{\mathcal{P}}$ are the feature mean and covariance of \mathcal{P}

Evaluation of GANs

The Frechet inception distance? Heusel et al. [NIPS 2017]

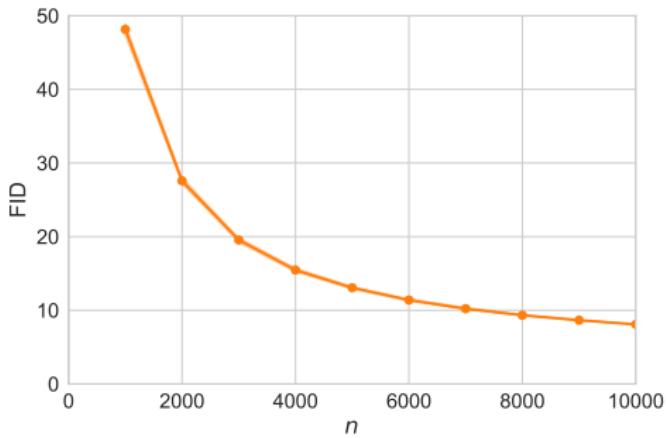
Fits Gaussians to features in the inception architecture (pool3 layer):

$$FID(P, Q) = \|\mu_P - \mu_Q\|^2 + \text{tr}(\Sigma_P) + \text{tr}(\Sigma_Q) - 2\text{tr}\left((\Sigma_P \Sigma_Q)^{\frac{1}{2}}\right)$$

where μ_P and Σ_P are the feature mean and covariance of P

Problem: bias. For finite samples can consistently give incorrect answer.

- Bias demo,
CIFAR-10 train vs
test



Evaluation of GANs

The FID can give the wrong answer in theory.

Assume m samples from P and $n \rightarrow \infty$ samples from Q .

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2 ,

$$FID(\widehat{P}_1, Q) < FID(\widehat{P}_2, Q).$$

Evaluation of GANs

The FID can give the wrong answer in theory.

Assume m samples from P and $n \rightarrow \infty$ samples from Q .

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2 ,

$$FID(\widehat{P}_1, Q) < FID(\widehat{P}_2, Q).$$

Evaluation of GANs

The FID can give the wrong answer in theory.

Assume m samples from P and $n \rightarrow \infty$ samples from Q .

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2 ,

$$FID(\widehat{P}_1, Q) < FID(\widehat{P}_2, Q).$$

Evaluation of GANs

The FID can give the wrong answer in theory.

Assume m samples from P and $n \rightarrow \infty$ samples from Q .

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2 ,

$$FID(\widehat{P_1}, Q) < FID(\widehat{P_2}, Q).$$

Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2I_d)) \quad Q = \text{relu}(\mathcal{N}(1, I_d))$$

where $\Sigma = \frac{4}{d}CC^T$, with C a $d \times d$ matrix with iid standard normal entries.

For a random draw of C :

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

$$FID(\widehat{P}_1, Q) \approx 1133.7 < 1136.2 \approx FID(\widehat{P}_2, Q)$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C .

Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$\textcolor{blue}{P}_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad \textcolor{blue}{P}_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2I_d)) \quad \textcolor{red}{Q} = \text{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d}CC^T$, with C a $d \times d$ matrix with iid standard normal entries.

For a random draw of C :

$$FID(\textcolor{blue}{P}_1, \textcolor{red}{Q}) \approx 1123.0 > 1114.8 \approx FID(\textcolor{blue}{P}_2, \textcolor{red}{Q})$$

With $m = 50\,000$ samples,

$$FID(\widehat{\textcolor{blue}{P}}_1, \textcolor{red}{Q}) \approx 1133.7 < 1136.2 \approx FID(\widehat{\textcolor{blue}{P}}_2, \textcolor{red}{Q})$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C .

Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$\textcolor{blue}{P}_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad \textcolor{blue}{P}_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2I_d)) \quad \textcolor{red}{Q} = \text{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d}CC^T$, with C a $d \times d$ matrix with iid standard normal entries.

For a random draw of C :

$$FID(\textcolor{blue}{P}_1, \textcolor{red}{Q}) \approx 1123.0 > 1114.8 \approx FID(\textcolor{blue}{P}_2, \textcolor{red}{Q})$$

With $m = 50\,000$ samples,

$$FID(\widehat{\textcolor{blue}{P}_1}, \textcolor{red}{Q}) \approx 1133.7 < 1136.2 \approx FID(\widehat{\textcolor{blue}{P}_2}, \textcolor{red}{Q})$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C .

Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$\textcolor{blue}{P}_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad \textcolor{blue}{P}_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2I_d)) \quad \textcolor{red}{Q} = \text{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d}CC^T$, with C a $d \times d$ matrix with iid standard normal entries.

For a random draw of C :

$$FID(\textcolor{blue}{P}_1, \textcolor{red}{Q}) \approx 1123.0 > 1114.8 \approx FID(\textcolor{blue}{P}_2, \textcolor{red}{Q})$$

With $m = 50\,000$ samples,

$$FID(\widehat{\textcolor{blue}{P}_1}, \textcolor{red}{Q}) \approx 1133.7 < 1136.2 \approx FID(\widehat{\textcolor{blue}{P}_2}, \textcolor{red}{Q})$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C .

The kernel inception distance (KID)

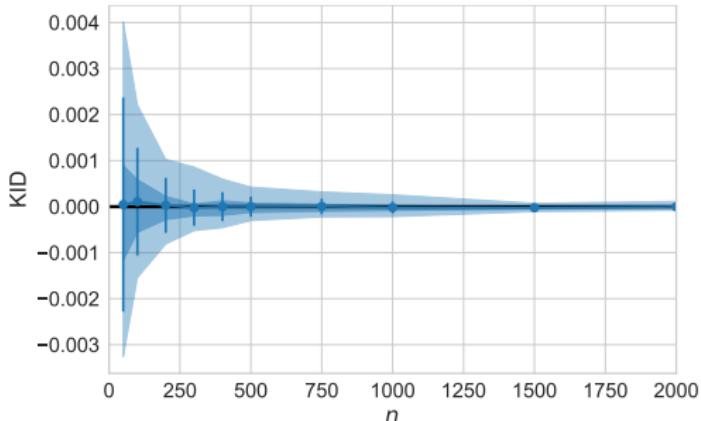
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1 \right)^3.$$

- Checks match for feature means, variances, skewness
- **Unbiased** : eg CIFAR-10 train/test



The kernel inception distance (KID)

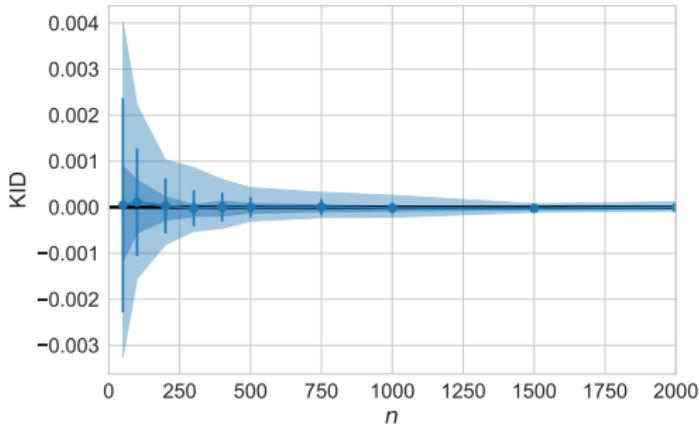
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1 \right)^3.$$

- Checks match for feature means, variances, skewness
- **Unbiased** : eg CIFAR-10 train/test



...“but isn't KID computationally costly?”

The kernel inception distance (KID)

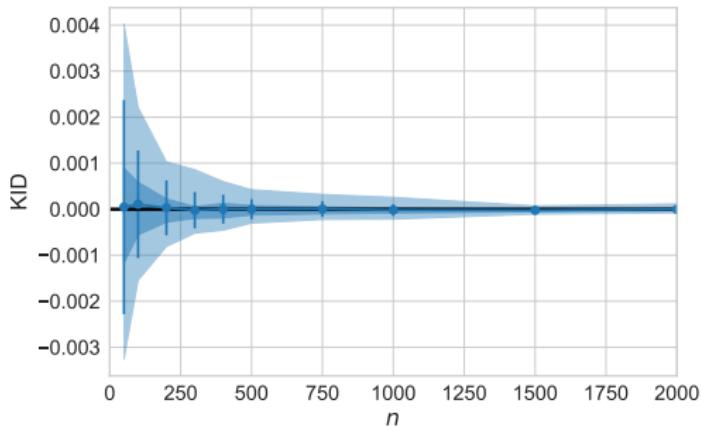
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1 \right)^3.$$

- Checks match for feature means, variances, skewness
- **Unbiased** : eg CIFAR-10 train/test



...“but isn't KID is computationally costly?”

“Block” KID implementation is cheaper than FID: see paper
(or use Tensorflow implementation)!

The kernel inception distance (KID)

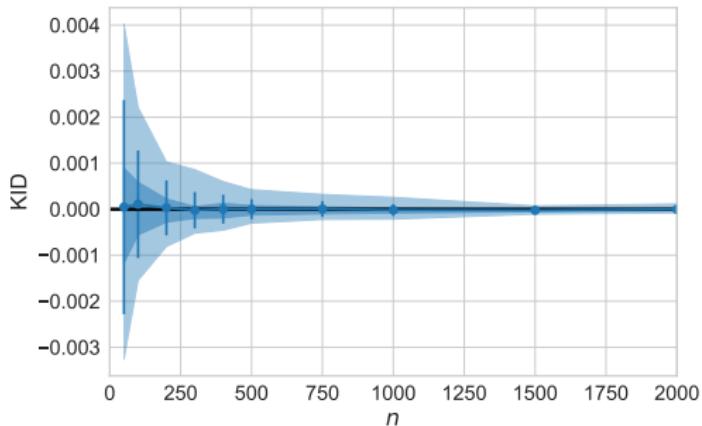
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1 \right)^3.$$

- Checks match for feature means, variances, skewness
- Unbiased : eg CIFAR-10 train/test



Also used for automatic learning rate adjustment: if $KID(\hat{P}_{t+1}, Q)$ not significantly better than $KID(\hat{P}_t, Q)$ then reduce learning rate.

[Bounliphone et al. ICLR 2016]

Benchmarks for comparison (all from ICLR 2018)

SPECTRAL NORMALIZATION FOR GENERATIVE ADVERSARIAL NETWORKS

Takeru Miyato¹, Toshiki Kataoka¹, Masanori Koyama², Yuichi Yoshida³

{miyato, kataoka}@preferred.jp

toyama.masanori@gmail.com

yoshi@li.ac.jp

Preferred Networks, Inc.¹Ritsumeikan University²National Institute of Informatics

We combine with scaled MMD

DEMYSTIFYING MMD GANS

Mikolaj Binkowski*

Department of Mathematics

Imperial College London

mikbinkowski@gmail.com

Dougal J. Sutherland, Michael Arbel & Arthur Gretton

Gatsby Computational Neuroscience Unit

University College London

dougal.sutherland, michael.n.arbel, arthur.gretton@gmail.com

Our ICLR
2018
paper

SOBOLEV GAN

Youssef Mroueh¹, Chun-Liang Li^{2,*}, Tom Sercombe^{1,*}, Anant Raj^{3,*} & Yu Cheng¹

† IBM Research AI

◦ Carnegie Mellon University

◊ Max Planck Institute for Intelligent Systems

* denotes Equal Contribution

{mroueh, chengyu}@us.ibm.com, chunliu@cs.cmu.edu,

tom.sercombe@ibm.com, anant.raj@tuebingen.mpg.de

BOUNDARY-SEEKING GENERATIVE ADVERSARIAL NETWORKS

R Devon Hjelm*

MILA, University of Montréal, IVADO

erroneous@gmail.com

Athul Paul Jacob*

MILA, MSR, University of Waterloo

apjacob@edu.uwaterloo.ca

Tong Che

MILA, University of Montréal

tong.che@umontreal.ca

Adam Trischler

MSR

adam.trischler@microsoft.com

Kyunghyun Cho

New York University

CIFAR Azrieli Global Scholar

kyunghyun.cho@nyu.edu

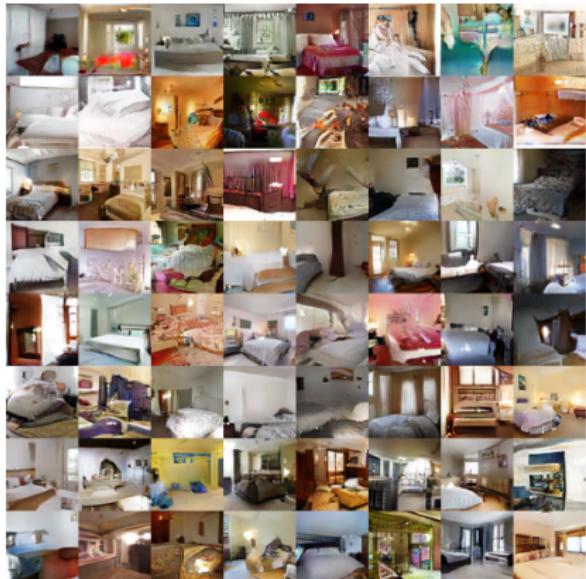
Yoshua Bengio

MILA, University of Montréal, CIFAR, IVADO

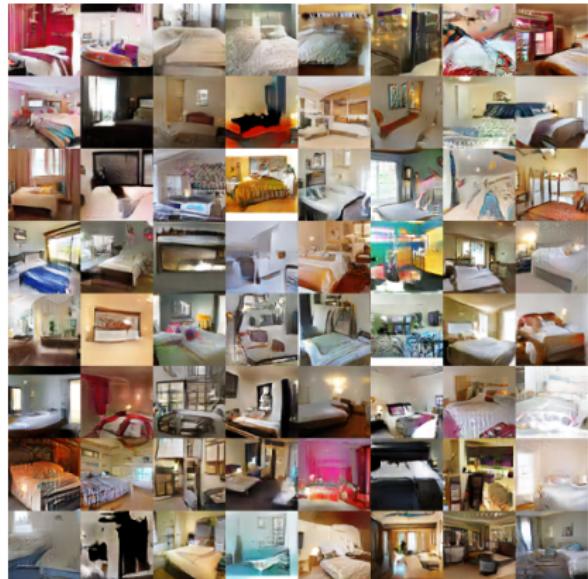
yoshua.bengio@umontreal.ca

Results: what does MMD buy you?

- Critic features from DCGAN: an f -filter critic has f , $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN 64×64 .



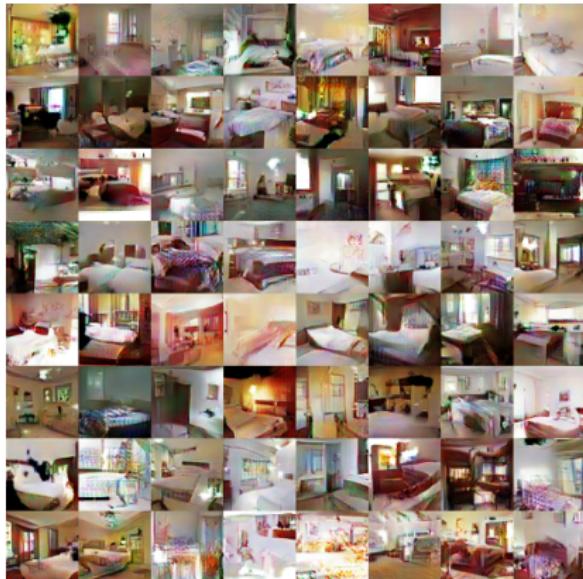
MMD GAN samples, $f = 64$,
FID=32, KID=3



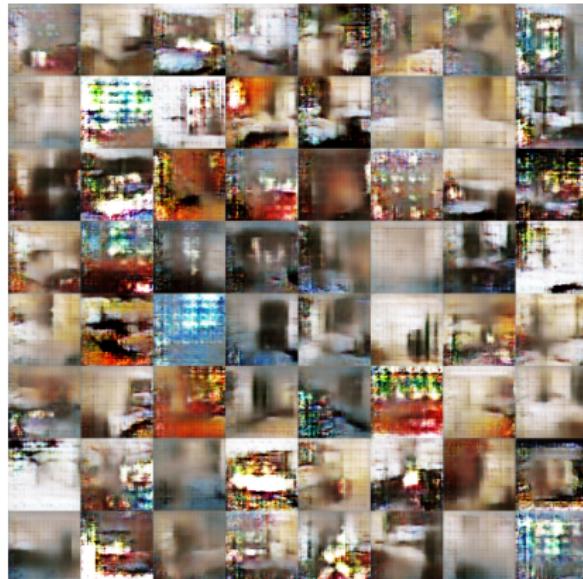
WGAN samples, $f = 64$,
FID=41, KID=4 19/71

Results: what does MMD buy you?

- Critic features from DCGAN: an f -filter critic has f , $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN 64×64 .



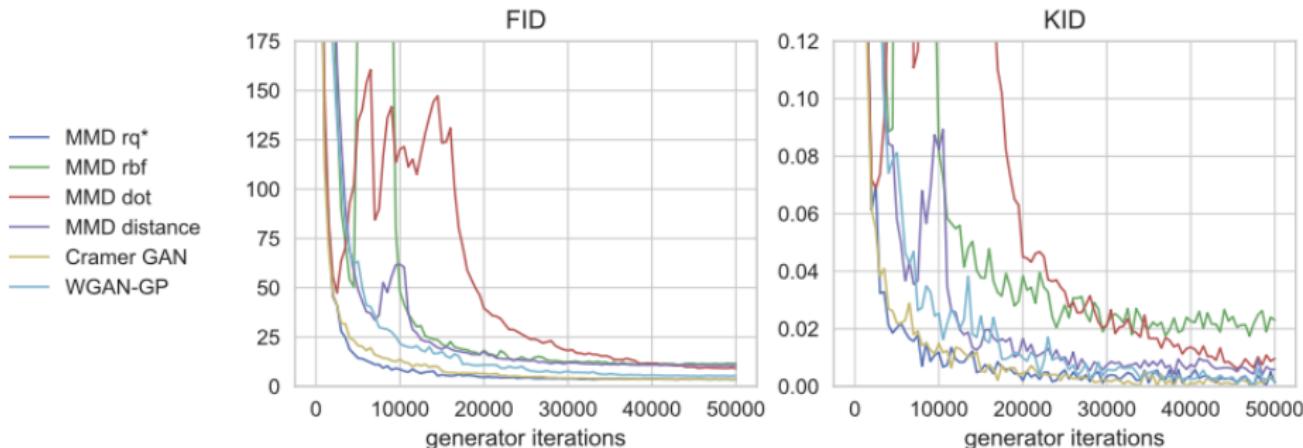
MMD GAN samples, $f = 16$,
FID=86, KID=9



WGAN samples, $f = 16$,
 $f = 64$, FID=293, KID=¹⁹/₇₁

The kernel inception distance (KID)

Faster training: performance scores vs generator iterations on MNIST



Results: celebrity faces 160×160

KID (FID)
scores:

- Sobolev GAN:
14 (20)
- SN-GAN:
18 (28)
- Old MMD
GAN:
13 (21)
- SMMD GAN:
6 (12)

202 599 face images, re-sized and cropped to 160 × 160

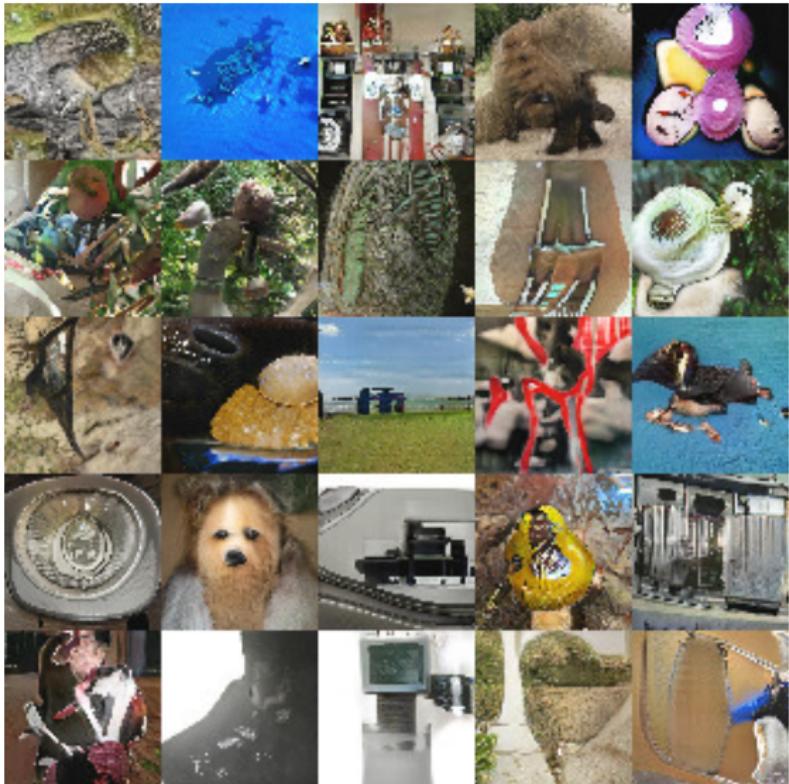


Results: imagenet 64×64

KID (FID)
scores:

- BGAN:
47 (44)
- SN-GAN:
44 (48)
- SMMD GAN:
35 (37)

ILSVRC2012 (ImageNet)
dataset, 1 281 167 im-
ages, resized to 64 × 64.
Around 20 000 classes.

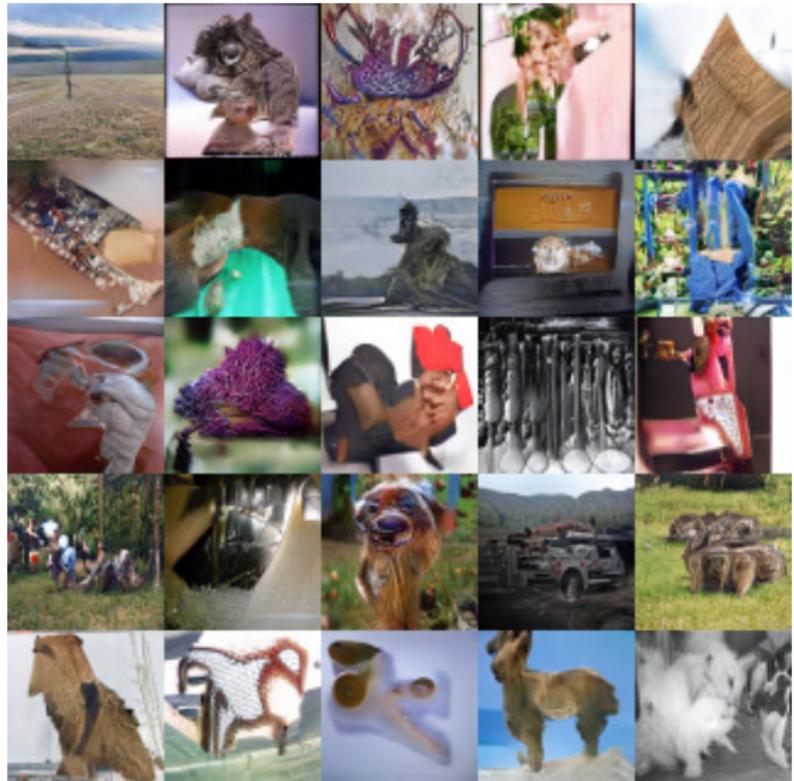


Results: imagenet 64×64

KID (FID)
scores:

- BGAN:
47 (44)
- SN-GAN:
44 (48)
- SMMD GAN:
35 (37)

ILSVRC2012 (ImageNet)
dataset, 1 281 167 im-
ages, resized to 64 × 64.
Around 20 000 classes.



Results: imagenet 64×64

KID (FID)
scores:

- BGAN:
47 (44)
- SN-GAN:
44 (48)
- SMMD GAN:
35 (37)

ILSVRC2012 (ImageNet)
dataset, 1 281 167 im-
ages, resized to 64 × 64.
Around 20 000 classes.



Summary

- MMD critic gives state-of-the-art performance for GAN training (FID and KID)
 - use convolutional input features
 - train with new gradient regulariser
- Faster training, simpler critic network
- Reasons for good performance:
 - Unlike WGAN-GP, MMD loss still a valid critic when features not optimal
 - Kernel features do some of the “work”, so simpler h_ψ features possible.
 - Better gradient/feature regulariser gives better critic

Code for “Demystifying MMD GANs,” ICLR 2018, including KID score: <https://github.com/mbinkowski/MMD-GAN>

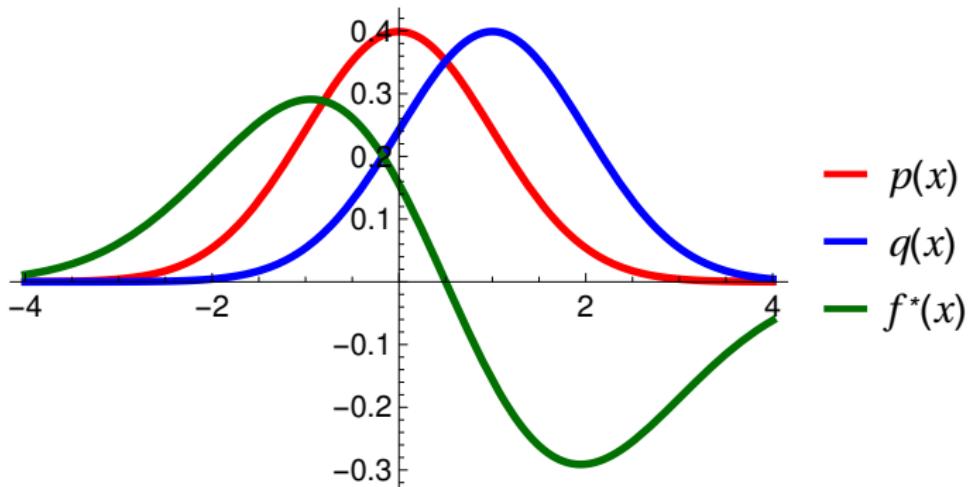
Code for new SMMD:

<https://github.com/MichaelArbel/Scaled-MMD-GAN>

Testing against a probabilistic model

Statistical model criticism

$$MMD(P, Q) = \|f^*\|^2 = \sup_{\|f\|_{\mathcal{F}} \leq 1} [E_Q f - E_P f]$$



$f^*(x)$ is the witness function

Can we compute MMD with samples from Q and a **model** P ?

Problem: usually can't compute $E_P f$ in closed form.

Stein idea

To get rid of $E_{\textcolor{red}{p}} f$ in

$$\sup_{\|f\|_{\mathcal{F}} \leq 1} [E_{\textcolor{blue}{q}} f - E_{\textcolor{red}{p}} f]$$

we define the **Stein operator**

$$[T_{\textcolor{red}{p}} f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x))$$

Then

$$E_P T_{\textcolor{red}{P}} f = 0$$

subject to appropriate boundary conditions. (Oates, Girolami, Chopin, 2016)

Stein idea: proof

$$\begin{aligned} E_{\textcolor{red}{p}} [T_{\textcolor{red}{p}} f] &= \int \left[\frac{1}{\textcolor{red}{p}(x)} \frac{d}{dx} (f(x) \textcolor{red}{p}(x)) \right] \textcolor{red}{p}(x) dx \\ &\quad \int \left[\frac{d}{dx} (f(x) p(x)) \right] dx \\ &= [f(x)p(x)]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Stein idea: proof

$$\begin{aligned} E_{\color{red}p} [T_{\color{red}p} f] &= \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx \\ &\quad \int \left[\frac{d}{dx} (f(x)p(x)) \right] dx \\ &= [f(x)p(x)]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Stein idea: proof

$$\begin{aligned} E_{\color{red}p} [T_{\color{red}p} f] &= \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx \\ &\quad \int \left[\frac{d}{dx} (f(x)p(x)) \right] dx \\ &= [f(x)p(x)]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Stein idea: proof

$$\begin{aligned} E_{\color{red}p} [T_{\color{red}p} f] &= \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx \\ &\quad \int \left[\frac{d}{dx} (f(x)p(x)) \right] dx \\ &= [f(x)p(x)]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Stein idea: proof

$$\begin{aligned} E_{\color{red}p} [T_{\color{red}p} f] &= \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx \\ &\quad \int \left[\frac{d}{dx} (f(x)p(x)) \right] dx \\ &= [f(x)p(x)]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Kernel Stein Discrepancy

Stein operator

$$T_{\textcolor{red}{p}} f = \frac{1}{\textcolor{red}{p}(x)} \frac{d}{dx} (f(x) \textcolor{red}{p}(x))$$

Kernel Stein Discrepancy (KSD)

$$KSD(\textcolor{red}{p}, \textcolor{blue}{q}, \mathcal{F}) = \sup_{\|\textcolor{teal}{g}\|_{\mathcal{F}} \leq 1} E_{\textcolor{blue}{q}} T_{\textcolor{red}{p}} \textcolor{teal}{g} - E_{\textcolor{red}{p}} T_{\textcolor{red}{p}} \textcolor{teal}{g}$$

Kernel Stein Discrepancy

Stein operator

$$T_{\textcolor{red}{p}} f = \frac{1}{\textcolor{red}{p}(x)} \frac{d}{dx} (f(x) \textcolor{red}{p}(x))$$

Kernel Stein Discrepancy (KSD)

$$KSD(\textcolor{red}{p}, \textcolor{blue}{q}, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_{\textcolor{blue}{q}} T_{\textcolor{red}{p}} g - \underline{E}_{\textcolor{red}{p}} T_{\textcolor{red}{p}} \overline{g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_{\textcolor{blue}{q}} T_{\textcolor{red}{p}} g$$

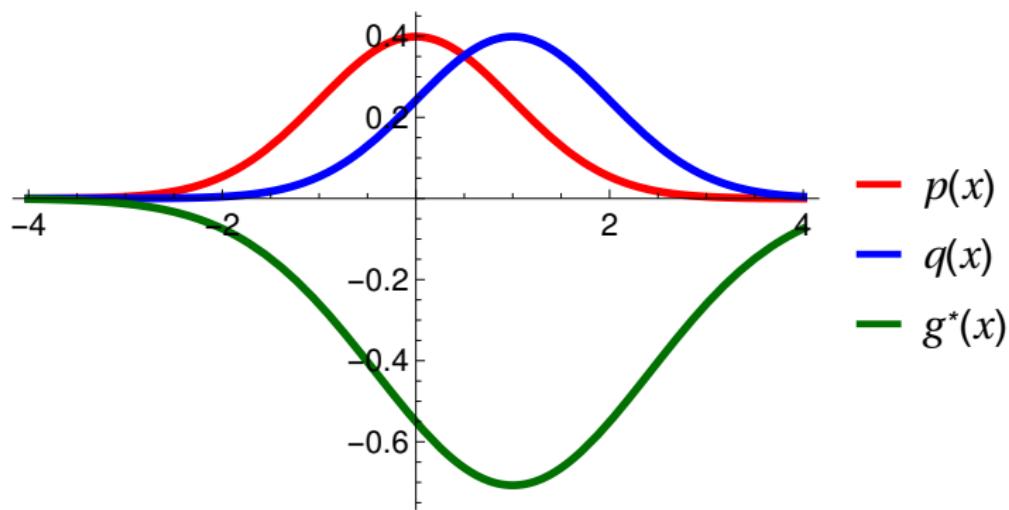
Kernel Stein Discrepancy

Stein operator

$$T_{\textcolor{red}{p}} f = \frac{1}{p(x)} \frac{d}{dx} (f(x) \textcolor{red}{p}(x))$$

Kernel Stein Discrepancy (KSD)

$$KSD(\textcolor{red}{p}, \textcolor{blue}{q}, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_{\textcolor{blue}{q}} T_{\textcolor{red}{p}} g - \cancel{E_{\textcolor{red}{p}} T_{\textcolor{red}{p}} g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_{\textcolor{blue}{q}} T_{\textcolor{red}{p}} g$$



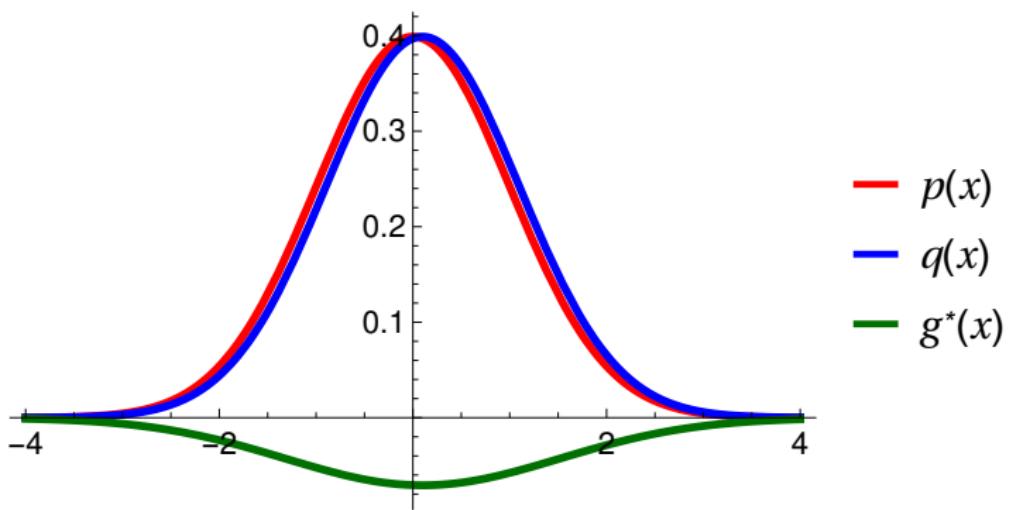
Kernel Stein Discrepancy

Stein operator

$$T_{\textcolor{red}{p}} f = \frac{1}{p(x)} \frac{d}{dx} (f(x) \textcolor{red}{p}(x))$$

Kernel Stein Discrepancy (KSD)

$$KSD(\textcolor{red}{p}, \textcolor{blue}{q}, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_{\textcolor{blue}{q}} T_{\textcolor{red}{p}} g - \cancel{E_{\textcolor{red}{p}} T_{\textcolor{red}{p}} g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_{\textcolor{blue}{q}} T_{\textcolor{red}{p}} g$$



Kernel stein discrepancy

Closed-form expression for KSD: given $Z, Z' \sim \textcolor{blue}{q}$, then

(Chwialkowski, Strathmann, G., ICML 2016) (Liu, Lee, Jordan ICML 2016)

$$\text{KSD}(\textcolor{red}{p}, \textcolor{blue}{q}, \mathcal{F}) = E_{\textcolor{blue}{q}} h_{\textcolor{red}{p}}(Z, Z')$$

where

$$\begin{aligned} h_{\textcolor{red}{p}}(x, y) := & \partial_x \log \textcolor{red}{p}(x) \partial_x \log \textcolor{red}{p}(y) k(x, y) \\ & + \partial_y \log \textcolor{red}{p}(y) \partial_x k(x, y) \\ & + \partial_x \log \textcolor{red}{p}(x) \partial_y k(x, y) \\ & + \partial_x \partial_y k(x, y) \end{aligned}$$

and k is RKHS kernel for \mathcal{F}

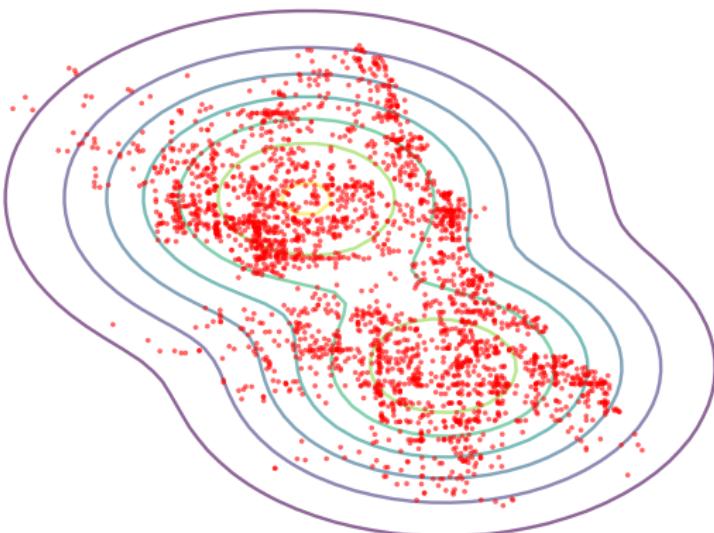
Only depends on kernel and $\partial_x \log \textcolor{red}{p}(x)$. Do not need to normalize $\textcolor{red}{p}$, or sample from it.

Statistical model criticism



Chicago crime data

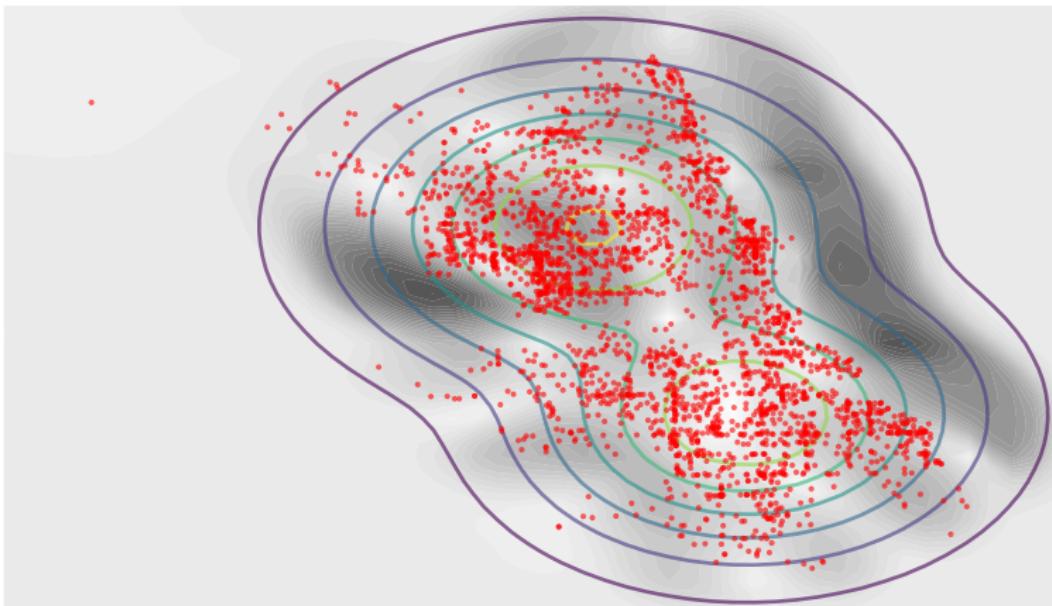
Statistical model criticism



Chicago crime data

Model is Gaussian mixture with **two** components.

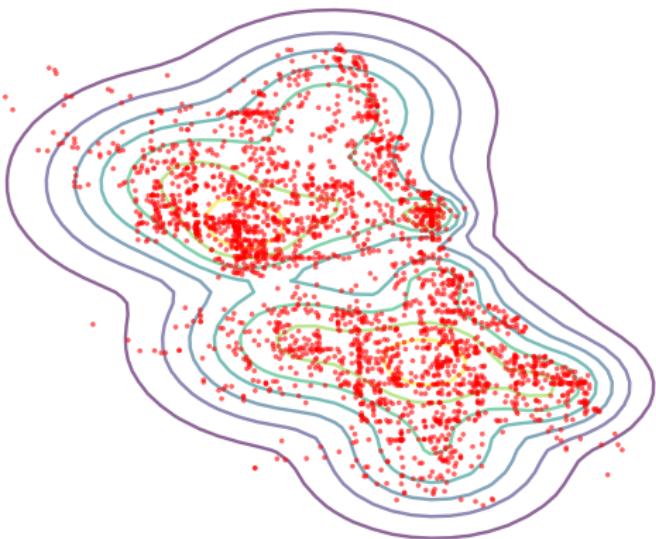
Statistical model criticism



Chicago crime data

Model is Gaussian mixture with **two** components
Stein witness function

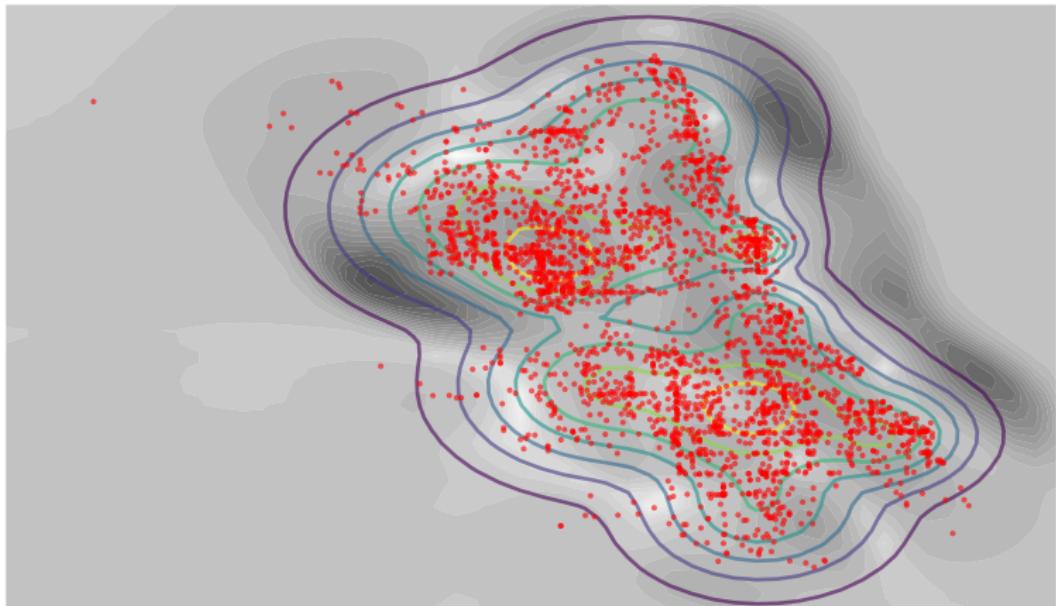
Statistical model criticism



Chicago crime data

Model is Gaussian mixture with **ten** components.

Statistical model criticism



Chicago crime data

Model is Gaussian mixture with **ten** components

Stein witness function

Code: https://github.com/karlnapf/kernel_goodness_of_fit

Kernel stein discrepancy

Further applications:

- Evaluation of approximate MCMC methods.
(Chwialkowski, Strathmann, G., ICML 2016; Gorham, Mackey, ICML 2017)

What kernel to use?

- The inverse multiquadric kernel,

$$k(x, y) = \left(c + \|x - y\|_2^2 \right)^\beta$$

for $\beta \in (-1, 0)$.

The image shows a screenshot of an arXiv.org page. The URL in the address bar is "arXiv.org > stat > arXiv:1703.01717". The page title is "Measuring Sample Quality with Kernels" by Jackson Gorham, Lester Mackey. It is categorized under "Statistics > Machine Learning". The conference information indicates it was presented at "ICML 2017". The submission date is "Submitted on 6 Mar 2017 (v1), last revised 3 Aug 2017 (this version, v6)".

Testing statistical dependence

Dependence testing

- Given: Samples from a distribution $P_{X Y}$
- Goal: Are X and Y independent?

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

Text from dogtime.com and petfinder.com

MMD as a dependence measure?

Could we use MMD?

$$MMD(\underbrace{P_{XY}}_P, \underbrace{P_X P_Y}_Q, \mathcal{H}_\kappa)$$

- We don't have samples from $\mathcal{Q} := P_X P_Y$, only pairs $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$
 - Solution: simulate \mathcal{Q} with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS \mathcal{H}_κ ?

MMD as a dependence measure?

Could we use MMD?

$$MMD(\underbrace{P_{XY}}_P, \underbrace{P_X P_Y}_Q, \mathcal{H}_\kappa)$$

- We don't have samples from $\textcolor{red}{Q} := P_X P_Y$, only pairs $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$
 - **Solution:** simulate $\textcolor{red}{Q}$ with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS \mathcal{H}_κ ?

MMD as a dependence measure?

Could we use MMD?

$$MMD(\underbrace{P_{XY}}_P, \underbrace{P_X P_Y}_Q, \mathcal{H}_\kappa)$$

- We don't have samples from $\textcolor{red}{Q} := P_X P_Y$, only pairs $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$
 - **Solution:** simulate $\textcolor{red}{Q}$ with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS \mathcal{H}_κ ?

MMD as a dependence measure

Kernel k on **images** with feature space \mathcal{F} ,

$$k(\text{dog}, \text{cat})$$

Kernel l on **captions** with feature space \mathcal{G} ,

$$l(\boxed{\text{A large animal who slings slobber, ...}}, \boxed{\text{A responsive, interactive pet --}})$$

MMD as a dependence measure

Kernel k on **images** with feature space \mathcal{F} ,

$$k\left(\text{dog} , \text{cat}\right)$$

Kernel l on **captions** with feature space \mathcal{G} ,

$$l\left(\boxed{\text{A large animal who slings slobber, ...}} , \boxed{\text{A responsive, interactive pet, ...}}\right)$$

Kernel κ on **image-text pairs**: are images and captions similar?

$$\kappa\left(\text{dog} , \boxed{\text{A large animal who slings slobber, ...}} , \text{cat} , \boxed{\text{A responsive, interactive pet, ...}}\right)$$

$$= k\left(\text{dog} , \text{cat}\right) \times l\left(\boxed{\text{A large animal who slings slobber, ...}} , \boxed{\text{A responsive, interactive pet, ...}}\right)$$

MMD as a dependence measure

- Given: Samples from a distribution $P_{\textcolor{blue}{X} \textcolor{red}{Y}}$
- Goal: Are $\textcolor{blue}{X}$ and $\textcolor{red}{Y}$ independent?

$$MMD^2(\hat{P}_{XY}, \hat{P}_X \hat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2} \text{trace}(\textcolor{blue}{K} \textcolor{red}{L})$$

($\textcolor{blue}{K}$, $\textcolor{red}{L}$ column centered)

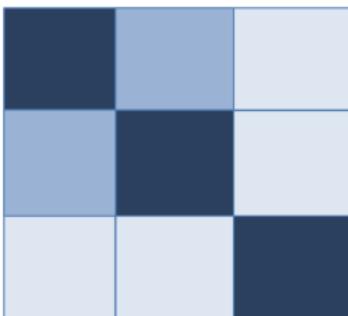
MMD as a dependence measure

- Given: Samples from a distribution P_{XY}
- Goal: Are X and Y independent?

$$MMD^2(\hat{P}_{XY}, \hat{P}_X \hat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2} \text{trace}(\mathbf{K} \mathbf{L})$$

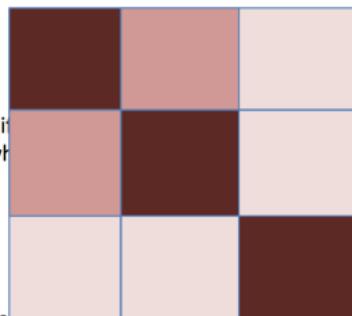


K



A large animal who slings slobber, exudes a distinctive houndy odor, ...

L



Their noses guide them through life and they're never happier than when following an interesting scent.

A responsive, interactive pet, one that will blow in your ear and follow you everywhere.

MMD as a dependence measure

Two questions:

- Why the product kernel? Many ways to combine kernels - why not eg a sum?
- Is there a more interpretable way of defining this dependence measure?

Illustration: dependence \neq correlation

- Given: Samples from a distribution $P_{X,Y}$
- Goal: Are X and Y dependent?

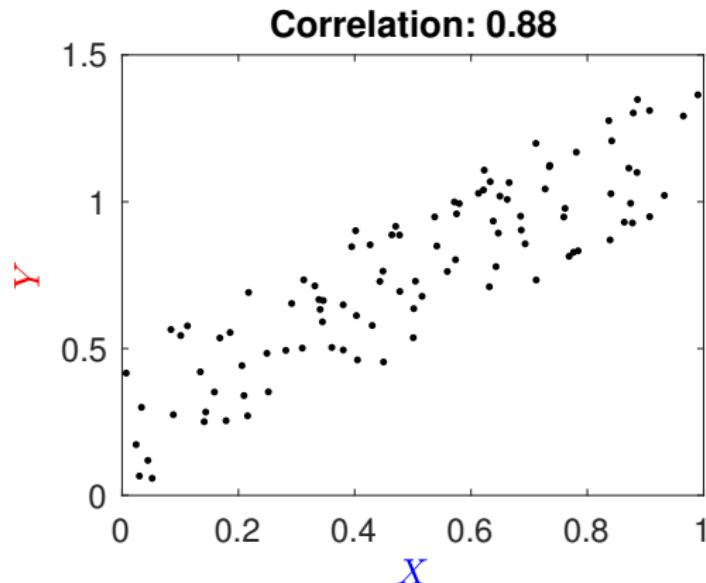


Illustration: dependence \neq correlation

- Given: Samples from a distribution $P_{X,Y}$
- Goal: Are X and Y dependent?

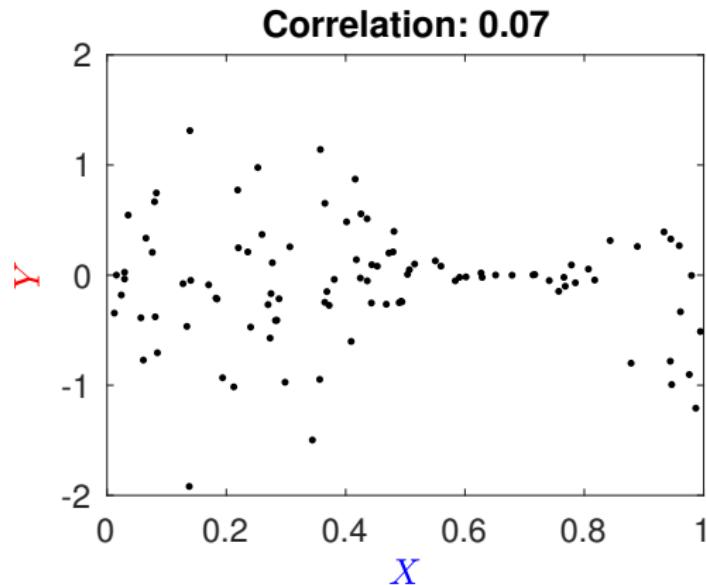
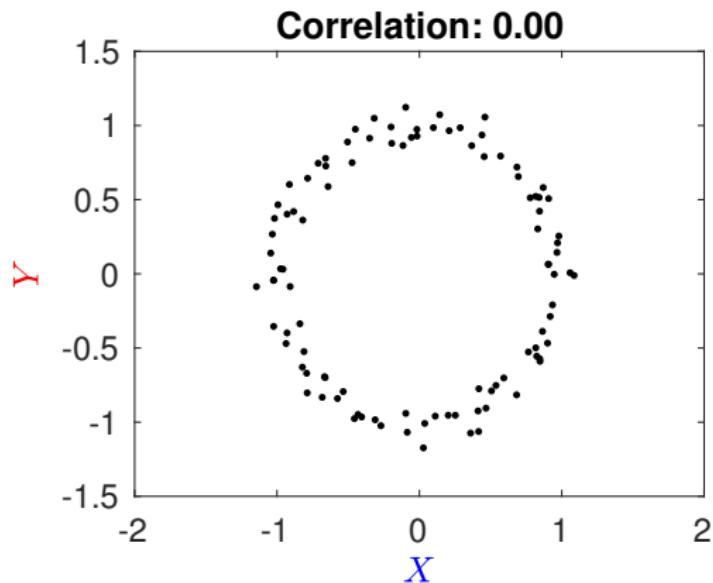


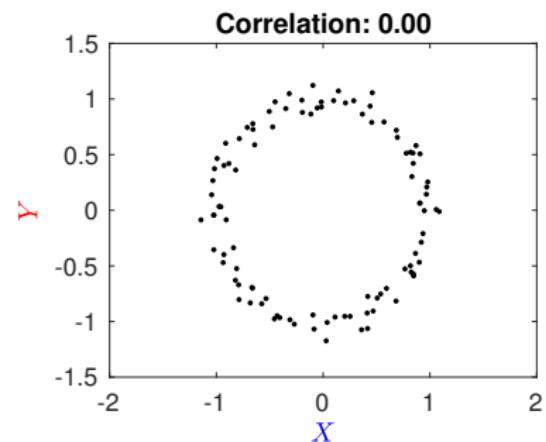
Illustration: dependence \neq correlation

- Given: Samples from a distribution $P_{X,Y}$
- Goal: Are X and Y dependent?



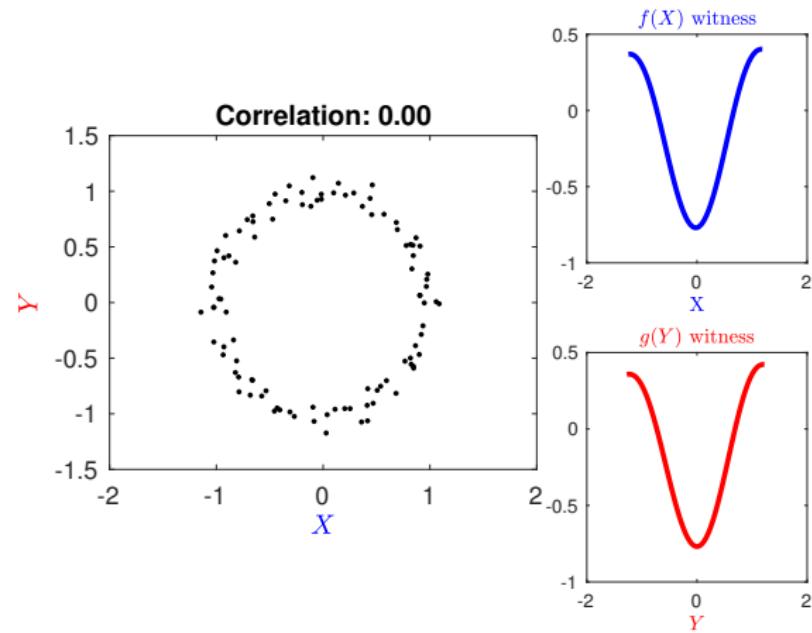
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



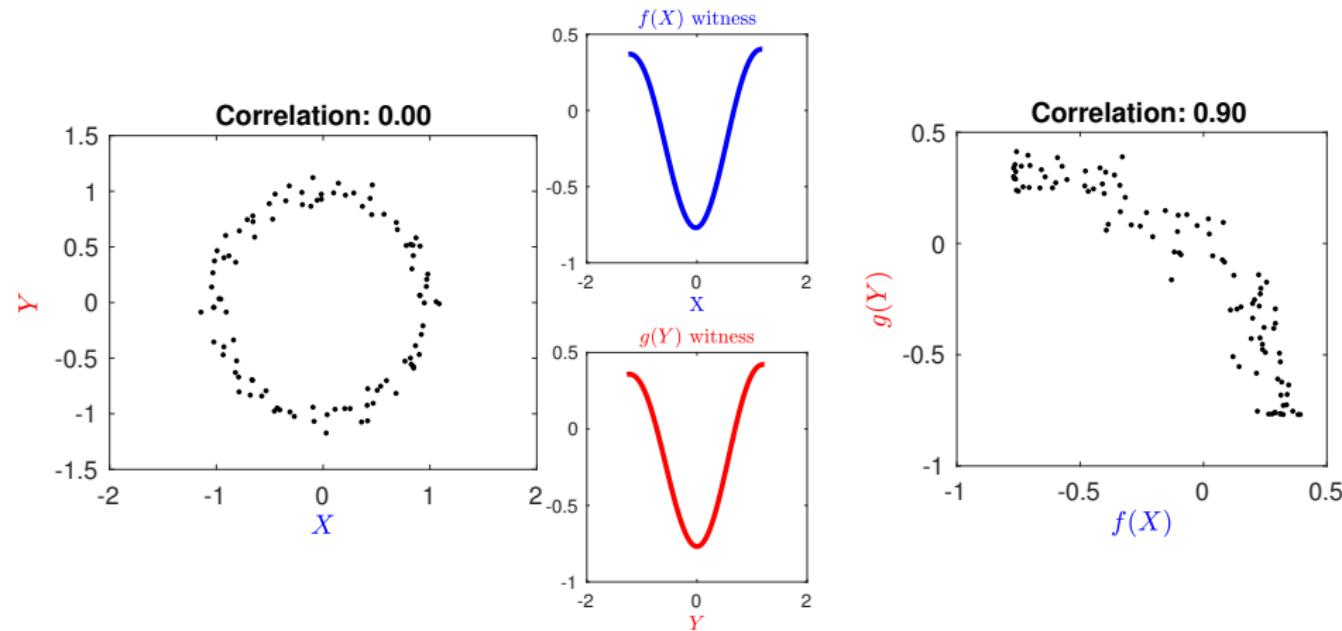
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.

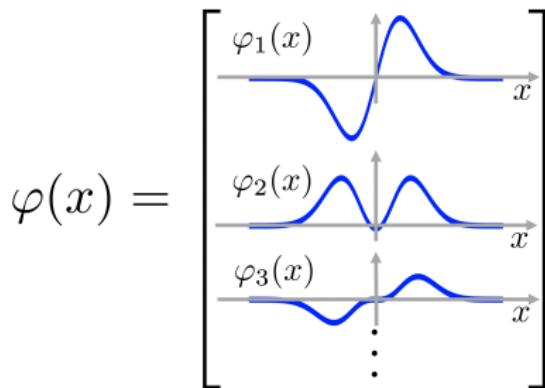


Define two spaces, one for each witness

Function in \mathcal{F}

$$f(x) = \sum_{j=1}^{\infty} f_j \varphi_j(x)$$

Feature map



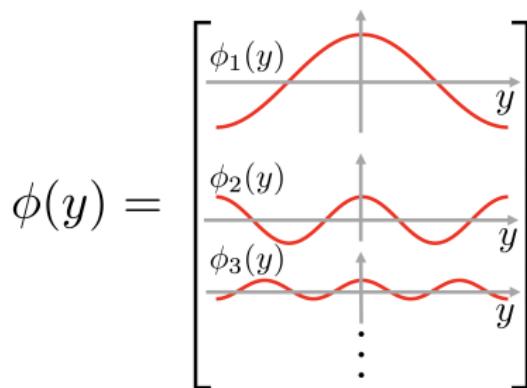
Kernel for RKHS \mathcal{F} on \mathcal{X} :

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

Function in \mathcal{G}

$$g(y) = \sum_{j=1}^{\infty} g_j \phi_j(y)$$

Feature map



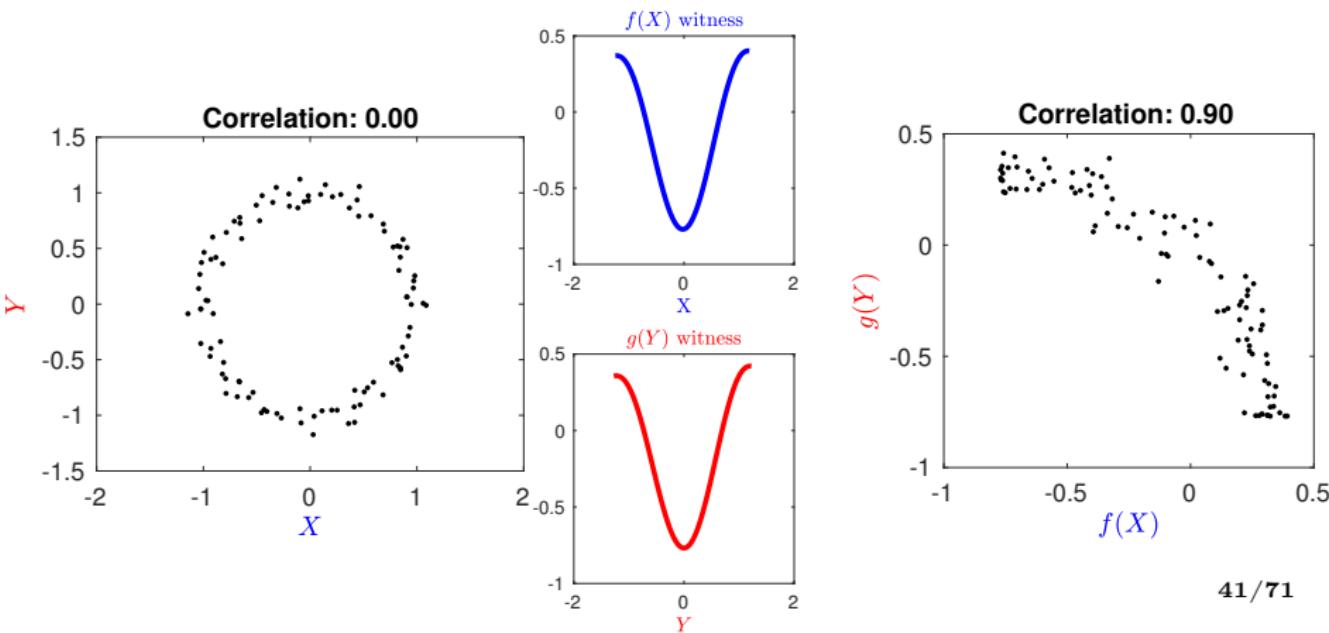
Kernel for RKHS \mathcal{G} on \mathcal{Y} :

$$l(x, x') = \langle \phi(y), \phi(y') \rangle_{\mathcal{G}}$$

The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\begin{array}{l} \|\mathbf{f}\|_{\mathcal{F}} \leq 1 \\ \|\mathbf{g}\|_{\mathcal{G}} \leq 1 \end{array}} \text{cov}[\mathbf{f}(x)\mathbf{g}(y)]$$



The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\begin{array}{l} \|\textcolor{blue}{f}\|_{\mathcal{F}} \leq 1 \\ \|\textcolor{red}{g}\|_{\mathcal{G}} \leq 1 \end{array}} \text{cov} \left[\left(\sum_{j=1}^{\infty} \textcolor{blue}{f}_j \varphi_j(x) \right) \left(\sum_{j=1}^{\infty} \textcolor{red}{g}_j \phi_j(y) \right) \right]$$

The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\begin{array}{l} \|\textcolor{blue}{f}\|_{\mathcal{F}} \leq 1 \\ \|\textcolor{red}{g}\|_{\mathcal{G}} \leq 1 \end{array}} E_{xy} \left[\left(\sum_{j=1}^{\infty} \textcolor{blue}{f}_j \varphi_j(x) \right) \left(\sum_{j=1}^{\infty} \textcolor{red}{g}_j \phi_j(y) \right) \right]$$

Fine print: feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.

The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\begin{array}{l} \|\mathbf{f}\|_{\mathcal{F}} \leq 1 \\ \|\mathbf{g}\|_{\mathcal{G}} \leq 1 \end{array}} E_{xy} \left[\left(\sum_{j=1}^{\infty} \mathbf{f}_j \varphi_j(x) \right) \left(\sum_{j=1}^{\infty} \mathbf{g}_j \phi_j(y) \right) \right]$$

Fine print: feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.

Rewriting:

$$E_{xy}[\mathbf{f}(x)\mathbf{g}(y)] = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \end{bmatrix}^\top \underbrace{\mathbf{E}_{xy} \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \end{bmatrix} \begin{bmatrix} \phi_1(y) & \phi_2(y) & \dots \end{bmatrix}}_{C_{\varphi(x)\phi(y)}} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \end{bmatrix}$$

The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\begin{array}{l} \|\mathbf{f}\|_{\mathcal{F}} \leq 1 \\ \|\mathbf{g}\|_{\mathcal{G}} \leq 1 \end{array}} E_{xy} \left[\left(\sum_{j=1}^{\infty} \mathbf{f}_j \varphi_j(x) \right) \left(\sum_{j=1}^{\infty} \mathbf{g}_j \phi_j(y) \right) \right]$$

Fine print: feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.

Rewriting:

$$E_{xy}[\mathbf{f}(x)\mathbf{g}(y)] = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \end{bmatrix}^\top \underbrace{\mathbf{E}_{xy} \left(\begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \end{bmatrix} \begin{bmatrix} \phi_1(y) & \phi_2(y) & \dots \end{bmatrix} \right)}_{C_{\varphi(x)\phi(y)}} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \end{bmatrix}$$

COCO: max singular value of feature covariance $C_{\varphi(x)\phi(y)}$ 41/71

Computing COCO in practice

Given sample $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical \widehat{COCO} ?

Computing COCO in practice

Given sample $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical \widehat{COCO} ?

\widehat{COCO} is largest eigenvalue γ_{\max} of

$$\begin{bmatrix} 0 & \frac{1}{n}KL \\ \frac{1}{n}LK & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} K & 0 \\ 0 & L \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$.

Fine print: kernels are computed with empirically centered features $\varphi(x) - \frac{1}{n} \sum_{i=1}^n \varphi(x_i)$ and $\phi(y) - \frac{1}{n} \sum_{i=1}^n \phi(y_i)$.

G., Smola., Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Schoelkopf, and Logothetis,
AISTATS'05

Computing COCO in practice

Given sample $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical \widehat{COCO} ?

\widehat{COCO} is largest eigenvalue γ_{\max} of

$$\begin{bmatrix} 0 & \frac{1}{n}KL \\ \frac{1}{n}LK & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} K & 0 \\ 0 & L \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$.

Witness functions (singular vectors):

$$f(x) \propto \sum_{i=1}^n \alpha_i k(x_i, x) \quad g(y) \propto \sum_{i=1}^n \beta_i l(y_i, y)$$

Fine print: kernels are computed with empirically centered features $\varphi(x) - \frac{1}{n} \sum_{i=1}^n \varphi(x_i)$ and $\phi(y) - \frac{1}{n} \sum_{i=1}^n \phi(y_i)$.

G., Smola., Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Schoelkopf, and Logothetis,
AISTATS'05

Empirical COCO: proof (1)

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \underbrace{\frac{1}{n} \sum_{i=1}^n [f(x_i)g(y_i)]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2} (\|f\|_{\mathcal{F}}^2 - 1)}_{\text{smoothness constraints}} - \frac{\gamma}{2} (\|g\|_{\mathcal{G}}^2 - 1).$$

Fine print: $f(x_i)g(y_i)$ centered to have zero empirical mean.

Empirical COCO: proof (1)

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \underbrace{\frac{1}{n} \sum_{i=1}^n [f(x_i)g(y_i)]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2} (\|f\|_{\mathcal{F}}^2 - 1)}_{\text{smoothness constraints}} - \frac{\gamma}{2} (\|g\|_{\mathcal{G}}^2 - 1).$$

Fine print: $f(x_i)g(y_i)$ centered to have zero empirical mean.

Assume (cf representer theorem):

$$f = \sum_{i=1}^n \alpha_i \varphi(x_i) \quad g = \sum_{i=1}^n \beta_i \psi(y_i)$$

for centered $\varphi(x_i), \psi(y_i)$.

Empirical COCO: proof (1)

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \underbrace{\frac{1}{n} \sum_{i=1}^n [f(x_i)g(y_i)]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2} (\|f\|_{\mathcal{F}}^2 - 1)}_{\text{smoothness constraints}} - \frac{\gamma}{2} (\|g\|_{\mathcal{G}}^2 - 1).$$

Fine print: $f(x_i)g(y_i)$ centered to have zero empirical mean.

Assume (cf representer theorem):

$$f = \sum_{i=1}^n \alpha_i \varphi(x_i) \quad g = \sum_{i=1}^n \beta_i \psi(y_i)$$

for centered $\varphi(x_i), \psi(y_i)$.

First step is smoothness constraint:

$$\|f\|_{\mathcal{F}}^2 - 1 = \langle f, f \rangle_{\mathcal{F}} - 1$$

Empirical COCO: proof (1)

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \underbrace{\frac{1}{n} \sum_{i=1}^n [f(x_i)g(y_i)]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2} (\|f\|_{\mathcal{F}}^2 - 1)}_{\text{smoothness constraints}} - \frac{\gamma}{2} (\|g\|_{\mathcal{G}}^2 - 1).$$

Fine print: $f(x_i)g(y_i)$ centered to have zero empirical mean.

Assume (cf representer theorem):

$$f = \sum_{i=1}^n \alpha_i \varphi(x_i) \quad g = \sum_{i=1}^n \beta_i \psi(y_i)$$

for centered $\varphi(x_i)$, $\psi(y_i)$.

First step is smoothness constraint:

$$\begin{aligned} \|f\|_{\mathcal{F}}^2 - 1 &= \langle f, f \rangle_{\mathcal{F}} - 1 \\ &= \left\langle \sum_{i=1}^n \alpha_i \varphi(x_i), \sum_{i=1}^n \alpha_i \varphi(x_i) \right\rangle_{\mathcal{F}} - 1 \end{aligned}$$

Empirical COCO: proof (1)

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \underbrace{\frac{1}{n} \sum_{i=1}^n [f(x_i)g(y_i)]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2} (\|f\|_{\mathcal{F}}^2 - 1)}_{\text{smoothness constraints}} - \frac{\gamma}{2} (\|g\|_{\mathcal{G}}^2 - 1).$$

Fine print: $f(x_i)g(y_i)$ centered to have zero empirical mean.

Assume (cf representer theorem):

$$f = \sum_{i=1}^n \alpha_i \varphi(x_i) \quad g = \sum_{i=1}^n \beta_i \psi(y_i)$$

for centered $\varphi(x_i)$, $\psi(y_i)$.

First step is smoothness constraint:

$$\begin{aligned} \|f\|_{\mathcal{F}}^2 - 1 &= \langle f, f \rangle_{\mathcal{F}} - 1 \\ &= \left\langle \sum_{i=1}^n \alpha_i \varphi(x_i), \sum_{i=1}^n \alpha_i \varphi(x_i) \right\rangle_{\mathcal{F}} - 1 \\ &= \alpha^T K \alpha - 1 \end{aligned}$$

Proof sketch (2)

Second step is covariance:

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n [\mathbf{f}(x_i) \mathbf{g}(y_i)] &= \frac{1}{n} \sum_{i=1}^n \langle \mathbf{f}, \varphi(x_i) \rangle_{\mathcal{F}} \langle \mathbf{g}, \varphi(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \sum_{i=1}^n \left\langle \sum_{\ell=1}^n \alpha_{\ell} \varphi(x_{\ell}), \varphi(x_i) \right\rangle_{\mathcal{F}} \langle \mathbf{g}, \varphi(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \boldsymbol{\alpha}^{\top} K L \boldsymbol{\beta}\end{aligned}$$

Proof sketch (2)

Second step is covariance:

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n [\mathbf{f}(x_i) \mathbf{g}(y_i)] &= \frac{1}{n} \sum_{i=1}^n \langle \mathbf{f}, \varphi(x_i) \rangle_{\mathcal{F}} \langle \mathbf{g}, \varphi(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \sum_{i=1}^n \left\langle \sum_{\ell=1}^n \alpha_{\ell} \varphi(x_{\ell}), \varphi(x_i) \right\rangle_{\mathcal{F}} \langle \mathbf{g}, \varphi(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \boldsymbol{\alpha}^T K L \boldsymbol{\beta}\end{aligned}$$

Proof sketch (2)

Second step is covariance:

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n [\mathbf{f}(x_i) \mathbf{g}(y_i)] &= \frac{1}{n} \sum_{i=1}^n \langle \mathbf{f}, \varphi(x_i) \rangle_{\mathcal{F}} \langle \mathbf{g}, \varphi(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \sum_{i=1}^n \left\langle \sum_{\ell=1}^n \alpha_{\ell} \varphi(x_{\ell}), \varphi(x_i) \right\rangle_{\mathcal{F}} \langle \mathbf{g}, \varphi(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \boldsymbol{\alpha}^{\top} K L \boldsymbol{\beta}\end{aligned}$$

where $K_{ij} = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{F}}$ $L_{ij} = l(y_i, y_j)$.

Proof sketch (2)

Second step is covariance:

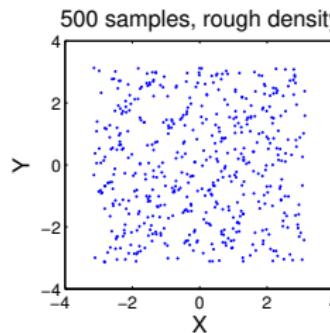
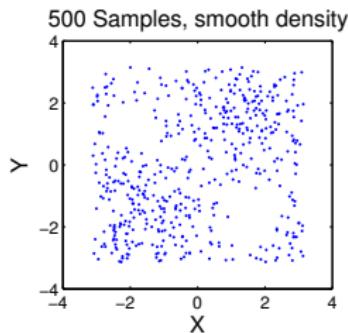
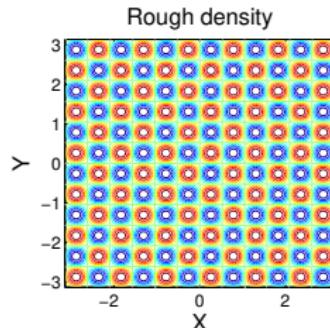
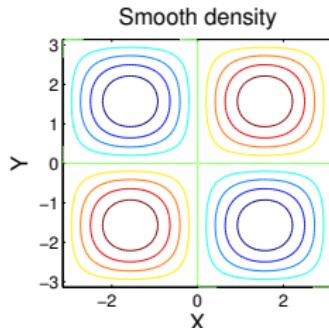
$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n [\mathbf{f}(x_i) \mathbf{g}(y_i)] &= \frac{1}{n} \sum_{i=1}^n \langle \mathbf{f}, \varphi(x_i) \rangle_{\mathcal{F}} \langle \mathbf{g}, \varphi(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \sum_{i=1}^n \left\langle \sum_{\ell=1}^n \alpha_{\ell} \varphi(x_{\ell}), \varphi(x_i) \right\rangle_{\mathcal{F}} \langle \mathbf{g}, \varphi(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \boldsymbol{\alpha}^T K L \boldsymbol{\beta}\end{aligned}$$

where $K_{ij} = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{F}}$ $L_{ij} = l(y_i, y_j)$.

The Lagrangian is now:

$$\mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \boldsymbol{\alpha}^T K L \boldsymbol{\beta} - \frac{\lambda}{2} (\boldsymbol{\alpha}^T K \boldsymbol{\alpha} - 1) - \frac{\gamma}{2} (\boldsymbol{\beta}^T L \boldsymbol{\beta} - 1)$$

What is a large dependence with COCO?



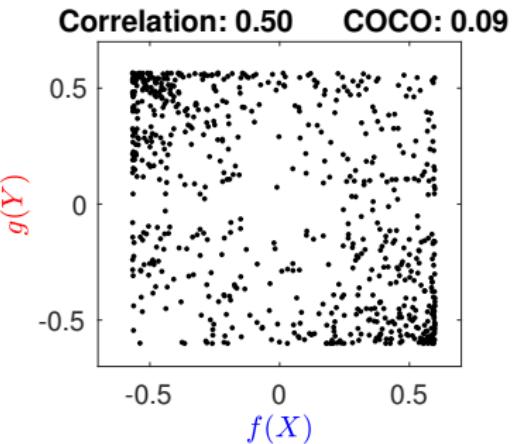
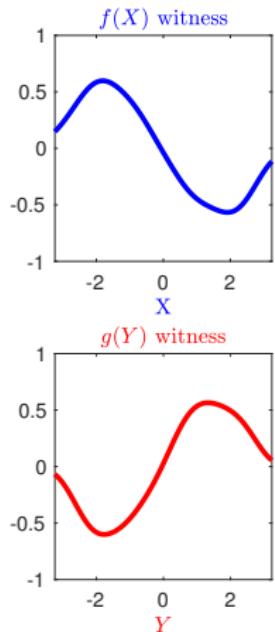
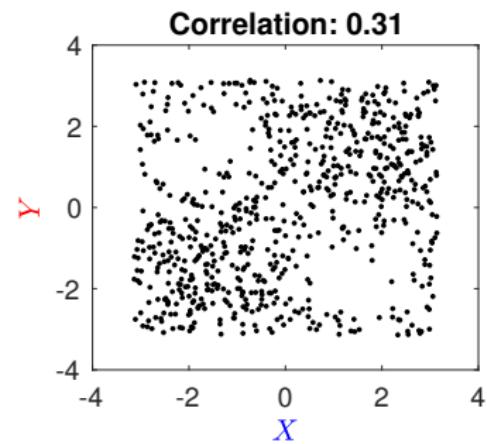
Density takes the form:

$$P_{XY} \propto 1 + \sin(\omega x) \sin(\omega y)$$

Which of these is the more “dependent”?

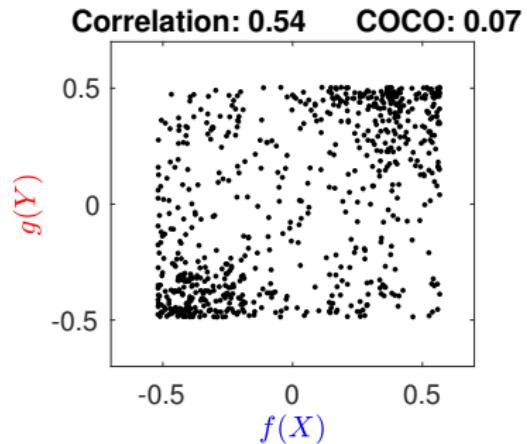
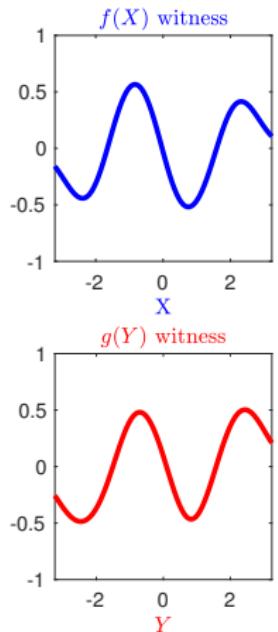
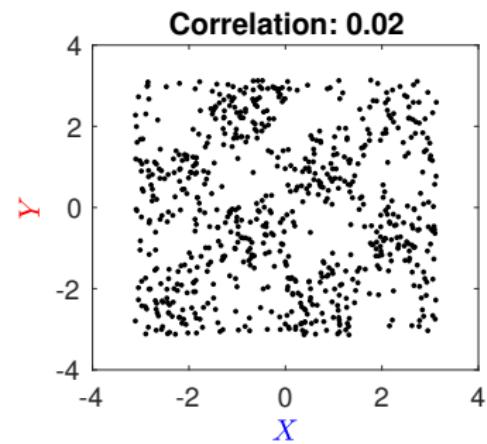
Finding covariance with smooth transformations

Case of $\omega = 1$:



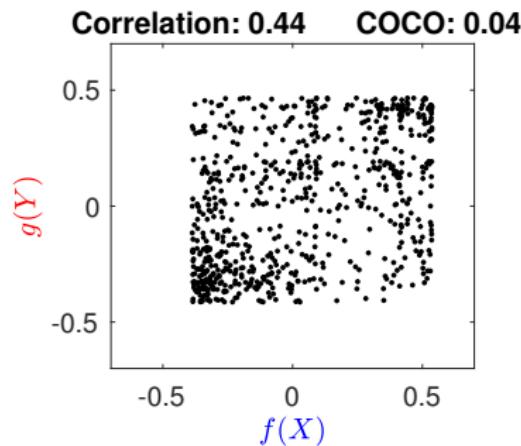
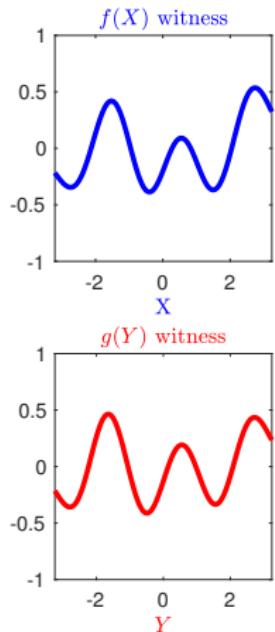
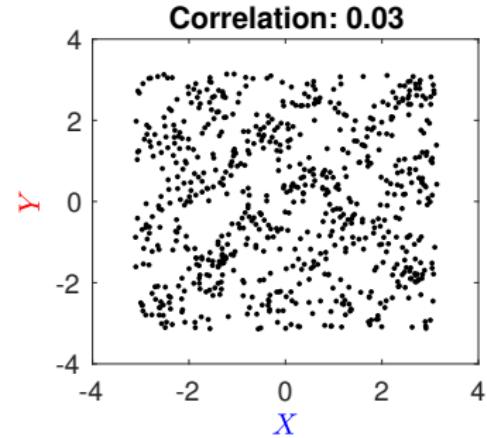
Finding covariance with smooth transformations

Case of $\omega = 2$:



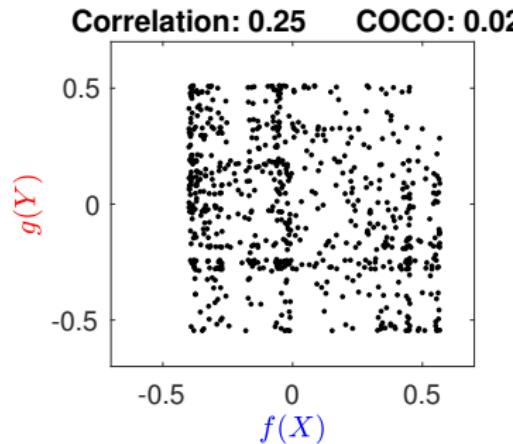
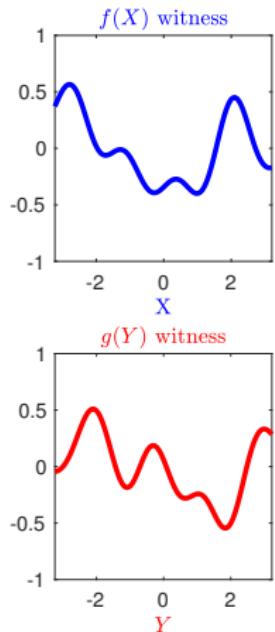
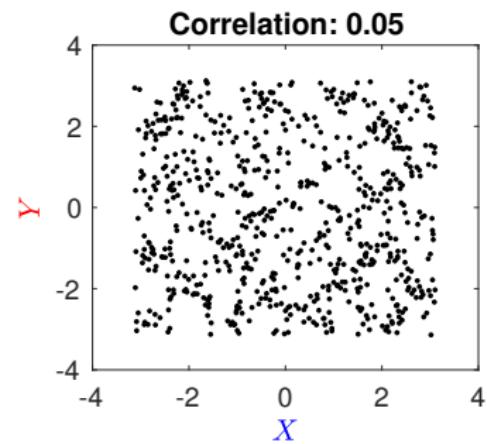
Finding covariance with smooth transformations

Case of $\omega = 3$:



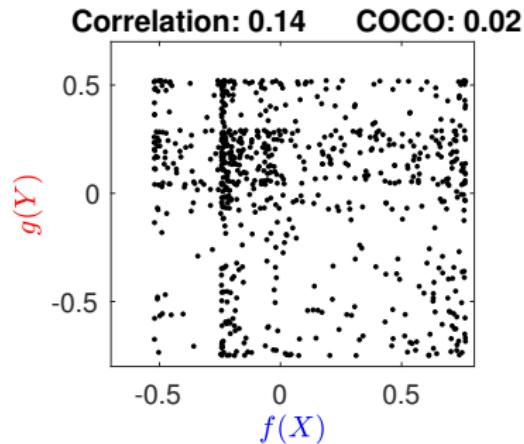
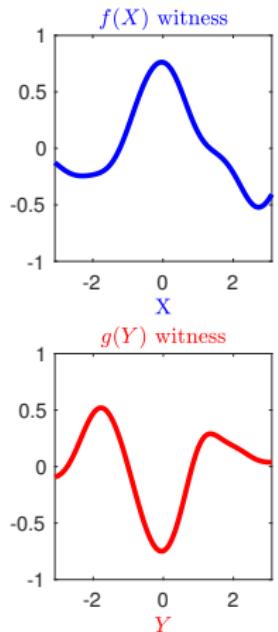
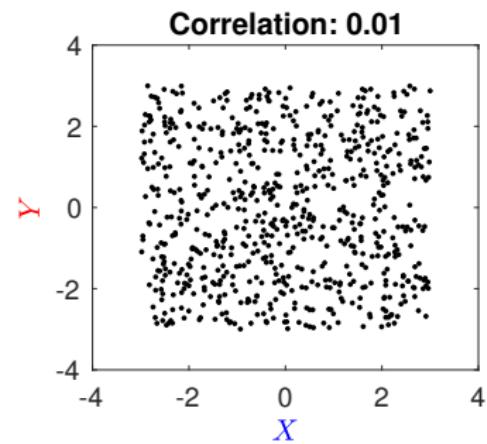
Finding covariance with smooth transformations

Case of $\omega = 4$:



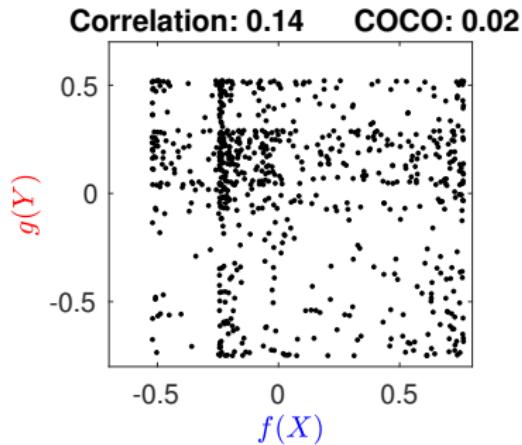
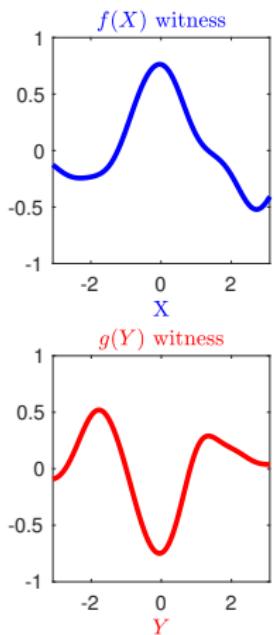
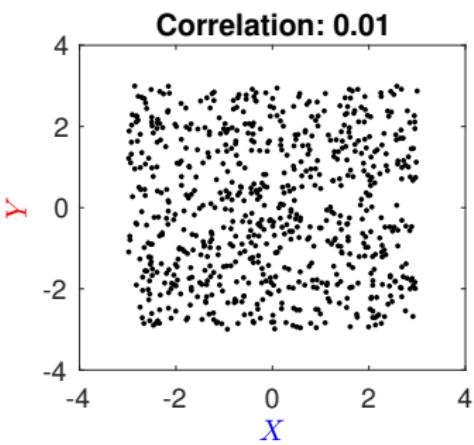
Finding covariance with smooth transformations

Case of $\omega = ??$:



Finding covariance with smooth transformations

Case of $w = 0$: uniform noise! (shows bias)



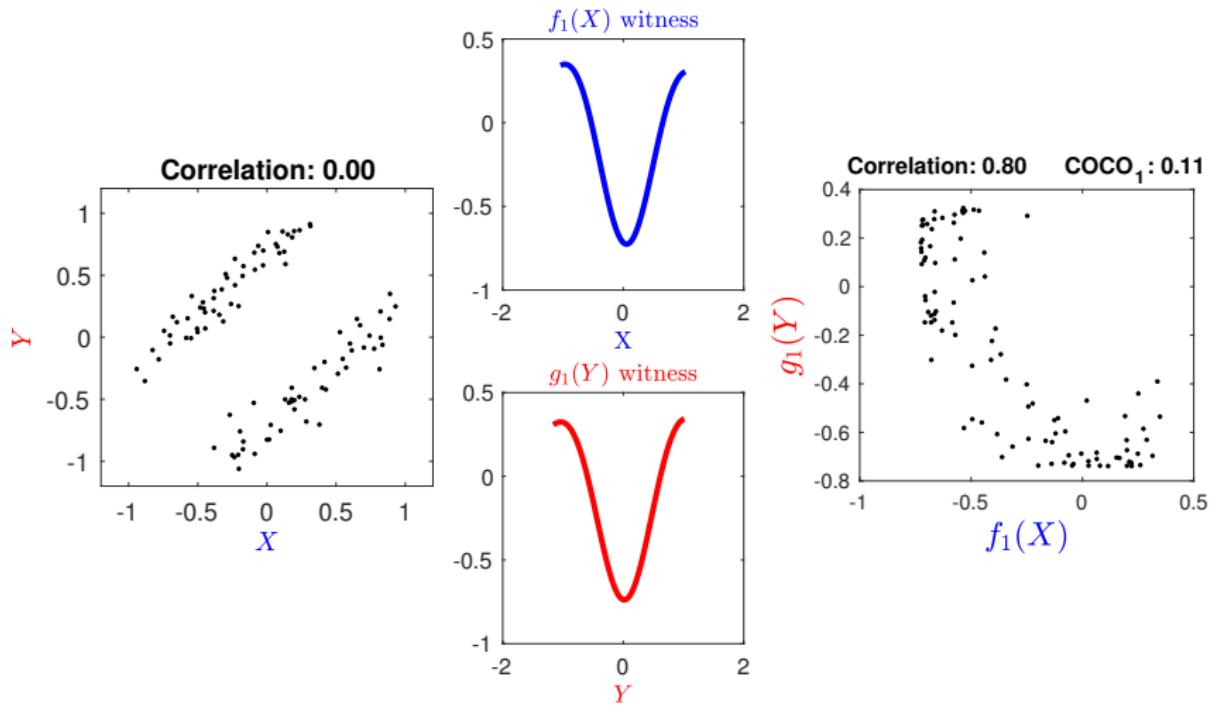
Dependence largest when at “low” frequencies

- As dependence is encoded at **higher frequencies**, the **smooth mappings** f, g achieve lower linear dependence.
- Even for **independent variables**, COCO will not be zero at **finite sample sizes**, since some mild linear dependence will be found by f, g (**bias**)
- This **bias** will decrease with increasing sample size.

Can we do better than COCO?

A second example with zero correlation.

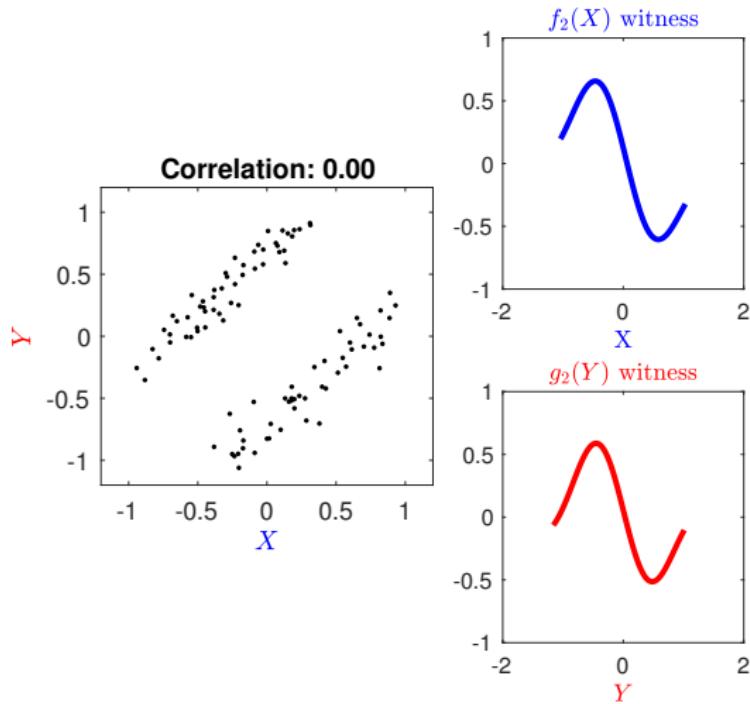
First singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



Can we do better than COCO?

A second example with zero correlation.

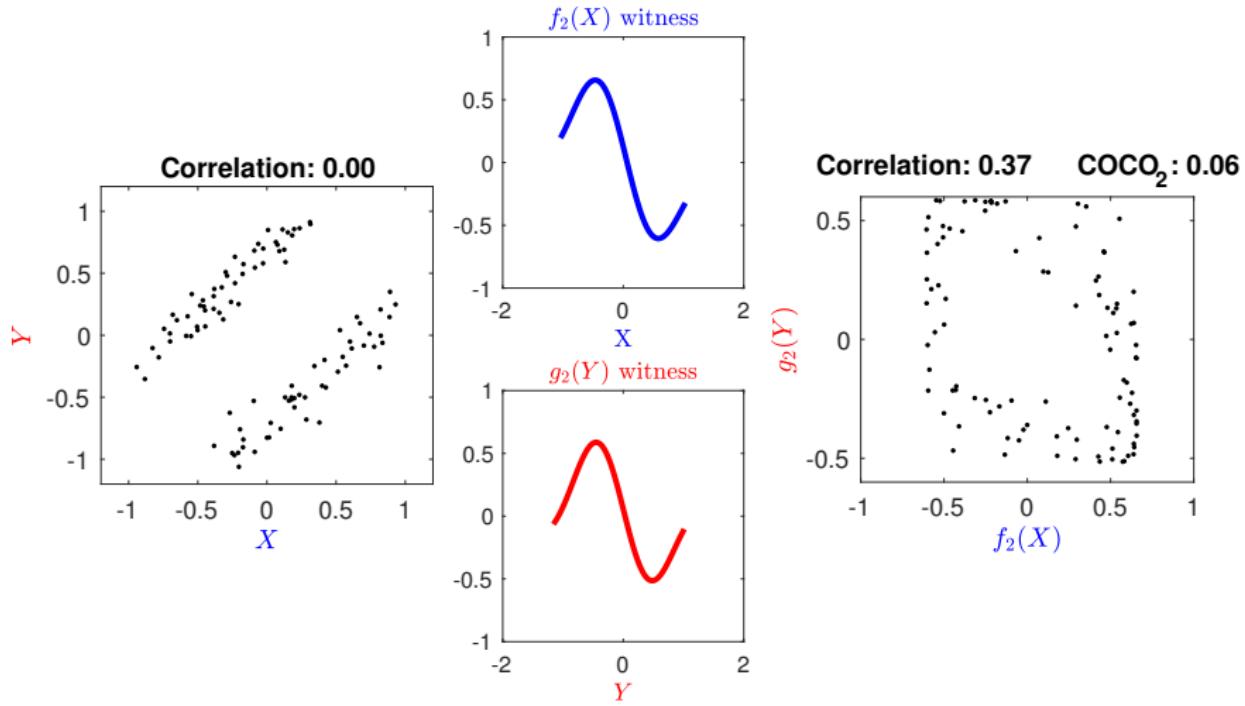
Second singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



Can we do better than COCO?

A second example with zero correlation.

Second singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



The Hilbert-Schmidt Independence Criterion

Writing the i th singular value of the feature covariance $C_{\varphi(x)\phi(y)}$ as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define **Hilbert-Schmidt Independence Criterion (HSIC)**

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \sum_{i=1}^{\infty} \gamma_i^2.$$

G, Bousquet , Smola., and Schoelkopf, ALT05; G.,, Fukumizu, Teo., Song., Schoelkopf., and Smola, NIPS 2007,.

The Hilbert-Schmidt Independence Criterion

Writing the i th singular value of the feature covariance $C_{\varphi(x)\phi(y)}$ as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define **Hilbert-Schmidt Independence Criterion (HSIC)**

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \sum_{i=1}^{\infty} \gamma_i^2.$$

G, Bousquet , Smola., and Schoelkopf, ALT05; G., Fukumizu, Teo., Song., Schoelkopf., and Smola, NIPS 2007,.

HSIC is MMD with product kernel!

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = MMD^2(P_{XY}, P_X P_Y; \mathcal{H}_\kappa)$$

where $\kappa((x, y), (x', y')) = k(x, x')l(y, y')$.

Asymptotics of HSIC under independence

- Given sample $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical \widehat{HSIC} ?
- Empirical HSIC (biased)

$$\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ (K and L computed with empirically centered features)

- Statistical testing: given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α ?
- Asymptotics of \widehat{HSIC} when $P_{XY} = P_X P_Y$:

$$n \widehat{HSIC} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \quad z_l \sim \mathcal{N}(0, 1) \text{i.i.d.}$$

where $\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv}$

Asymptotics of HSIC under independence

- Given sample $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical \widehat{HSIC} ?
- Empirical HSIC (biased)

$$\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ (K and L computed with empirically centered features)

- Statistical testing: given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α ?
- Asymptotics of \widehat{HSIC} when $P_{XY} = P_X P_Y$:

$$n \widehat{HSIC} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \quad z_l \sim \mathcal{N}(0, 1) \text{i.i.d.}$$

where $\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu} l_{tu} + k_{tv} l_{vw} - 2k_{tu} l_{tv}$

Asymptotics of HSIC under independence

- Given sample $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical \widehat{HSIC} ?
- Empirical HSIC (biased)

$$\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ (K and L computed with empirically centered features)

- Statistical testing: given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α ?
- Asymptotics of \widehat{HSIC} when $P_{XY} = P_X P_Y$:

$$n \widehat{HSIC} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \quad z_l \sim \mathcal{N}(0, 1) \text{i.i.d.}$$

where $\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv}$

Asymptotics of HSIC under independence

- Given sample $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical \widehat{HSIC} ?
- Empirical HSIC (biased)

$$\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ (K and L computed with empirically centered features)

- Statistical testing: given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α ?
- Asymptotics of \widehat{HSIC} when $P_{XY} = P_X P_Y$:

$$n \widehat{HSIC} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \quad z_l \sim \mathcal{N}(0, 1) \text{i.i.d.}$$

where $\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}$, $h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv}$

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α (prob. of false positive)?
- Original time series:

$X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10}$
 $Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 Y_7 Y_8 Y_9 Y_{10}$

- Permutation:

$X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10}$
 $Y_7 Y_3 Y_9 Y_2 Y_4 Y_8 Y_5 Y_1 Y_6 Y_{10}$

- Null distribution via permutation

- Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \dots, n\}$. This gives HSIC for independent variables.
- Repeat for many different permutations, get empirical CDF
- Threshold c_α is $1 - \alpha$ quantile of empirical CDF

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α (prob. of false positive)?
- Original time series:

X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10}
 Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 Y_7 Y_8 Y_9 Y_{10}

- Permutation:

X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10}
 Y_7 Y_3 Y_9 Y_2 Y_4 Y_8 Y_5 Y_1 Y_6 Y_{10}

- Null distribution via permutation

- Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \dots, n\}$. This gives HSIC for independent variables.
- Repeat for many different permutations, get empirical CDF
- Threshold c_α is $1 - \alpha$ quantile of empirical CDF

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α (prob. of false positive)?
- Original time series:

X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10}
 Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 Y_7 Y_8 Y_9 Y_{10}

- Permutation:

X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10}
 Y_7 Y_3 Y_9 Y_2 Y_4 Y_8 Y_5 Y_1 Y_6 Y_{10}

- Null distribution via **permutation**

- Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \dots, n\}$. This gives HSIC for independent variables.
- Repeat for many different permutations, get empirical CDF
- Threshold c_α is $1 - \alpha$ quantile of empirical CDF

Application: dependence detection across languages

Testing task: detect dependence between English and French text

X	Y
Honourable senators, I have a question for the Leader of the Government in the Senate	Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat
No doubt there is great pressure on provincial and municipal governments	Les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions
In fact, we have increased federal investments for early childhood development.	Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes
• • •	• • •

Application: dependence detection across languages

Testing task: detect dependence between English and French text

k-spectrum kernel, $k = 10$, sample size $n = 10$

X

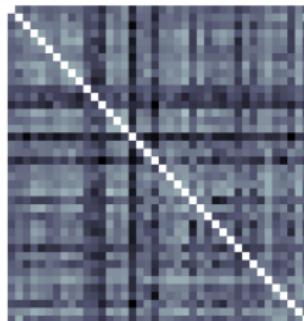
Honourable senators, I have a question for the Leader of the Government in the Senate

No doubt there is great pressure on provincial and municipal governments

In fact, we have increased federal investments for early childhood development.

⋮

K



Y

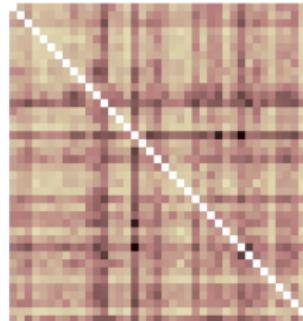
Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat

Les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions

Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes

⋮

L



$$\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

(K and L column centered)

Application: Dependence detection across languages

Results (for $\alpha = 0.05$)

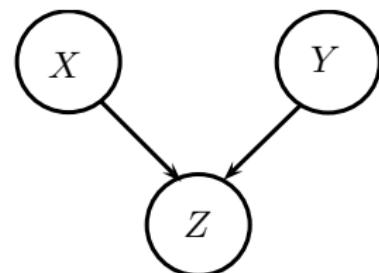
- k-spectrum kernel: average Type II error 0
- Bag of words kernel: average Type II error 0.18

Settings: Five line extracts, averaged over 300 repetitions, for “Agriculture” transcripts. Similar results for Fisheries and Immigration transcripts.

Testing higher order interactions

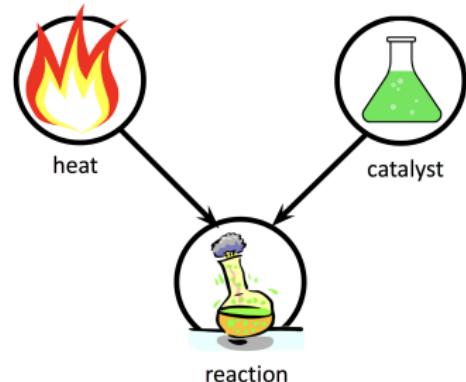
Detecting higher order interaction

How to detect V-structures with pairwise weak individual dependence?



Detecting higher order interaction

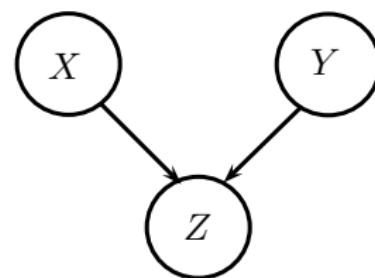
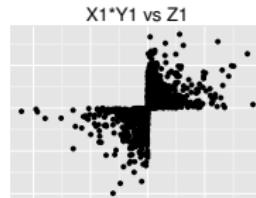
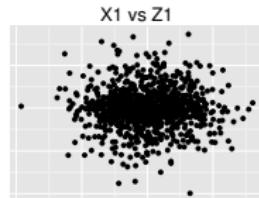
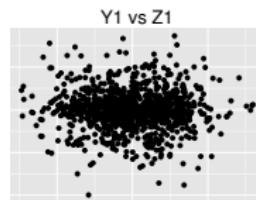
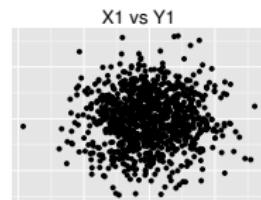
How to detect V-structures with pairwise weak individual dependence?



Detecting higher order interaction

How to detect V-structures with pairwise weak individual dependence?

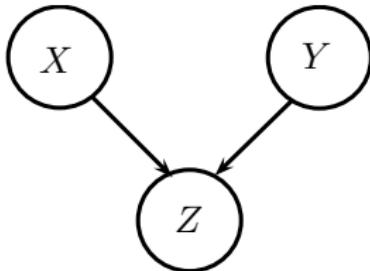
$$X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Z$$



- $X, Y \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$
- $Z | X, Y \sim \text{sign}(XY) \text{Exp}\left(\frac{1}{\sqrt{2}}\right)$

Fine print: Faithfulness violated here!

V-structure discovery



Assume $X \perp\!\!\!\perp Y$ has been established.

V-structure can then be detected by:

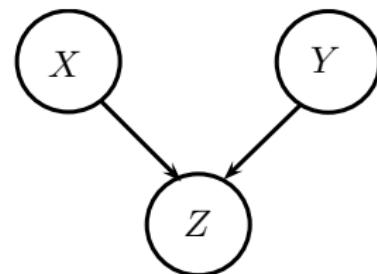
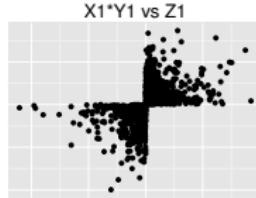
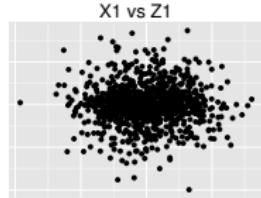
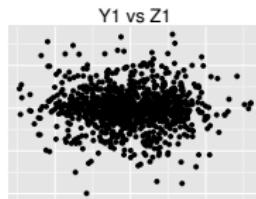
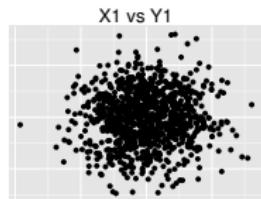
- Consistent CI test: $H_0 : X \perp\!\!\!\perp Y | Z$ [Fukumizu et al. 2008, Zhang et al. 2011]
- Factorisation test: $H_0 : (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$
(multiple standard two-variable tests)

How well do these work?

Detecting higher order interaction

Generalise earlier example to p dimensions

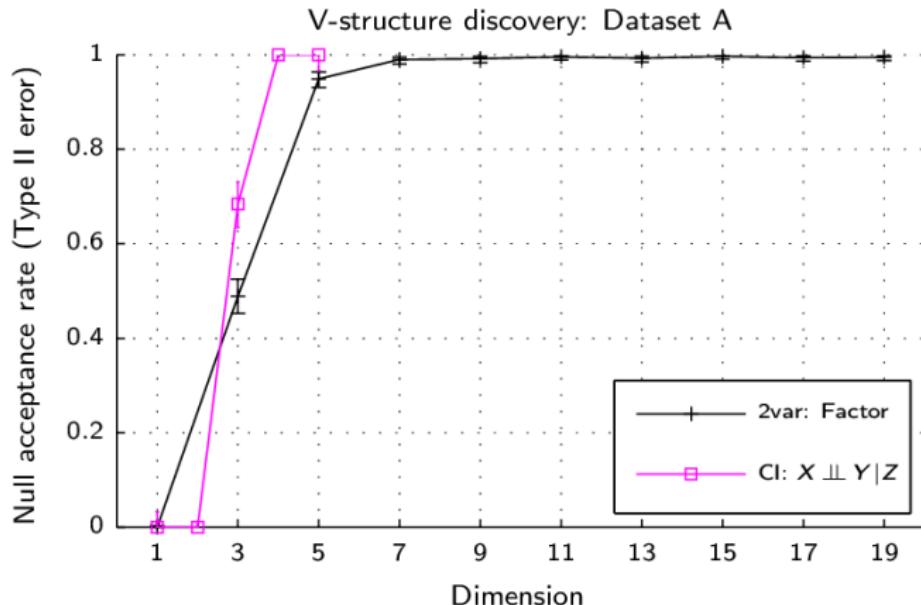
$$X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Z$$



- $X, Y \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$
- $Z | X, Y \sim \text{sign}(XY) \text{Exp}\left(\frac{1}{\sqrt{2}}\right)$
- $X_{2:p}, Y_{2:p}, Z_{2:p} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1})$

Fine print: Faithfulness violated here!

V-structure discovery



CI test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al. (2011), and a factorisation test,
 $n = 500$ 64/71

Lancaster interaction measure

Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that **vanishes** whenever P can be factorised non-trivially.

$$D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y$$

Lancaster interaction measure

Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that **vanishes** whenever P can be factorised non-trivially.

$$D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y$$

$$D = 3 : \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z$$

Lancaster interaction measure

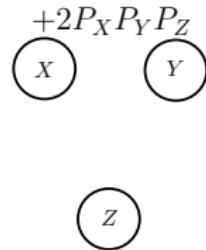
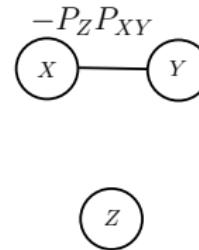
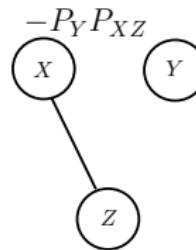
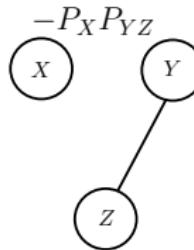
Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that **vanishes** whenever P can be factorised non-trivially.

$$D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y$$

$$D = 3 : \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z$$

$$\Delta_L P =$$

$$P_{XYZ}$$

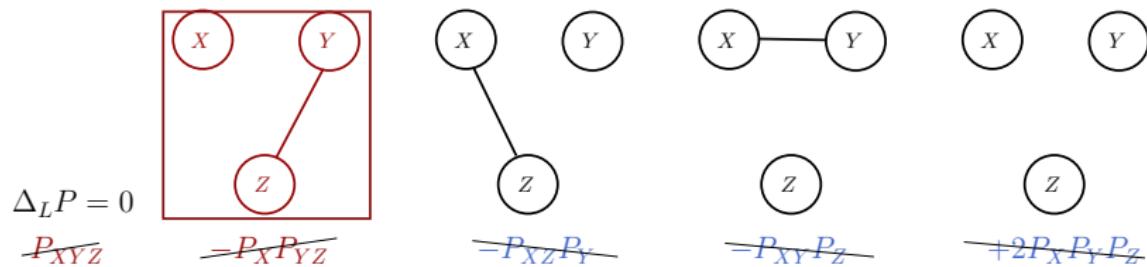


Lancaster interaction measure

Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure Δ_P that **vanishes** whenever P can be factorised non-trivially.

$$D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y$$

$$D = 3 : \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z$$



Case of $P_X \perp\!\!\!\perp P_{YZ}$

Lancaster interaction measure

Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that **vanishes** whenever P can be factorised non-trivially.

$$D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y$$

$$D = 3 : \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z$$

$$(X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X \Rightarrow \Delta_L P = 0.$$

...so what might be missed?

Lancaster interaction measure

Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that **vanishes** whenever P can be factorised non-trivially.

$$D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y$$

$$D = 3 : \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z$$

$$\Delta_L P = 0 \nRightarrow (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$$

Example:

$P(0, 0, 0) = 0.2$	$P(0, 0, 1) = 0.1$	$P(1, 0, 0) = 0.1$	$P(1, 0, 1) = 0.1$
$P(0, 1, 0) = 0.1$	$P(0, 1, 1) = 0.1$	$P(1, 1, 0) = 0.1$	$P(1, 1, 1) = 0.2$

A kernel test statistic using Lancaster Measure

Construct a test by estimating $\|\mu_\kappa(\Delta_L P)\|_{\mathcal{H}_\kappa}^2$, where $\kappa = \textcolor{red}{k} \otimes \textcolor{blue}{l} \otimes \textcolor{magenta}{m}$:

$$\begin{aligned}\|\mu_\kappa(P_{XYZ} - P_{XY}P_Z - \dots)\|_{\mathcal{H}_\kappa}^2 &= \\ \langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XYZ} \rangle_{\mathcal{H}_\kappa} - 2 \langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XY}P_Z \rangle_{\mathcal{H}_\kappa} - \dots\end{aligned}$$

A kernel test statistic using Lancaster Measure

$\nu \setminus \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_XP_YP_Z$
P_{XYZ}	$(K \circ L \circ M)_{++}$	$((K \circ L)M)_{++}$	$((K \circ M)L)_{++}$	$((M \circ L)K)_{++}$	$tr(K_+ \circ L_+ \circ M_+)$
$P_{XY}P_Z$		$(K \circ L)_{++} M_{++}$	$(MKL)_{++}$	$(KLM)_{++}$	$(KL)_{++} M_{++}$
$P_{XZ}P_Y$			$(K \circ M)_{++} L_{++}$	$(KML)_{++}$	$(KM)_{++} L_{++}$
$P_{YZ}P_X$				$(L \circ M)_{++} K_{++}$	$(LM)_{++} K_{++}$
$P_XP_YP_Z$					$K_{++} L_{++} M_{++}$

Table: V -statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$ (without terms $P_X P_Y P_Z$). H is centering matrix $I - n^{-1}$

Lancaster interaction statistic: Sejdinovic, G, Bergsma, NIPS13

$$\|\mu_\kappa(\Delta_L P)\|_{\mathcal{H}_\kappa}^2 = \frac{1}{n^2} \boxed{(HKH \circ H\textcolor{blue}{L}H \circ H\textcolor{red}{M}H)_{++}}.$$

A kernel test statistic using Lancaster Measure

$\nu \setminus \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_XP_YP_Z$
P_{XYZ}	$(K \circ L \circ M)_{++}$	$((K \circ L)M)_{++}$	$((K \circ M)L)_{++}$	$((M \circ L)K)_{++}$	$tr(K_+ \circ L_+ \circ M_+)$
$P_{XY}P_Z$		$(K \circ L)_{++} M_{++}$	$(MKL)_{++}$	$(KLM)_{++}$	$(KL)_{++} M_{++}$
$P_{XZ}P_Y$			$(K \circ M)_{++} L_{++}$	$(KML)_{++}$	$(KM)_{++} L_{++}$
$P_{YZ}P_X$				$(L \circ M)_{++} K_{++}$	$(LM)_{++} K_{++}$
$P_XP_YP_Z$					$K_{++} L_{++} M_{++}$

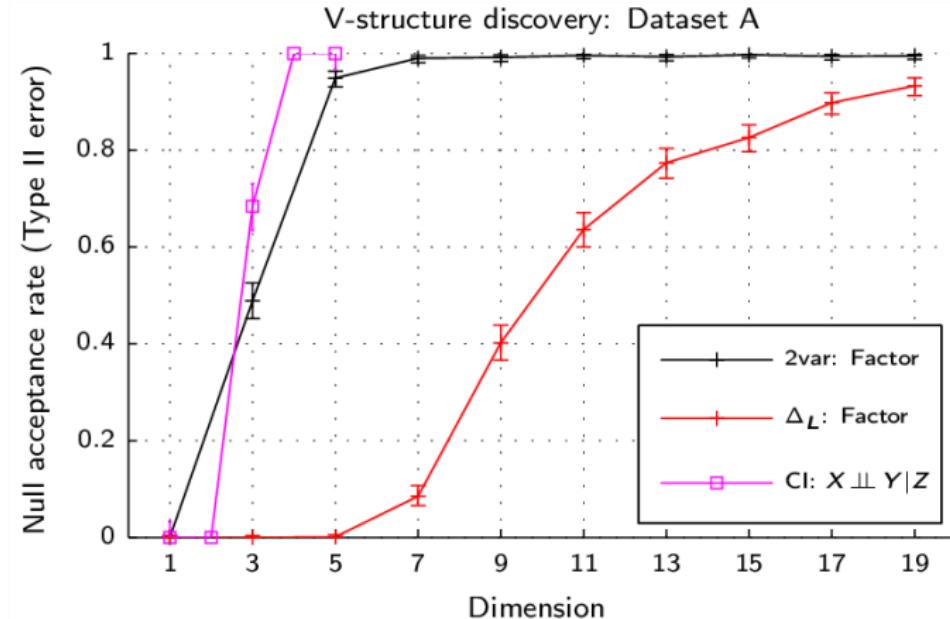
Table: V -statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$ (without terms $P_X P_Y P_Z$). H is centering matrix $I - n^{-1}$

Lancaster interaction statistic: Sejdinovic, G, Bergsma, NIPS13

$$\|\mu_\kappa(\Delta_L P)\|_{\mathcal{H}_\kappa}^2 = \frac{1}{n^2} \boxed{(HKH \circ H\textcolor{blue}{L}H \circ H\textcolor{magenta}{M}H)_{++}}.$$

Empirical joint central moment in the feature space

V-structure discovery



Lancaster test, CI test for $X \perp\!\!\! \perp Y|Z$ from Zhang et al. (2011), and a factorisation test, $n = 500$

Interaction for $D \geq 4$

- Interaction measure valid for all D :

(Streitberg, 1990)

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_{\pi} P$$

- For a partition π , J_{π} associates to the joint the corresponding factorisation,
e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

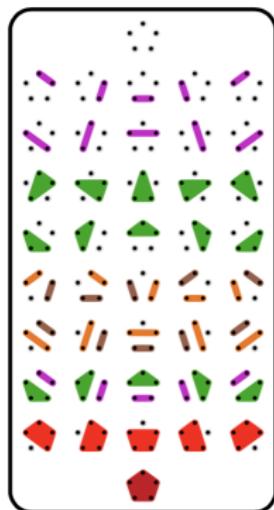
Interaction for $D \geq 4$

- Interaction measure valid for all D :

(Streitberg, 1990)

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_{\pi} P$$

- For a partition π , J_{π} associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.



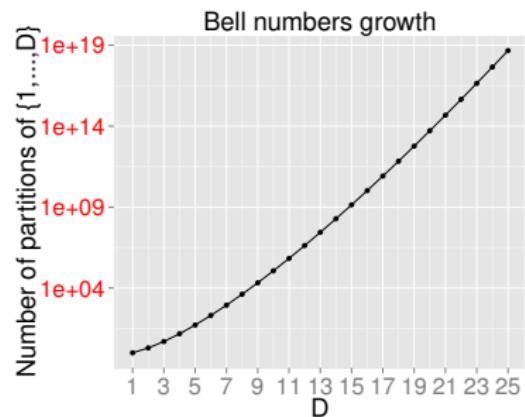
Interaction for $D \geq 4$

■ Interaction measure valid for all D :

(Streitberg, 1990)

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_{\pi} P$$

- For a partition π , J_{π} associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.



Co-authors

From Gatsby:

- Mikolaj Binkowski
- Kacper Chwialkowski
- Wittawat Jitkrittum
- Heiko Strathmann
- Dougal Sutherland
- Wenkai Xu

External collaborators:

- Kenji Fukumizu
- Bernhard Schoelkopf
- Dino Sejdinovic
- Bharath Sriperumbudur
- Alex Smola
- Zoltan Szabo

Questions?

