Introduction to Kernel Methods for Comparing Distributions

Wittawat Jitkrittum

Max Planck Institute for Intelligent Systems, Germany
wittawatj@gmail.com

Bangkok Machine Learning Meetup
7 March 2018

Comparing Distributions

Have: Two collections of samples X, Y from unknown distributions p and q.

Two goals: using only X, Y ...

1. Measure the distance between p and q.
2. Are p and q different (not just by chance)?

$\Longrightarrow$ **Two-sample testing**

Comparing Distributions

Have: Two collections of samples X, Y from unknown distributions p and q.

Two goals: using only X, Y ...

1. Measure the distance between p and q.
2. Are p and q different (not just by chance)?

$\Longrightarrow$ **Two-sample testing**

Comparing Distributions

Have: Two collections of samples X, Y from unknown distributions p and q.

Two goals: using only X, Y ...

1. Measure the distance between p and q.
2. Are p and q different (not just by chance)?

$\Longrightarrow$ Two-sample testing

$\sim P$

$\sim Q$

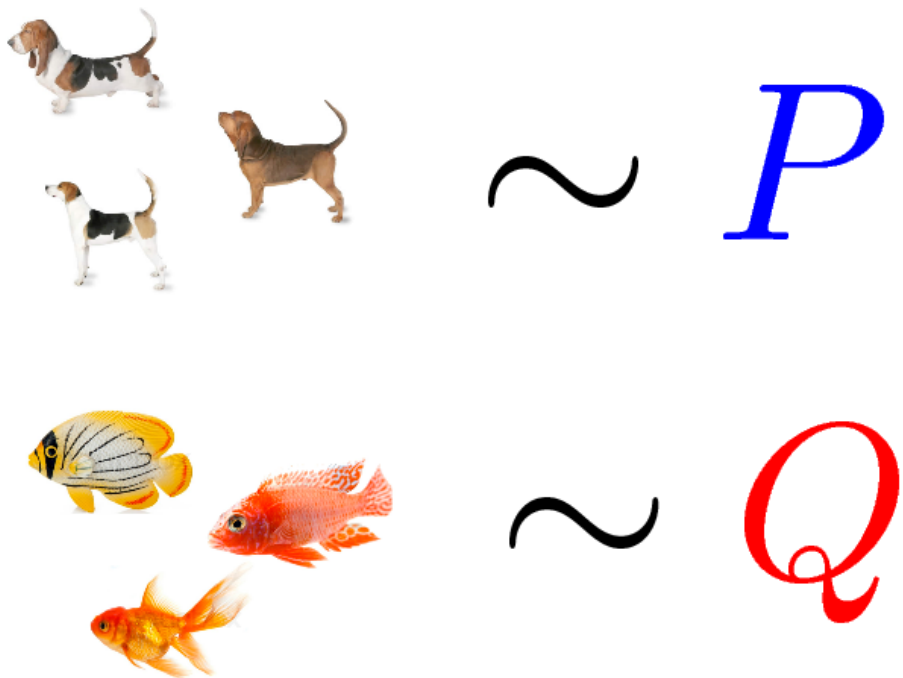
Comparing Distributions

Have: Two collections of samples X, Y from unknown distributions p and q.

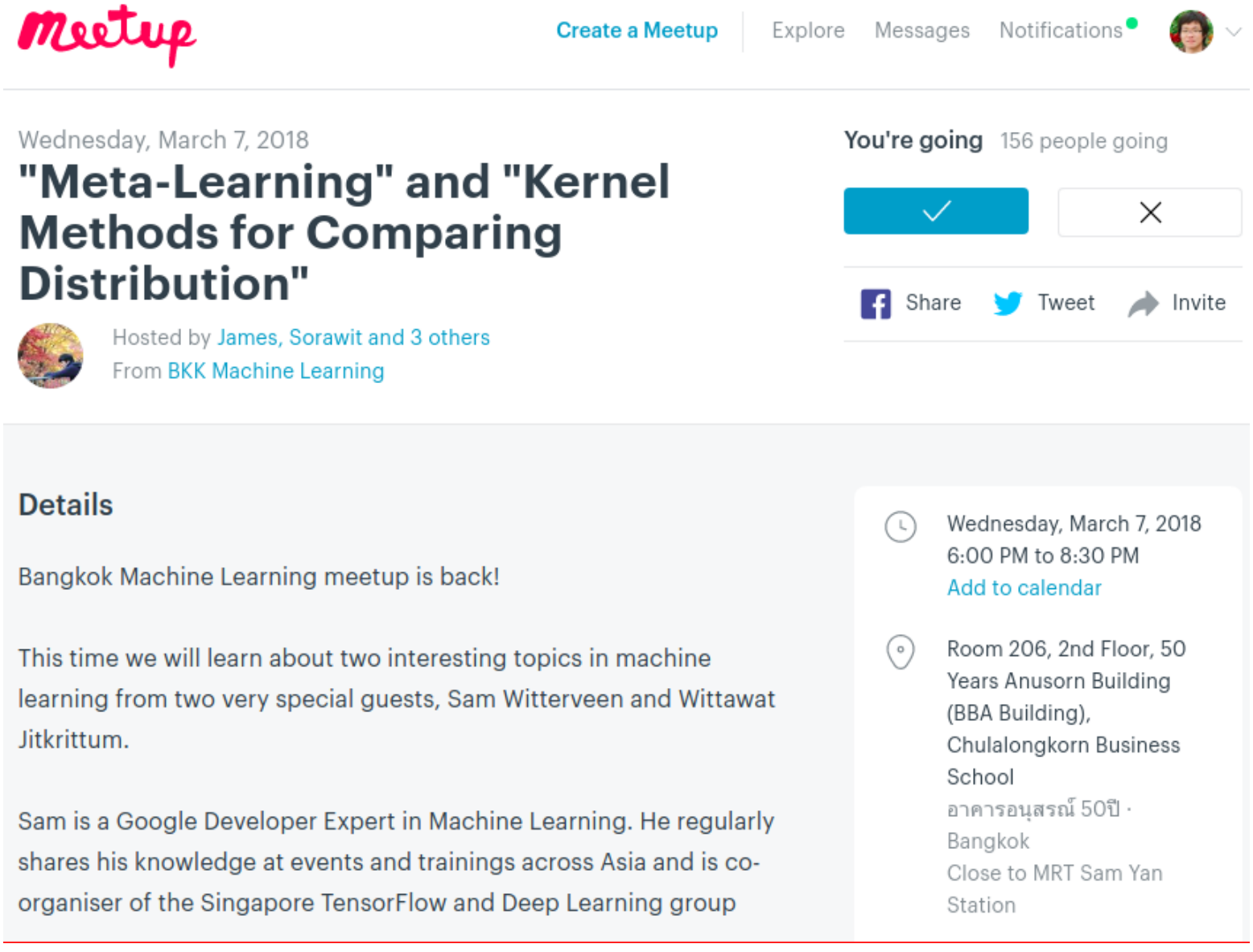
Two goals: using only X, Y ...

1. Measure the distance between p and q.
2. Are p and q different (not just by chance)?

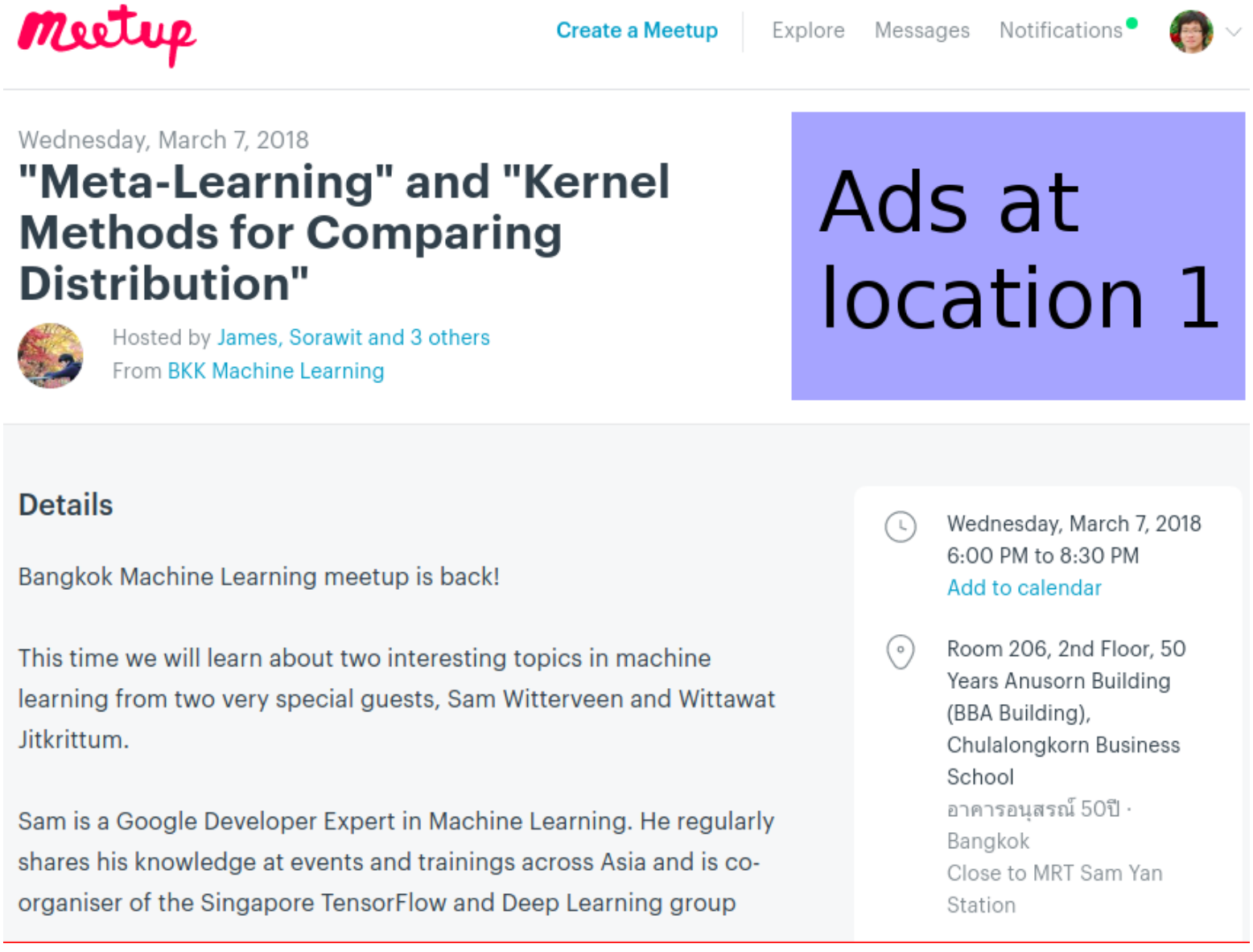
$\Longrightarrow$ **Two-sample testing**



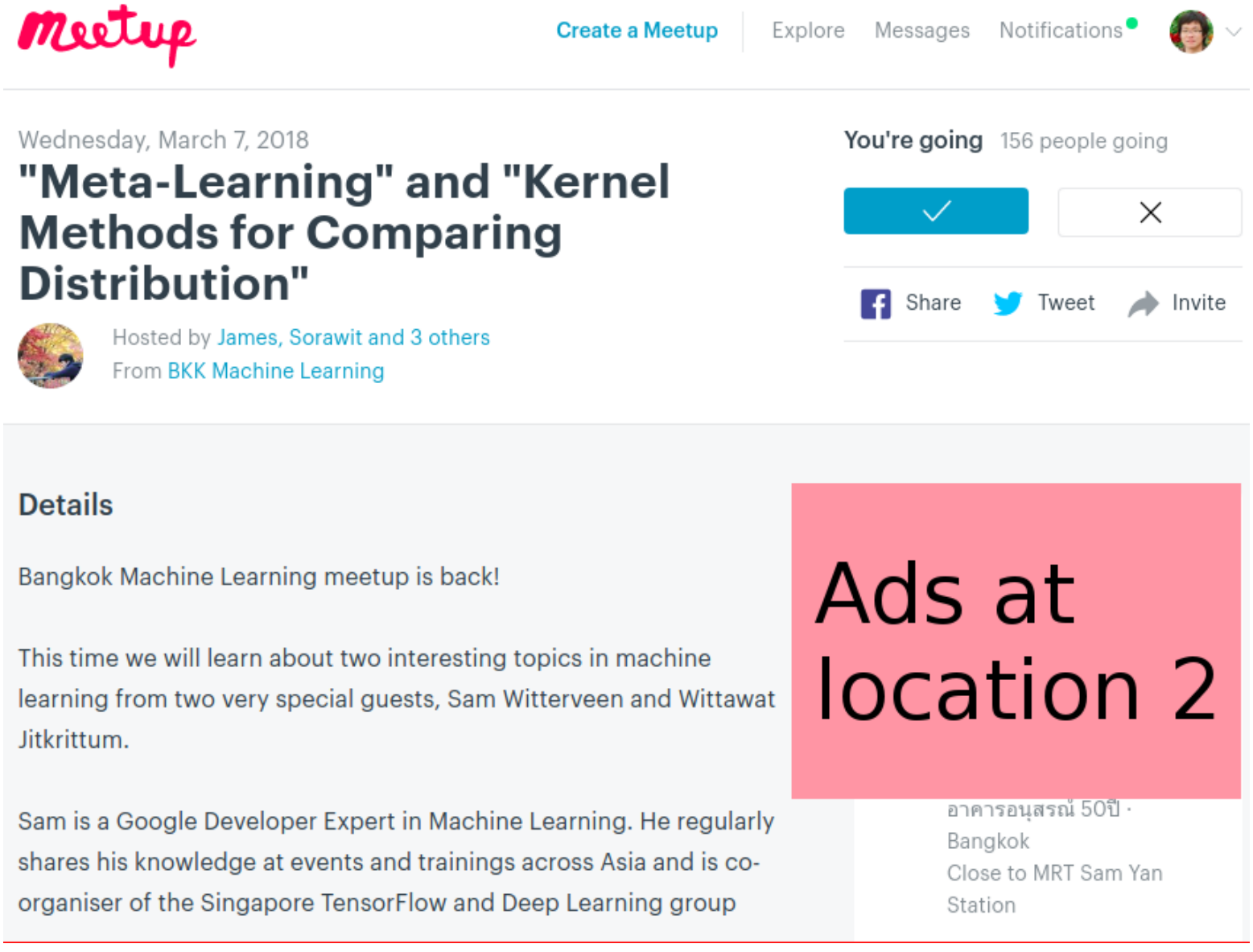
Application 1: Effects of Ads Placement



Application 1: Effects of Ads Placement



Application 1: Effects of Ads Placement



Application 1: Effects of Ads Placement

Does ads at location 1 have the same effect as ads at location 2?

Application 1: Effects of Ads Placement

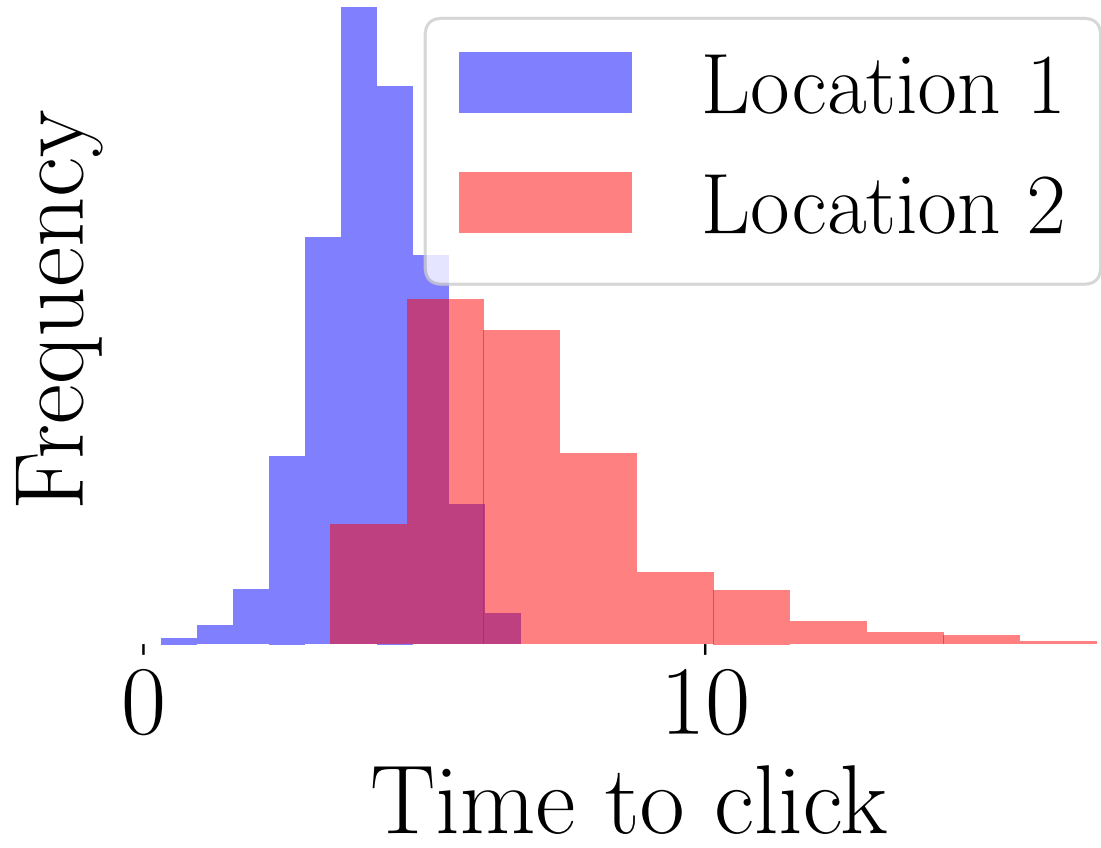
Does ads at location 1 have the same effect as ads at location 2?

- $X =$ time it takes users to click ads at location 1.
- $Y =$ time it takes users to click ads at location 2.

Application 1: Effects of Ads Placement

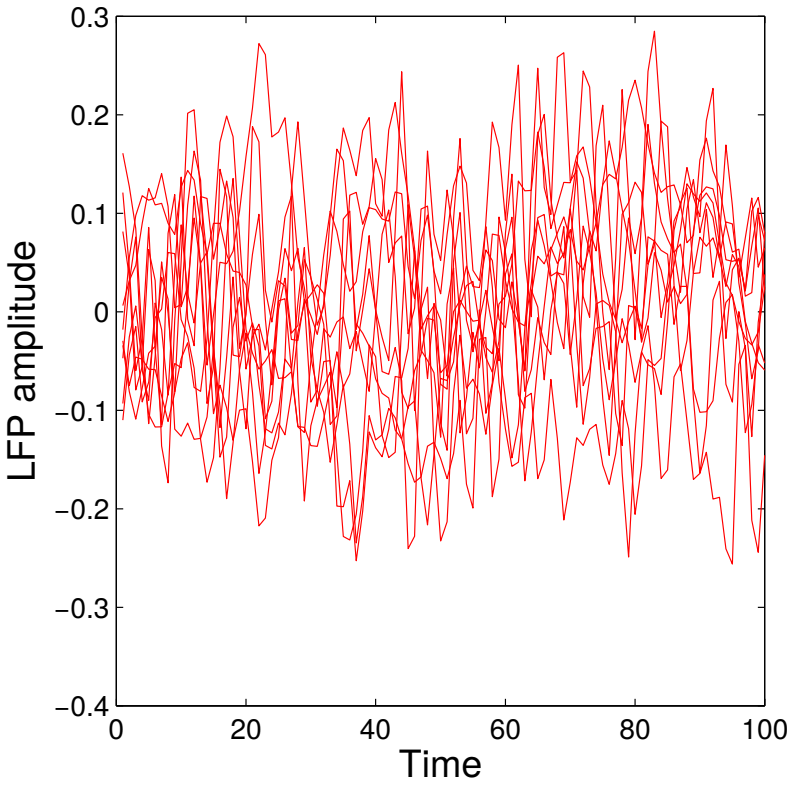
Does ads at location 1 have the same effect as ads at location 2?

- X = time it takes users to click ads at location 1.
- Y = time it takes users to click ads at location 2.

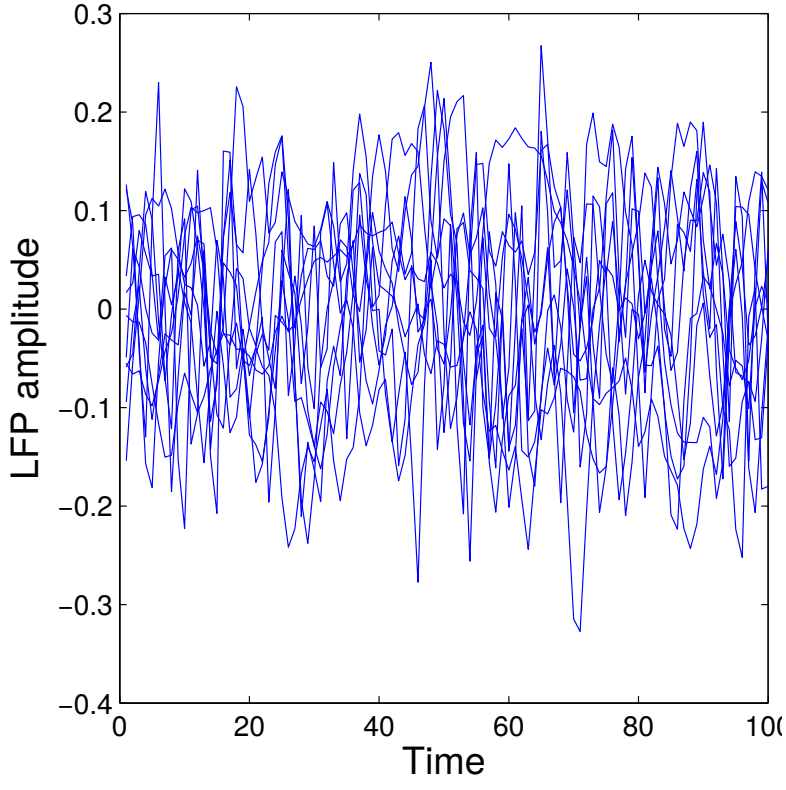


Application 2: Data Integration

- Data from two labs collected under the (supposedly) same setting.
- Should we merge the two databases into one? [Gretton et al., 2012a]



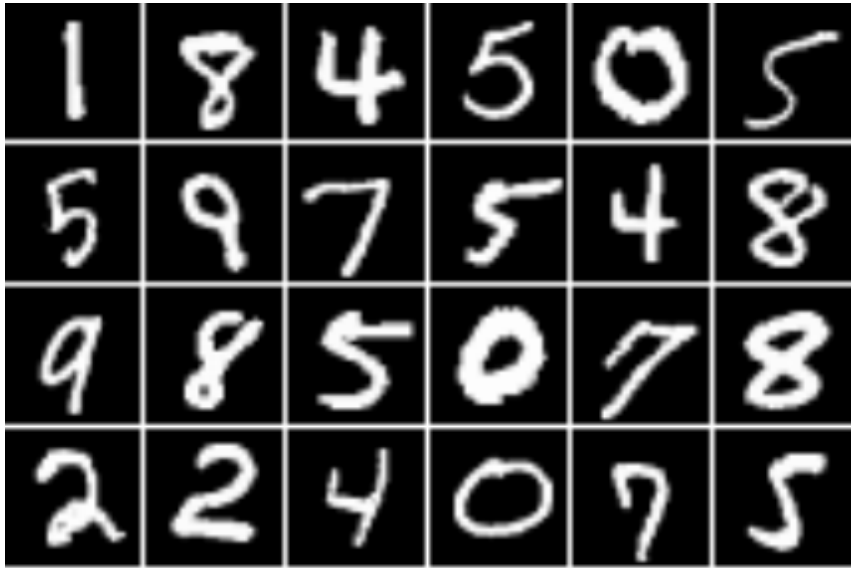
Data collected from lab 1: X.



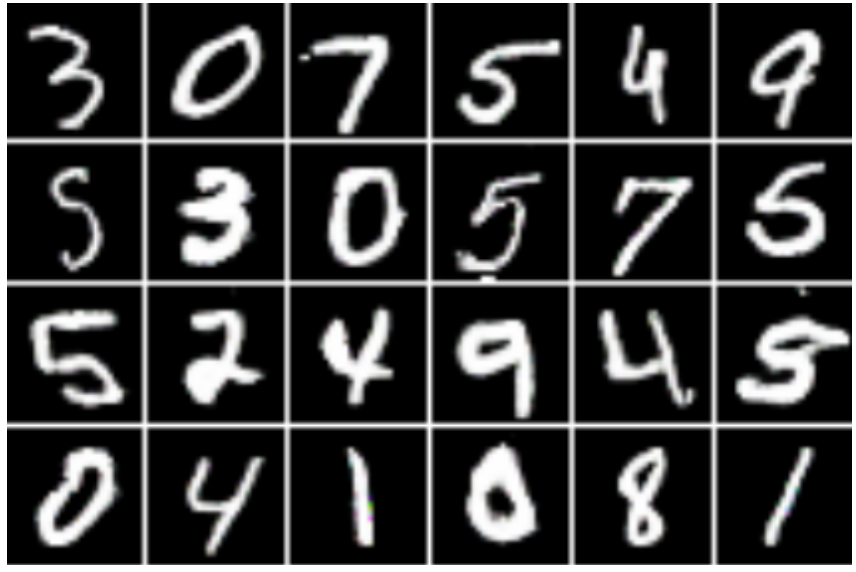
Data collected from lab 2: Y.

- If they have different distributions, do not merge.

Application 3: Benchmarking Generative Models



Observed MNIST handwritten digits. X.



Generated images from a model. Y.

Is Y similar to X?

- Distance between distributions can be used to train generative models.

Outline

Outline

Basic Linear Algebra

- Let $a = (a_1, \ldots, a_d)^\top$, b and c be vectors in $\mathbb{R}^d$.
- **Norm** (length): $\|a\| := \sqrt{\sum_{i=1}^d a_i^2}$.

Inner product (dot product)

$$a \cdot b = a^\top b = \langle a, b \rangle = \sum_{i=1}^d a_i b_i$$

$$= a_1 b_1 + a_2 b_2 + \cdots + a_d b_d.$$

- $\langle a, b \rangle =$ similarity between a and b.
- An inner product induces a norm: $\|a\| = \sqrt{\langle a, a \rangle}$.

Basic Linear Algebra

- Let $\boldsymbol{a} = (a_1, \ldots, a_d)^\top$, $\boldsymbol{b}$ and $\boldsymbol{c}$ be vectors in $\mathbb{R}^d$.
- **Norm** (length): $\|\boldsymbol{a}\| := \sqrt{\sum_{i=1}^d a_i^2}$.

Inner product (dot product)

$$a \cdot b = a^\top b = \langle a, b \rangle = \sum_{i=1}^{d} a_i b_i$$

$$= a_1 b_1 + a_2 b_2 + \cdots + a_d b_d.$$

- $\langle a, b \rangle =$ similarity between a and b.
- An inner product induces a norm: $\|a\| = \sqrt{\langle a, a \rangle}$.

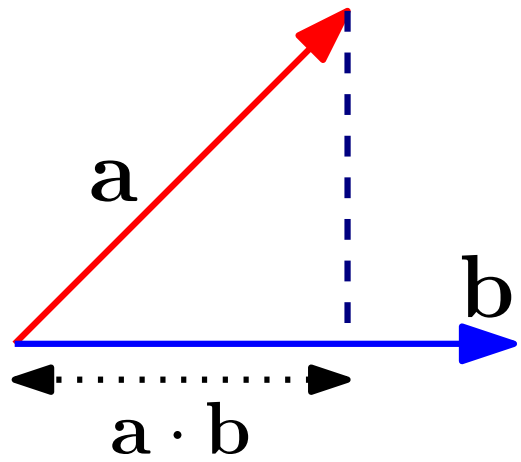
Basic Linear Algebra

- Let $a = (a_1, \ldots, a_d)^\top$, b and c be vectors in $\mathbb{R}^d$.
- **Norm** (length): $\|a\| := \sqrt{\sum_{i=1}^d a_i^2}$.

Inner product (dot product)

$$a \cdot b = a^\top b = \langle a, b \rangle = \sum_{i=1}^d a_i b_i$$

$$= a_1 b_1 + a_2 b_2 + \cdots + a_d b_d.$$



- $\langle a, b \rangle =$ similarity between a and b.
- An inner product induces a norm: $\|a\| = \sqrt{\langle a, a \rangle}$.

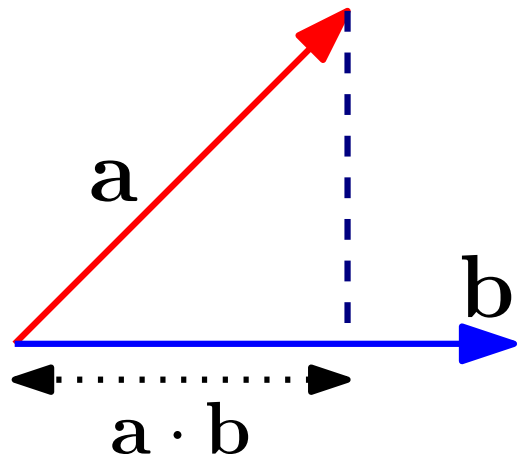
Basic Linear Algebra

- Let $a = (a_1, \ldots, a_d)^\top$, b and c be vectors in $\mathbb{R}^d$.
- **Norm** (length): $\|a\| := \sqrt{\sum_{i=1}^d a_i^2}$.

Inner product (dot product)

$$a \cdot b = a^\top b = \langle a, b \rangle = \sum_{i=1}^d a_i b_i$$

$$= a_1 b_1 + a_2 b_2 + \cdots + a_d b_d.$$



- $\langle a, b \rangle =$ similarity between a and b.
- An inner product induces a norm: $\|a\| = \sqrt{\langle a, a \rangle}$.

Properties of Inner Product

Three properties:

1. (Linear): $\langle \alpha a + \beta b, c \rangle = \alpha \langle a, c \rangle + \beta \langle b, c \rangle$
2. (Symmetric): $\langle a, b \rangle = \langle b, a \rangle$
3. $\langle a, a \rangle \geq 0$ and $\langle a, a \rangle = 0$ if and only if $a = \mathbf{0}$.

- For $x, y \in \mathbb{R}$, we have $(x - y)^2 = x^2 - 2xy + y^2$.
- In general: $\|a - b\|^2 = \langle a, a \rangle - 2 \langle a, b \rangle + \langle b, b \rangle$.
  - $\|a - b\|$ = distance between a and b.

Definition 1 (Hilbert space).

A **Hilbert space** $\mathcal{H}$ is a complete inner product space.

- Hilbert space $\approx$ a space with an inner product defined. $\mathbb{R}^d$ is a Hilbert space.
- In general, can be a space of generic objects.

Properties of Inner Product

Three properties:

1. (Linear): $\langle \alpha a + \beta b, c \rangle = \alpha \langle a, c \rangle + \beta \langle b, c \rangle$
2. (Symmetric): $\langle a, b \rangle = \langle b, a \rangle$
3. $\langle a, a \rangle \geq 0$ and $\langle a, a \rangle = 0$ if and only if $a = \mathbf{0}$.

- For $x, y \in \mathbb{R}$, we have $(x - y)^2 = x^2 - 2xy + y^2$.
- In general: $\|a - b\|^2 = \langle a, a \rangle - 2 \langle a, b \rangle + \langle b, b \rangle$.
  - $\|a - b\|$ = distance between a and b.

Definition 1 (Hilbert space).

A **Hilbert space** $\mathcal{H}$ is a complete inner product space.

- Hilbert space $\approx$ a space with an inner product defined. $\mathbb{R}^d$ is a Hilbert space.
- In general, can be a space of generic objects.

Properties of Inner Product

Three properties:

1. (Linear): $\langle \alpha a + \beta b, c \rangle = \alpha \langle a, c \rangle + \beta \langle b, c \rangle$
2. (Symmetric): $\langle a, b \rangle = \langle b, a \rangle$
3. $\langle a, a \rangle \geq 0$ and $\langle a, a \rangle = 0$ if and only if $a = \mathbf{0}$.

- For $x, y \in \mathbb{R}$, we have $(x - y)^2 = x^2 - 2xy + y^2$.
- In general: $\|a - b\|^2 = \langle a, a \rangle - 2 \langle a, b \rangle + \langle b, b \rangle$.
  - $\|a - b\|$ = distance between a and b.

Definition 1 (Hilbert space).

A **Hilbert space** $\mathcal{H}$ is a complete inner product space.

- Hilbert space $\approx$ a space with an inner product defined. $\mathbb{R}^d$ is a Hilbert space.
- In general, can be a space of generic objects.

Properties of Inner Product

Three properties:

1. (Linear): $\langle \alpha a + \beta b, c \rangle = \alpha \langle a, c \rangle + \beta \langle b, c \rangle$
2. (Symmetric): $\langle a, b \rangle = \langle b, a \rangle$
3. $\langle a, a \rangle \geq 0$ and $\langle a, a \rangle = 0$ if and only if $a = \mathbf{0}$.

- For $x, y \in \mathbb{R}$, we have $(x - y)^2 = x^2 - 2xy + y^2$.
- In general: $\|a - b\|^2 = \langle a, a \rangle - 2 \langle a, b \rangle + \langle b, b \rangle$.
  - $\|a - b\|$ = distance between a and b.

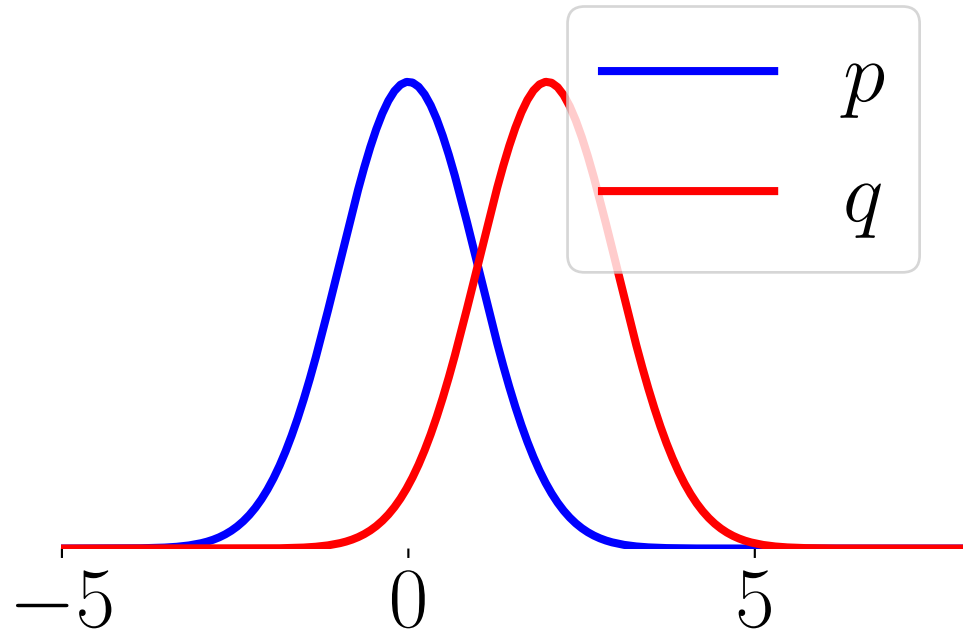
Definition 1 (Hilbert space).

A **Hilbert space** $\mathcal{H}$ is a complete inner product space.

- Hilbert space $\approx$ a space with an inner product defined. $\mathbb{R}^d$ is a Hilbert space.
- In general, can be a space of generic objects.

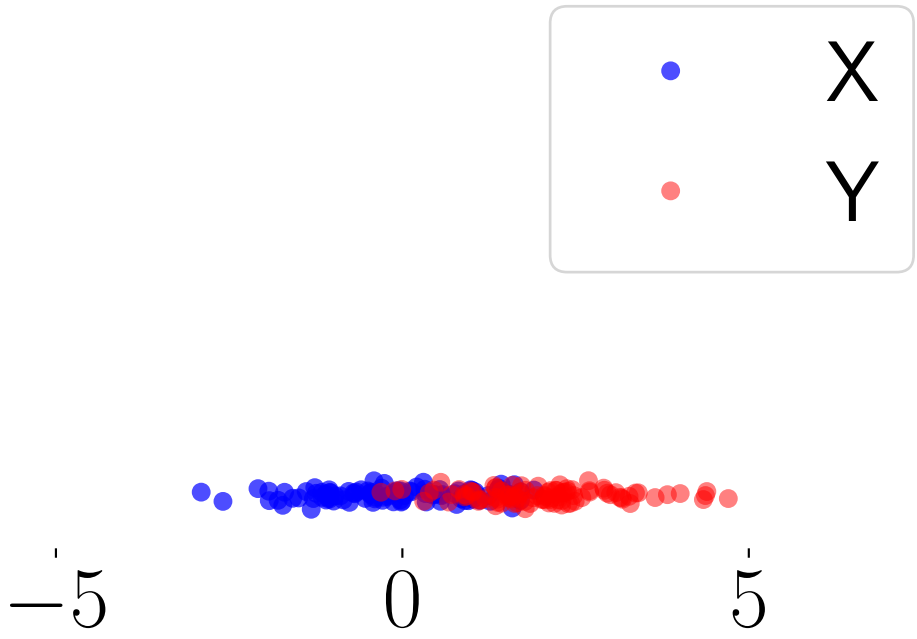
Outline

Case 1: Simple Mean Shift in 1D



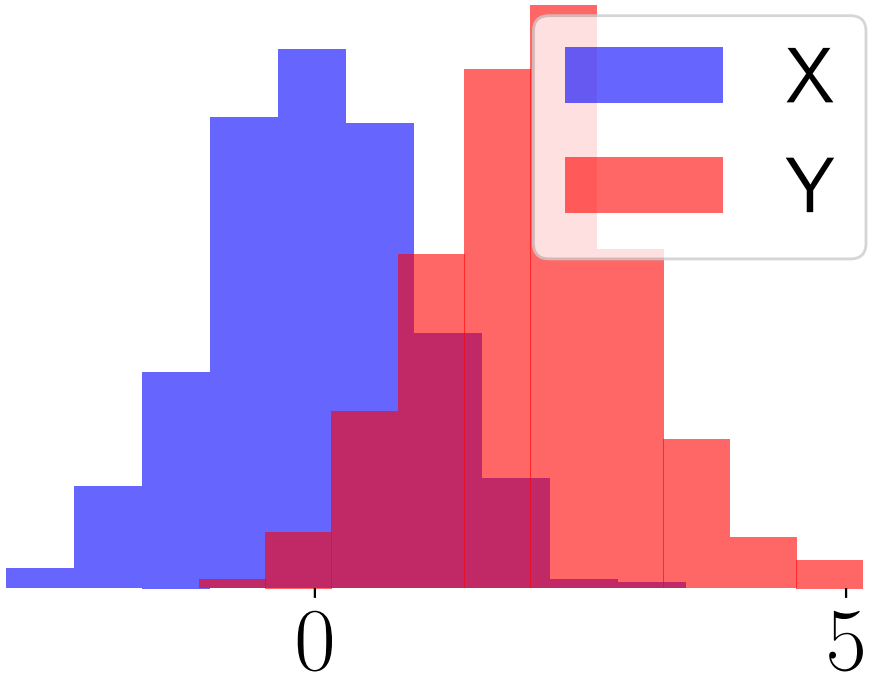
- Two Gaussian distributions.

Case 1: Simple Mean Shift in 1D



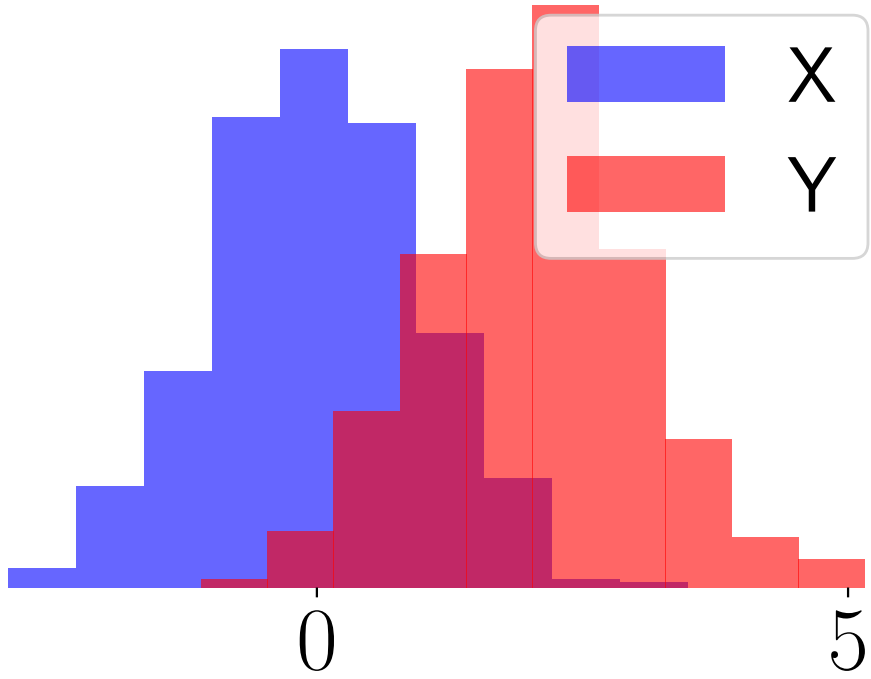
- We have only samples $\mathsf{X} \sim p$ and $\mathsf{Y} \sim q$.
- $\mathsf{X} = \{x_1, \ldots, x_n\}$ and $\mathsf{Y} = \{y_1, \ldots, y_n\}$. Sets of numbers.

Case 1: Simple Mean Shift in 1D



- We have only samples $\mathsf{X} \sim p$ and $\mathsf{Y} \sim q$.
- $\mathsf{X} = \{x_1, \ldots, x_n\}$ and $\mathsf{Y} = \{y_1, \ldots, y_n\}$. Sets of numbers.

Case 1: Simple Mean Shift in 1D

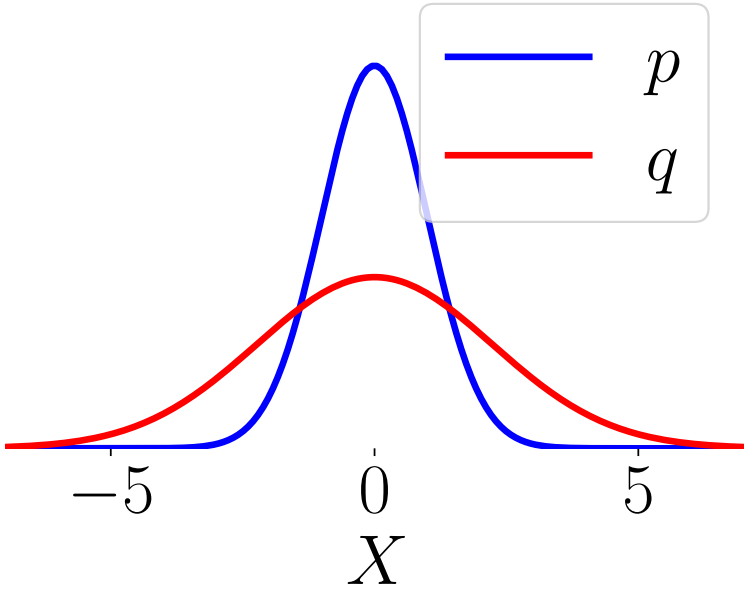


- Assume no differece in high-order moments.
- "Distance" = difference in the means. T-test.

$$\text{(population) } D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$$

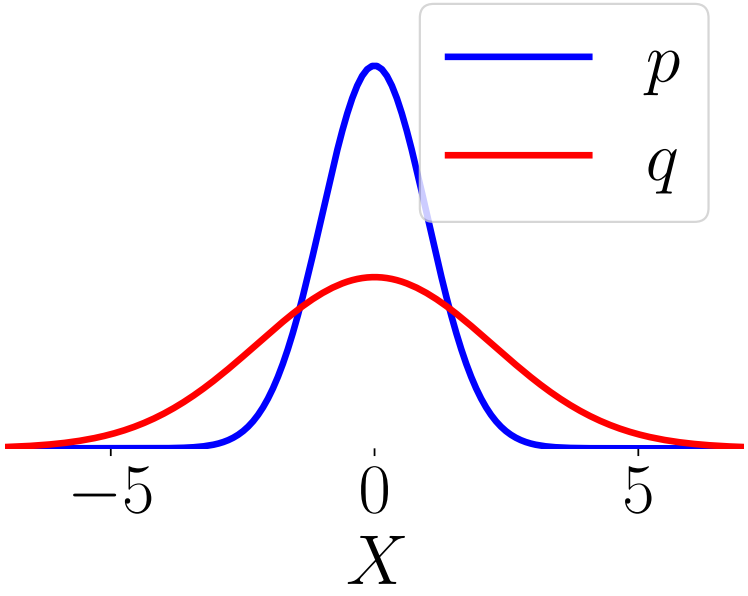
$$\text{(empirical) } \hat{D}_1(\mathsf{X}, \mathsf{Y}) = \left|\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right|$$

Case 2: Same Mean, Different Variances



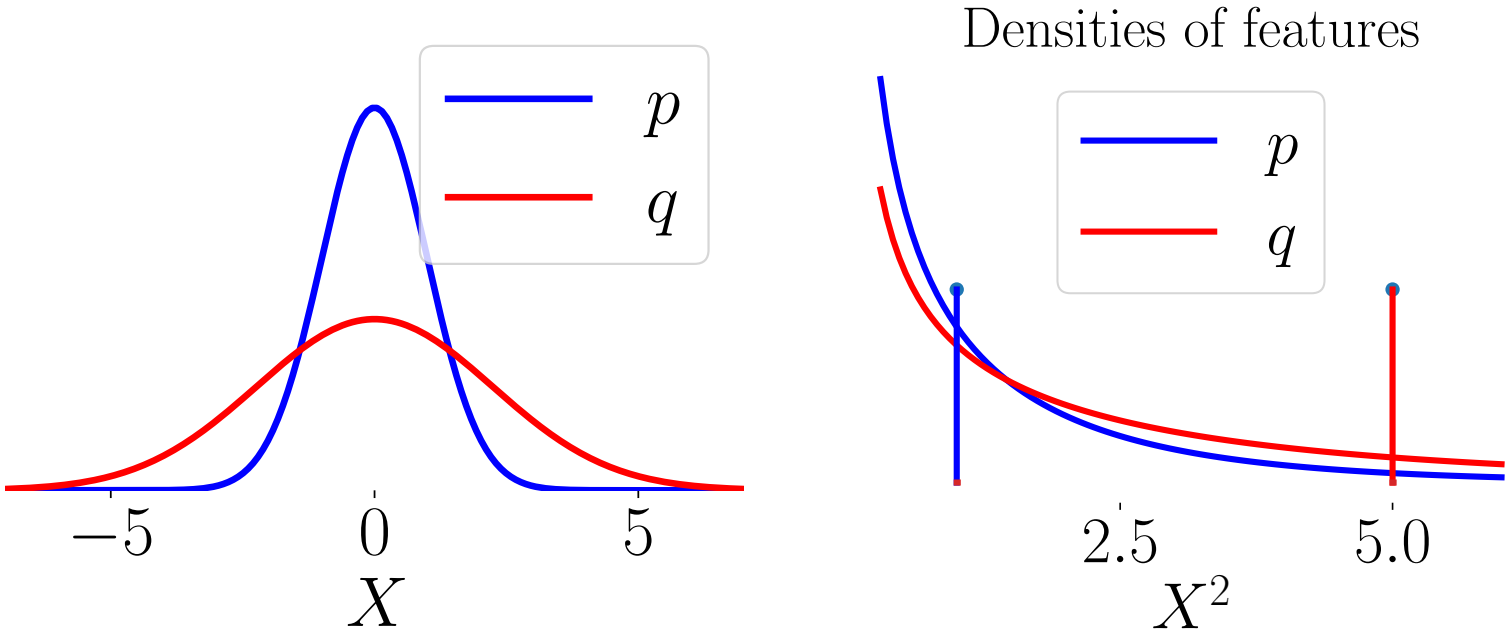
- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- Idea: look at difference in means of features $\phi(\cdot)$ of X and Y.

Case 2: Same Mean, Different Variances



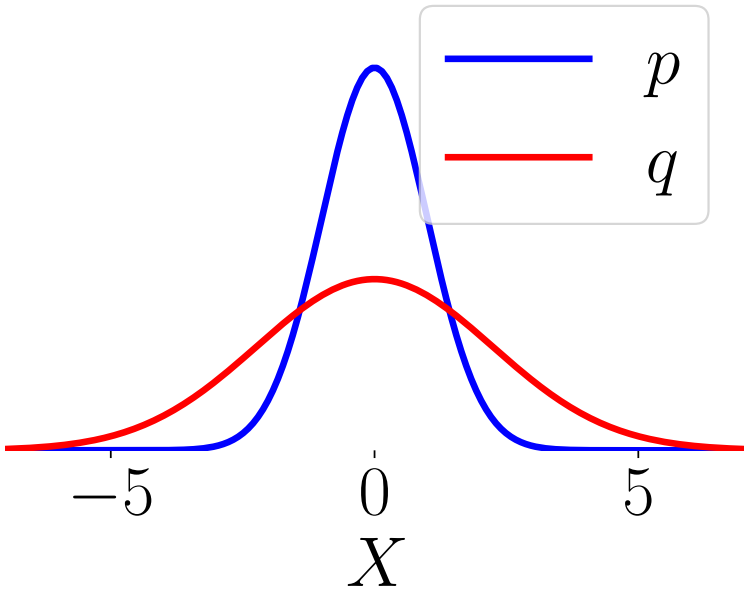
- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.

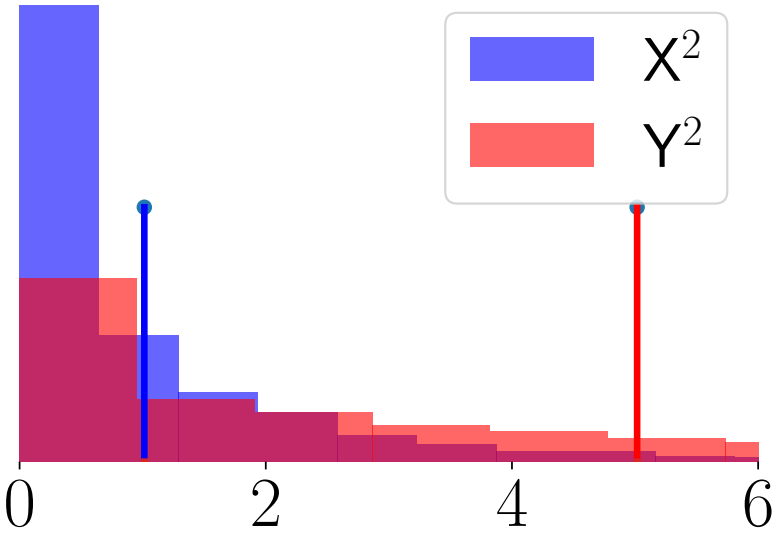
Case 2: Same Mean, Different Variances



- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.

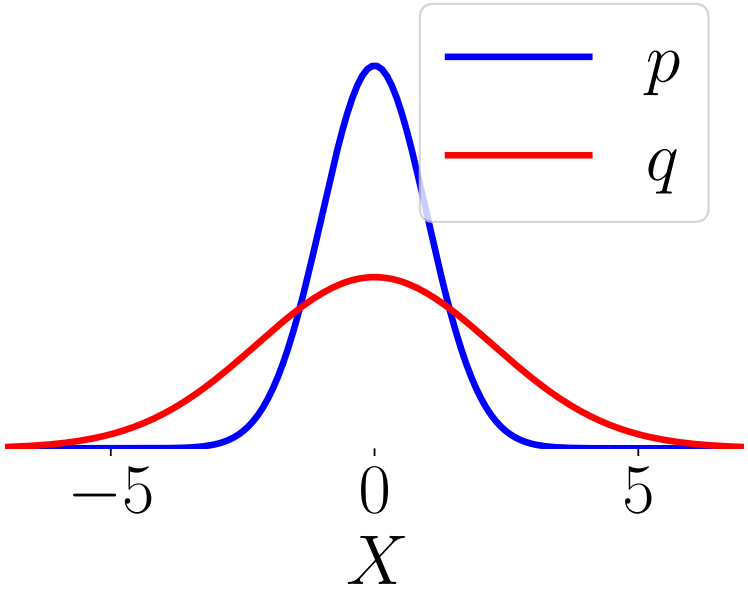
Case 2: Same Mean, Different Variances

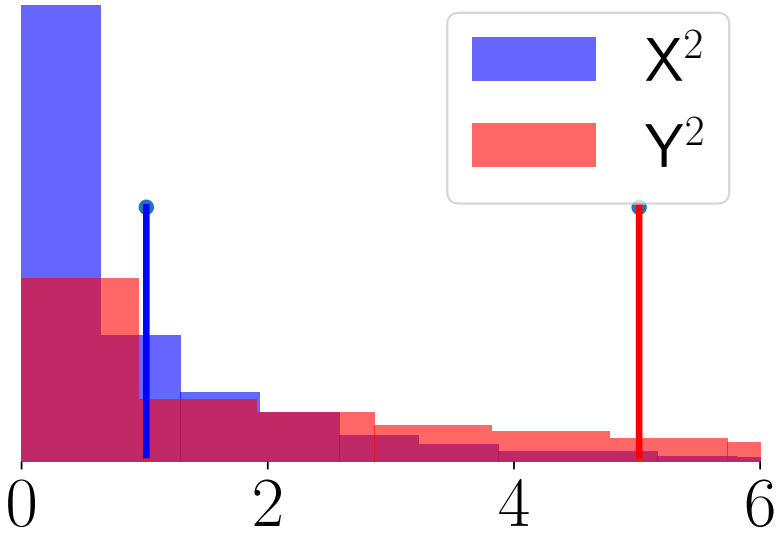




- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.

Case 2: Same Mean, Different Variances



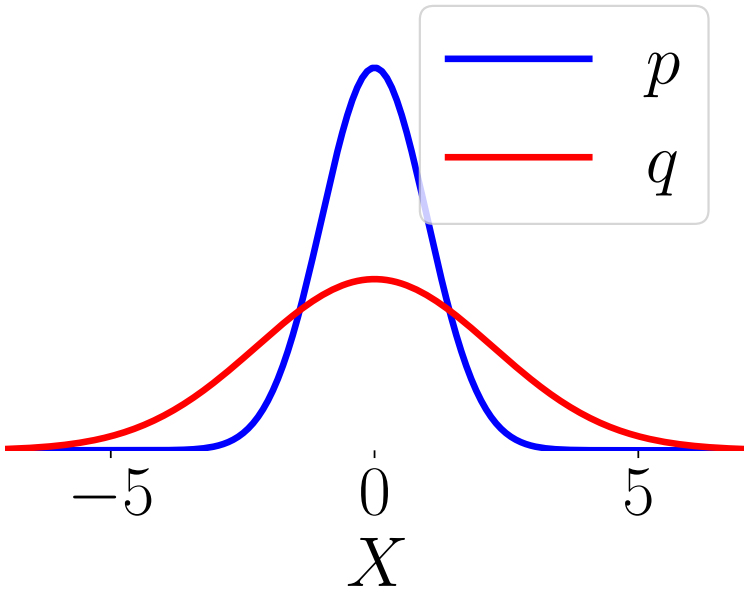


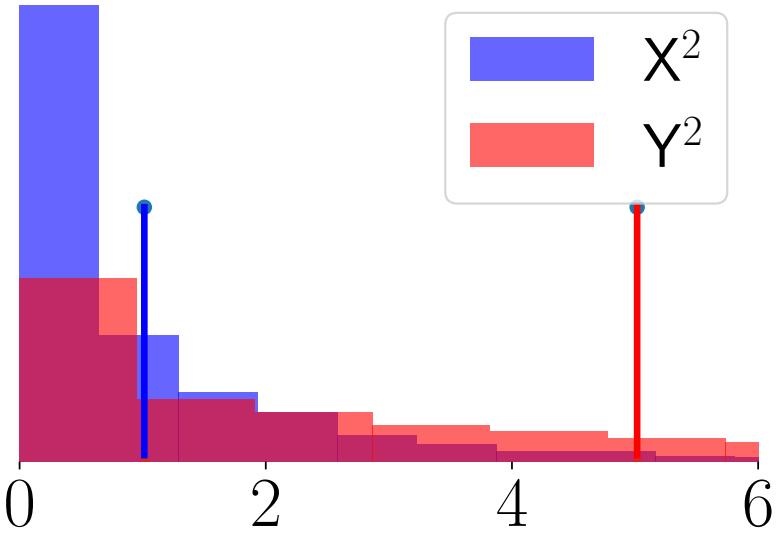
- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.
- New "distance":

$$D_2(p, q) = \|\mathbb{E}_{X\sim p}[\phi(X)] - \mathbb{E}_{Y\sim q}[\phi(X)]\|,$$

where $\phi(x) = (x, x^2)^\top$.

Case 2: Same Mean, Different Variances

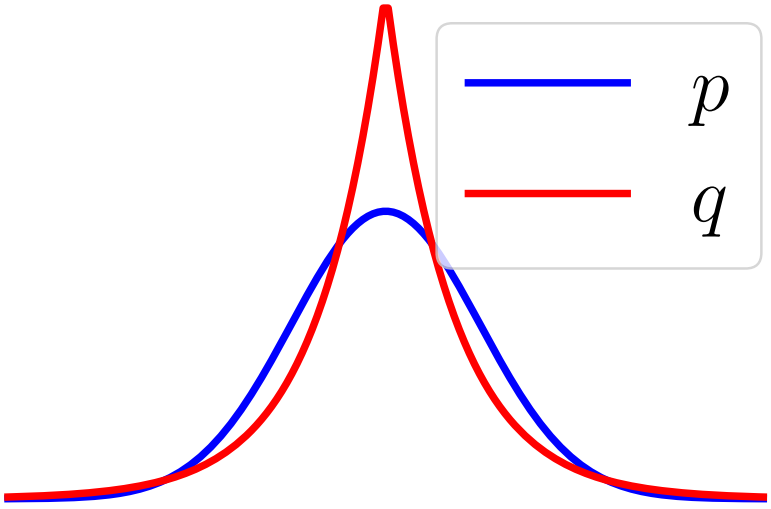




- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.

$$D_2(p, q) = \left\| \begin{pmatrix} \mathbb{E}_{X\sim p}[X] \\ \mathbb{E}_{X\sim p}[X^2] \end{pmatrix} - \begin{pmatrix} \mathbb{E}_{X\sim q}[X] \\ \mathbb{E}_{X\sim q}[X^2] \end{pmatrix} \right\|.$$

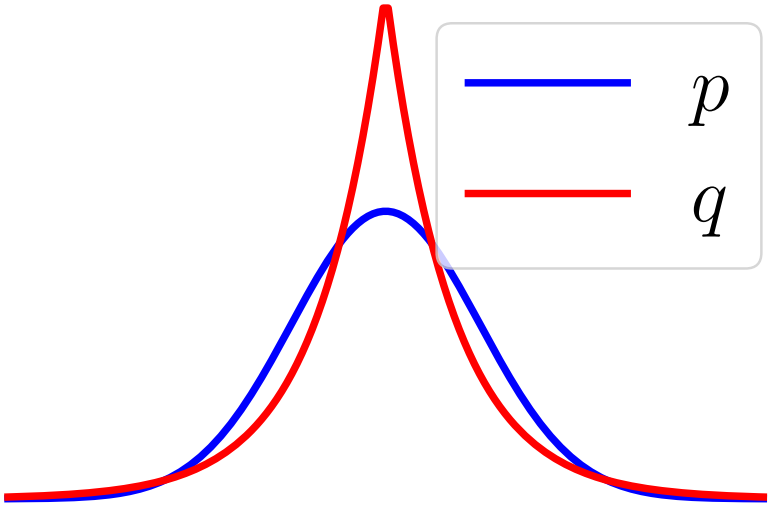
Case 3: Difference in High-Order Moments



- p = Gaussian distribution,
  q = Laplace distribution.
- Same mean **and** variance.
- D_2 fails.

- $\phi(x) = (x, x^2, x^4)^\top$ works. Difference is in kurtosis (4^{th} moment).
- $\phi(x) = (x, x^2, x^4, \cos x, e^x, \ldots)^\top$. But, when to stop?
- **Solution**: Use an infinite-dimensional feature map $\phi(\cdot)$ with the kernel trick.

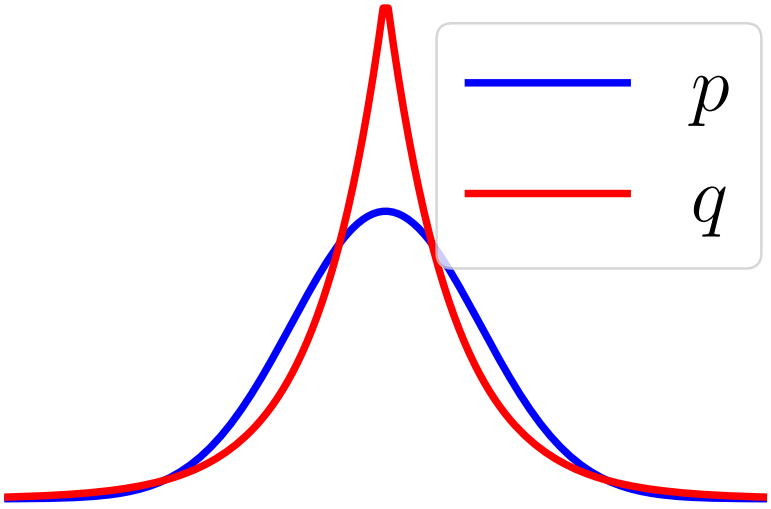
Case 3: Difference in High-Order Moments



- p = Gaussian distribution,
  q = Laplace distribution.
- Same mean **and** variance.
- D_2 fails.

- $\phi(x) = (x, x^2, x^4)^\top$ works. Difference is in kurtosis (4^{th} moment).
- $\phi(x) = (x, x^2, x^4, \cos x, e^x, \ldots)^\top$. But, when to stop?
- **Solution**: Use an infinite-dimensional feature map $\phi(\cdot)$ with the kernel trick.

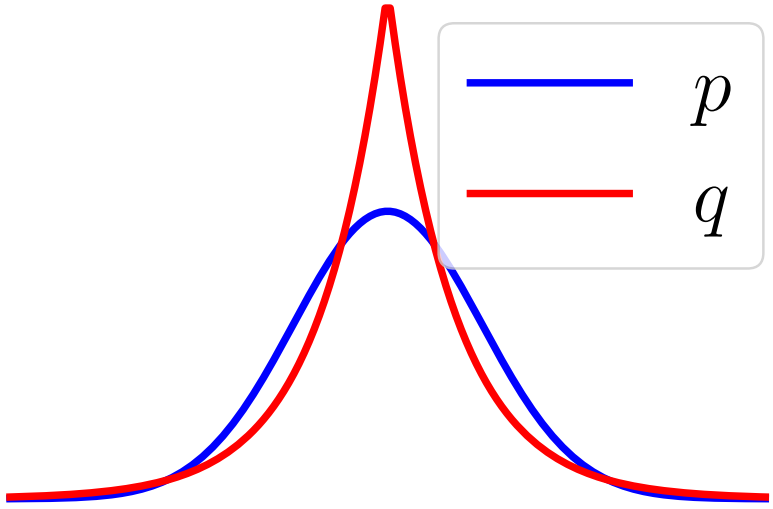
Case 3: Difference in High-Order Moments



- p = Gaussian distribution,
  q = Laplace distribution.
- Same mean **and** variance.
- D_2 fails.

- $\phi(x) = (x, x^2, x^4)^\top$ works. Difference is in kurtosis (4^{th} moment).
- $\phi(x) = (x, x^2, x^4, \cos x, e^x, \ldots)^\top$. But, when to stop?
- **Solution**: Use an infinite-dimensional feature map $\phi(\cdot)$ with the kernel trick.

Case 3: Difference in High-Order Moments



- p = Gaussian distribution,
  q = Laplace distribution.
- Same mean **and** variance.
- D_2 fails.

- $\phi(x) = (x, x^2, x^4)^\top$ works. Difference is in kurtosis (4^{th} moment).
- $\phi(x) = (x, x^2, x^4, \cos x, e^x, \ldots)^\top$. But, when to stop?
- **Solution**: Use an infinite-dimensional feature map $\phi(\cdot)$ with the kernel trick.

The (Kernel) Mean Embedding [Smola et al., 2007]

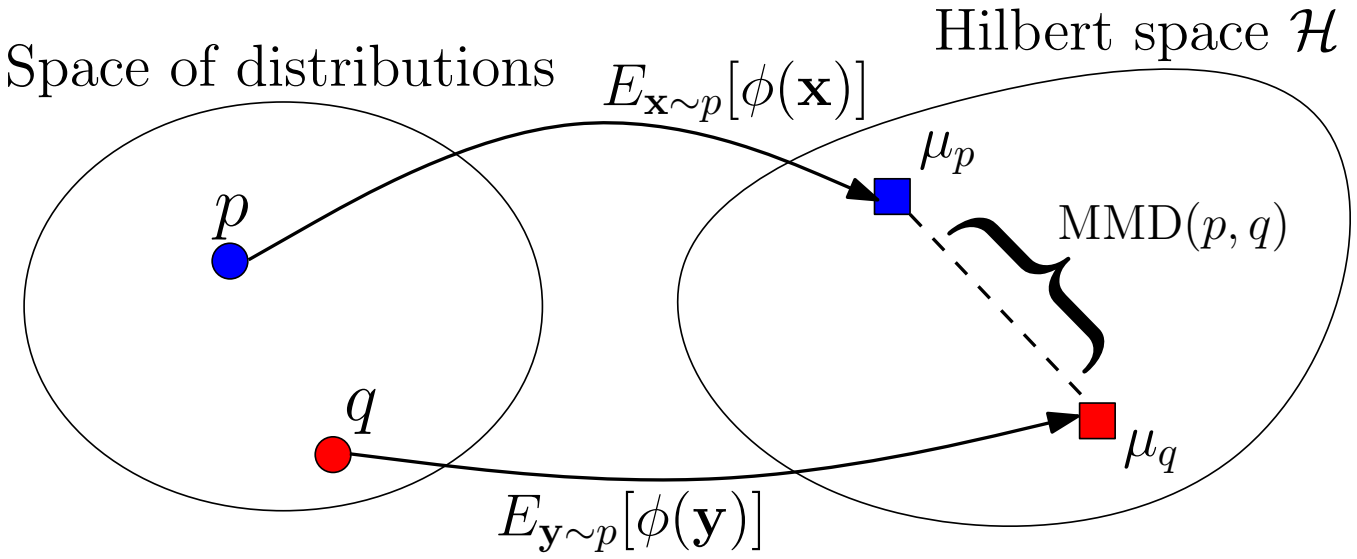
- Given a feature map $\phi(\cdot)$ mapping to a Hilbert space $\mathcal{H}$,
  - represent p with $\mu_p := \mathbb{E}_{\boldsymbol{x} \sim p}[\phi(\boldsymbol{x})]$ (i.e., the mean embedding of p),
  - represent q with $\mu_q := \mathbb{E}_{\boldsymbol{y} \sim q}[\phi(\boldsymbol{y})]$.

The (Kernel) Mean Embedding [Smola et al., 2007]

- Given a feature map $\phi(\cdot)$ mapping to a Hilbert space $\mathcal{H}$,
  - represent p with $\mu_p := \mathbb{E}_{\boldsymbol{x} \sim p}[\phi(\boldsymbol{x})]$ (i.e., the mean embedding of p),
  - represent q with $\mu_q := \mathbb{E}_{\boldsymbol{y} \sim q}[\phi(\boldsymbol{y})]$.

The (Kernel) Mean Embedding [Smola et al., 2007]

- Given a feature map $\phi(\cdot)$ mapping to a Hilbert space $\mathcal{H}$,
  - represent p with $\mu_p := \mathbb{E}_{\boldsymbol{x}\sim p}[\phi(\boldsymbol{x})]$ (i.e., the mean embedding of p),
  - represent q with $\mu_q := \mathbb{E}_{\boldsymbol{y}\sim q}[\phi(\boldsymbol{y})]$.



- $\mathcal{H}$ can be infinite dimensional. Depends on $\phi(\cdot)$.
  - If $\phi(x) = (x, x^2)^\top$, then $\mathcal{H} = \mathbb{R}^2$.
- Then, measure the distance in $\mathcal{H}$.
- The distance is called the "Maximum Mean Discrepancy" (MMD).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{\boldsymbol{x} \sim p}[\boldsymbol{\phi}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim q}[\boldsymbol{\phi}(\boldsymbol{y})] \right\|_{\mathcal{H}},$$

where $\boldsymbol{\phi}(\boldsymbol{x})$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2 \langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{\boldsymbol{x} \sim p}[\boldsymbol{\phi}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim q}[\boldsymbol{\phi}(\boldsymbol{y})] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{\boldsymbol{x} \sim p}[\boldsymbol{\phi}(\boldsymbol{x})], \mathbb{E}_{\boldsymbol{x}' \sim p}[\boldsymbol{\phi}(\boldsymbol{x}')] \rangle - 2 \langle \mathbb{E}_{\boldsymbol{x} \sim p}[\boldsymbol{\phi}(\boldsymbol{x})], \mathbb{E}_{\boldsymbol{y} \sim q}[\boldsymbol{\phi}(\boldsymbol{y})] \rangle \\
&\quad + \langle \mathbb{E}_{\boldsymbol{y} \sim q}[\boldsymbol{\phi}(\boldsymbol{y})], \mathbb{E}_{\boldsymbol{y}' \sim q}[\boldsymbol{\phi}(\boldsymbol{y}')] \rangle \\
&= \mathbb{E}_{\boldsymbol{x} \sim p} \mathbb{E}_{\boldsymbol{x}' \sim p} \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}') \rangle - 2 \mathbb{E}_{\boldsymbol{x} \sim p} \mathbb{E}_{\boldsymbol{y} \sim q} \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}) \rangle \\
&\quad + \mathbb{E}_{\boldsymbol{y} \sim q} \mathbb{E}_{\boldsymbol{y}' \sim q} \langle \boldsymbol{\phi}(\boldsymbol{y}), \boldsymbol{\phi}(\boldsymbol{y}') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}) \rangle$.
- Don't need $\boldsymbol{\phi}(\boldsymbol{x})$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{\boldsymbol{x} \sim p}[\boldsymbol{\phi}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim q}[\boldsymbol{\phi}(\boldsymbol{y})] \right\|_{\mathcal{H}},$$

where $\boldsymbol{\phi}(\boldsymbol{x})$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|\boldsymbol{a} - \boldsymbol{b}\|^2 = \langle \boldsymbol{a}, \boldsymbol{a} \rangle - 2 \langle \boldsymbol{a}, \boldsymbol{b} \rangle + \langle \boldsymbol{b}, \boldsymbol{b} \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{\boldsymbol{x} \sim p}[\boldsymbol{\phi}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim q}[\boldsymbol{\phi}(\boldsymbol{y})] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{\boldsymbol{x} \sim p}[\boldsymbol{\phi}(\boldsymbol{x})], \mathbb{E}_{\boldsymbol{x}' \sim p}[\boldsymbol{\phi}(\boldsymbol{x}')] \rangle - 2 \langle \mathbb{E}_{\boldsymbol{x} \sim p}[\boldsymbol{\phi}(\boldsymbol{x})], \mathbb{E}_{\boldsymbol{y} \sim q}[\boldsymbol{\phi}(\boldsymbol{y})] \rangle \\
&\quad + \langle \mathbb{E}_{\boldsymbol{y} \sim q}[\boldsymbol{\phi}(\boldsymbol{y})], \mathbb{E}_{\boldsymbol{y}' \sim q}[\boldsymbol{\phi}(\boldsymbol{y}')] \rangle \\
&= \mathbb{E}_{\boldsymbol{x} \sim p} \mathbb{E}_{\boldsymbol{x}' \sim p} \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}') \rangle - 2 \mathbb{E}_{\boldsymbol{x} \sim p} \mathbb{E}_{\boldsymbol{y} \sim q} \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}) \rangle \\
&\quad + \mathbb{E}_{\boldsymbol{y} \sim q} \mathbb{E}_{\boldsymbol{y}' \sim q} \langle \boldsymbol{\phi}(\boldsymbol{y}), \boldsymbol{\phi}(\boldsymbol{y}') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}) \rangle$.
- Don't need $\boldsymbol{\phi}(\boldsymbol{x})$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}},$$

where $\phi(x)$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2 \langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{x' \sim p}[\phi(x')] \rangle - 2 \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{y \sim q}[\phi(y)] \rangle \\
&\quad + \langle \mathbb{E}_{y \sim q}[\phi(y)], \mathbb{E}_{y' \sim q}[\phi(y')] \rangle \\
&= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} \langle \phi(x), \phi(x') \rangle - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} \langle \phi(x), \phi(y) \rangle \\
&\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} \langle \phi(y), \phi(y') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \phi(x), \phi(y) \rangle$.
- Don't need $\phi(x)$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}},$$

where $\phi(x)$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2 \langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{x' \sim p}[\phi(x')] \rangle - 2 \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{y \sim q}[\phi(y)] \rangle \\
&\quad + \langle \mathbb{E}_{y \sim q}[\phi(y)], \mathbb{E}_{y' \sim q}[\phi(y')] \rangle \\
&= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} \langle \phi(x), \phi(x') \rangle - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} \langle \phi(x), \phi(y) \rangle \\
&\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} \langle \phi(y), \phi(y') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \phi(x), \phi(y) \rangle$.
- Don't need $\phi(x)$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}},$$

where $\phi(x)$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2 \langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{x' \sim p}[\phi(x')] \rangle - 2 \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{y \sim q}[\phi(y)] \rangle \\
&\quad + \langle \mathbb{E}_{y \sim q}[\phi(y)], \mathbb{E}_{y' \sim q}[\phi(y')] \rangle \\
&= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} \langle \phi(x), \phi(x') \rangle - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} \langle \phi(x), \phi(y) \rangle \\
&\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} \langle \phi(y), \phi(y') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \phi(x), \phi(y) \rangle$.
- Don't need $\phi(x)$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}},$$

where $\phi(x)$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2\langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{x' \sim p}[\phi(x')] \rangle - 2\langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{y \sim q}[\phi(y)] \rangle \\
&\quad + \langle \mathbb{E}_{y \sim q}[\phi(y)], \mathbb{E}_{y' \sim q}[\phi(y')] \rangle \\
&= \mathbb{E}_{x \sim p}\mathbb{E}_{x' \sim p}\langle \phi(x), \phi(x') \rangle - 2\mathbb{E}_{x \sim p}\mathbb{E}_{y \sim q}\langle \phi(x), \phi(y) \rangle \\
&\quad + \mathbb{E}_{y \sim q}\mathbb{E}_{y' \sim q}\langle \phi(y), \phi(y') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \phi(x), \phi(y) \rangle$.
- Don't need $\phi(x)$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

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

$$\begin{aligned}\mathrm{MMD}_k^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\ &= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} k(x, x') - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} k(x, y') \\ &\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} k(y, y').\end{aligned}$$

- Unbiased estimator:

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

- $k(x, x') \approx$ similarity between x and x'.

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

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

$$\begin{aligned}\mathrm{MMD}_k^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\ &= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} k(x, x') - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} k(x, y') \\ &\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} k(y, y').\end{aligned}$$

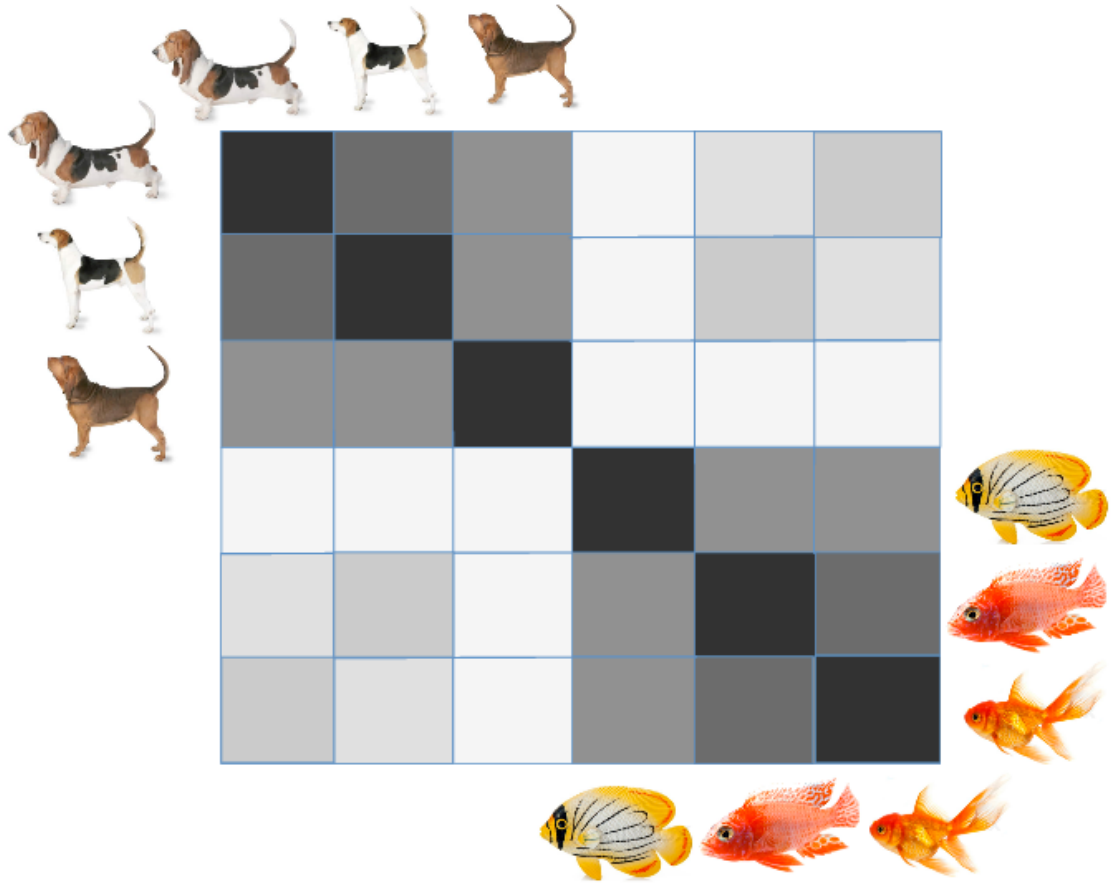
- Unbiased estimator:

$$\begin{aligned}\widehat{\mathrm{MMD}_k^2}(\mathrm{X}, \mathrm{Y}) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(x_i, x_j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, y_j) \\ &\quad + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(y_i, y_j).\end{aligned}$$

- $k(x, x') \approx$ similarity between x and x'.

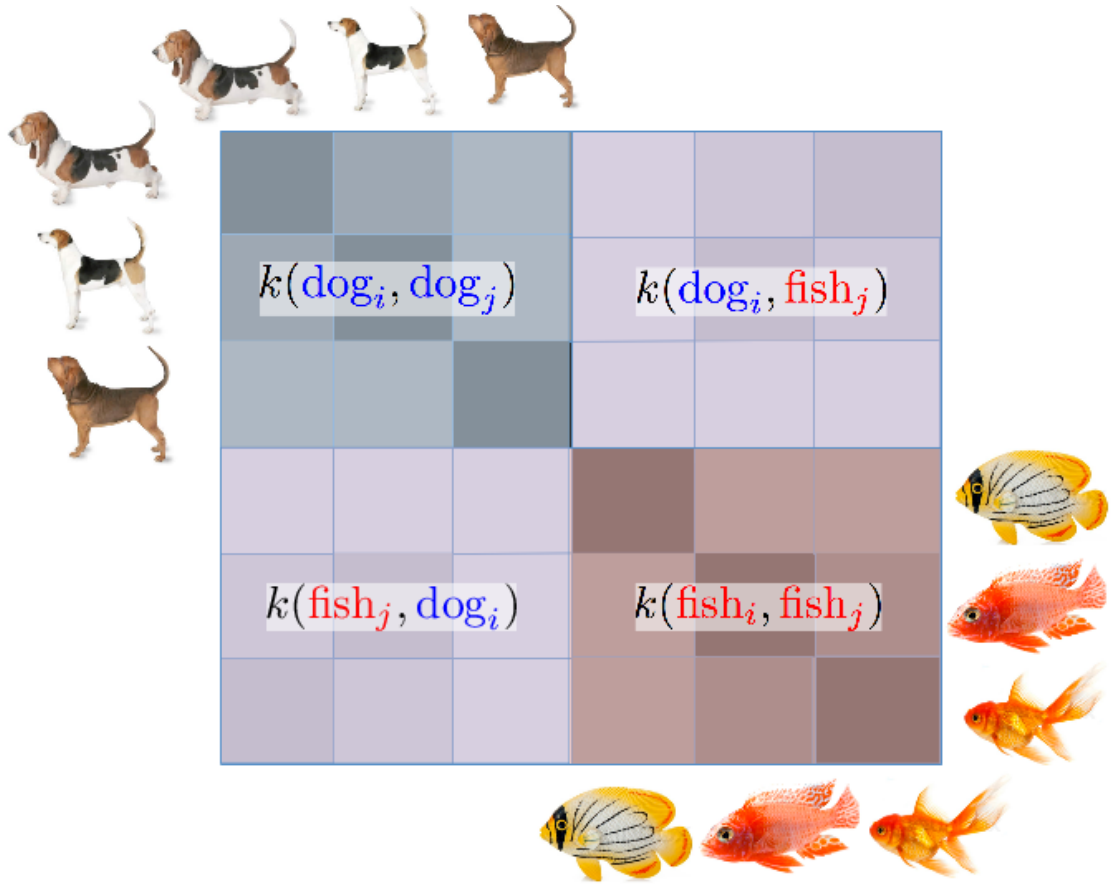
Intuition for the MMD

- Dogs $\sim p$ and fish $\sim q$.
- Each entry is one of $k(\text{dog}_i, \text{dog}_j)$, $k(\text{dog}_i, \text{fish}_j)$, or $k(\text{fish}_i, \text{fish}_j)$



Intuition for the MMD

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



Positive Definite Kernel

- Defining $k(x, x')$ from $\phi(\cdot)$ is always valid.
- Can start directly from $k(x, x')$ without specifying $\phi(\cdot)$.
- What k is valid?

Definition 2 (Positive definite kernel).

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite if, for any integer $n > 0$, $c_1, \ldots, c_n \in \mathbb{R}$, and $x_1, \ldots, x_n \in \mathcal{X}$, we have $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$.

- Equivalently, the Gram matrix $\boldsymbol{K}$ is a positive semi-definite matrix where $(\boldsymbol{K})_{ij} = k(x_i, x_j)$.
- x can be *anything* (e.g., vector, image, tree, string, graph, ...).

Positive Definite Kernel

- Defining $k(\boldsymbol{x}, \boldsymbol{x}')$ from $\boldsymbol{\phi}(\cdot)$ is always valid.
- Can start directly from $k(\boldsymbol{x}, \boldsymbol{x}')$ without specifying $\boldsymbol{\phi}(\cdot)$.
- What k is valid?

Definition 2 (Positive definite kernel).

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite if, for any integer $n > 0$, $c_1, \ldots, c_n \in \mathbb{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$, we have $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$.

- Equivalently, the Gram matrix $\boldsymbol{K}$ is a positive semi-definite matrix where $(\boldsymbol{K})_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
- $\boldsymbol{x}$ can be *anything* (e.g., vector, image, tree, string, graph, ...).

Positive Definite Kernel

- Defining $k(\boldsymbol{x}, \boldsymbol{x}')$ from $\phi(\cdot)$ is always valid.
- Can start directly from $k(\boldsymbol{x}, \boldsymbol{x}')$ without specifying $\phi(\cdot)$.
- What k is valid?

Definition 2 (Positive definite kernel).

A symmetric function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite if, for any integer $n > 0$, $c_1, \ldots, c_n \in \mathbb{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$, we have $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$.

- Equivalently, the Gram matrix $\boldsymbol{K}$ is a positive semi-definite matrix where $(\boldsymbol{K})_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
- $\boldsymbol{x}$ can be *anything* (e.g., vector, image, tree, string, graph, ...).

Positive Definite Kernel

- Defining $k(\boldsymbol{x}, \boldsymbol{x}')$ from $\phi(\cdot)$ is always valid.
- Can start directly from $k(\boldsymbol{x}, \boldsymbol{x}')$ without specifying $\phi(\cdot)$.
- What k is valid?

Definition 2 (Positive definite kernel).

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite if, for any integer $n > 0$, $c_1, \ldots, c_n \in \mathbb{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$, we have $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$.

- Equivalently, the Gram matrix $\boldsymbol{K}$ is a positive semi-definite matrix where $(\boldsymbol{K})_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
- $\boldsymbol{x}$ can be *anything* (e.g., vector, image, tree, string, graph, ...).

Positive Definite Kernel $\iff$ $\phi(\cdot)$ Exists

Theorem 1 (Moore-Aronszajn).

Assume $k(\cdot,\cdot)$ is positive definite.

1. *k is an inner product in some Hilbert space $\mathcal{H}$.*
2. *There exists $\phi(\cdot)$ such that $k(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.*

Summary: Pos. def. k automatically defines $\phi(\cdot)$ (implicitly).

- Defining k can be easier than defining $\phi(\cdot)$. Imagine strings.
- To study $\mathcal{H}$, can study $k(\cdot,\cdot)$ instead of $\phi(\cdot)$.
  - Reproducing kernel Hilbert spaces (RKHS).

Positive Definite Kernel $\iff$ $\phi(\cdot)$ Exists

Theorem 1 (Moore-Aronszajn).

Assume $k(\cdot,\cdot)$ is positive definite.

1. *k is an inner product in some Hilbert space $\mathcal{H}$.*
2. *There exists $\phi(\cdot)$ such that $k(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.*

Summary: Pos. def. k automatically defines $\phi(\cdot)$ (implicitly).

- Defining k can be easier than defining $\phi(\cdot)$. Imagine strings.
- To study $\mathcal{H}$, can study $k(\cdot,\cdot)$ instead of $\phi(\cdot)$.
  - Reproducing kernel Hilbert spaces (RKHS).

Positive Definite Kernel $\iff$ $\phi(\cdot)$ Exists

Theorem 1 (Moore-Aronszajn).

Assume $k(\cdot,\cdot)$ is positive definite.

1. *k is an inner product in some Hilbert space $\mathcal{H}$.*
2. *There exists $\phi(\cdot)$ such that $k(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.*

Summary: Pos. def. k automatically defines $\phi(\cdot)$ (implicitly).

- Defining k can be easier than defining $\phi(\cdot)$. Imagine strings.
- To study $\mathcal{H}$, can study $k(\cdot,\cdot)$ instead of $\phi(\cdot)$.
  - Reproducing kernel Hilbert spaces (RKHS).

Positive Definite Kernel $\iff$ $\phi(\cdot)$ Exists

Theorem 1 (Moore-Aronszajn).

Assume $k(\cdot,\cdot)$ is positive definite.

1. *k is an inner product in some Hilbert space $\mathcal{H}$.*
2. *There exists $\phi(\cdot)$ such that $k(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.*

Summary: Pos. def. k automatically defines $\phi(\cdot)$ (implicitly).

- Defining k can be easier than defining $\phi(\cdot)$. Imagine strings.
- To study $\mathcal{H}$, can study $k(\cdot,\cdot)$ instead of $\phi(\cdot)$.
  - Reproducing kernel Hilbert spaces (RKHS).

Example: Polynomial Kernel

Let $\mathcal{X} = \mathbb{R}^d$ (domain).

$$k(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^\top \boldsymbol{y}\right)^m$$

is positive definite, for $m \in \{1, 2, \ldots\}$.

- Consider $d = 2$ and $m = 3$.

$$k\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) = (x_1 y_1 + x_2 y_2)^3.$$

- Feature map

$$\phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1^3 \\ \sqrt{3} x_1^2 x_2 \\ \sqrt{3} x_1 x_2^2 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \\ \phi_4(x) \end{pmatrix}$$

  So, $\mathcal{H} = \mathbb{R}^4$.

- Show that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \phi(x)^\top \phi(y)$.

Example: Polynomial Kernel

Let $\mathcal{X} = \mathbb{R}^d$ (domain).

$$k(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^\top \boldsymbol{y}\right)^m$$

is positive definite, for $m \in \{1, 2, \ldots\}$.

- Consider $d = 2$ and $m = 3$.

$$k\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) = (x_1 y_1 + x_2 y_2)^3.$$

- Feature map

$$\phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1^3 \\ \sqrt{3} x_1^2 x_2 \\ \sqrt{3} x_1 x_2^2 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \\ \phi_4(x) \end{pmatrix}$$

  So, $\mathcal{H} = \mathbb{R}^4$.

- Show that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \phi(x)^\top \phi(y)$.

Example: Polynomial Kernel

Let $\mathcal{X} = \mathbb{R}^d$ (domain).

$$k(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^\top \boldsymbol{y}\right)^m$$

is positive definite, for $m \in \{1, 2, \ldots\}$.

- Consider $d = 2$ and $m = 3$.

$$k\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) = (x_1 y_1 + x_2 y_2)^3.$$

- Feature map

$$\phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1^3 \\ \sqrt{3} x_1^2 x_2 \\ \sqrt{3} x_1 x_2^2 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} \phi_1(\boldsymbol{x}) \\ \phi_2(\boldsymbol{x}) \\ \phi_3(\boldsymbol{x}) \\ \phi_4(\boldsymbol{x}) \end{pmatrix}$$

  So, $\mathcal{H} = \mathbb{R}^4$.

- Show that $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle_{\mathcal{H}} = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{y})$.

Example: Polynomial Kernel

Let $\mathcal{X} = \mathbb{R}^d$ (domain).

$$k(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^\top \boldsymbol{y}\right)^m$$

is positive definite, for $m \in \{1, 2, \ldots\}$.

- Consider $d = 2$ and $m = 3$.

$$k\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) = (x_1 y_1 + x_2 y_2)^3.$$

- Feature map

$$\phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1^3 \\ \sqrt{3} x_1^2 x_2 \\ \sqrt{3} x_1 x_2^2 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} \phi_1(\boldsymbol{x}) \\ \phi_2(\boldsymbol{x}) \\ \phi_3(\boldsymbol{x}) \\ \phi_4(\boldsymbol{x}) \end{pmatrix}$$

  So, $\mathcal{H} = \mathbb{R}^4$.

- Show that $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle_{\mathcal{H}} = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{y})$.

New Kernels from Old [Shawe-Taylor and Cristianini, 2004, p. 75]

- Assume k_1, k_2 are pos. def. kernels with feature maps ϕ_1 and ϕ_2.
- New kernel k with feature map ϕ.

$k(x, y)$	features
$k_1(x, y) + k_2(x, y)$	$\phi(x) =$ stack of $\phi_1(x)$ and $\phi_2(x)$
$k_1(f(x), f(x))$	$\phi(x) = \phi_1(f(x))$
$k_1(x, y)k_2(x, y)$	$\phi(x) =$ tensor product of $\phi_1(x), \phi_2(x)$
$\exp(k_1(x, y))$	$\approx$ weighted polynomial features of all orders
$\vdots$	$\vdots$

- $k(x, y) = -k_1(x, y)$ is NOT valid. Why?

New Kernels from Old [Shawe-Taylor and Cristianini, 2004, p. 75]

- Assume k_1, k_2 are pos. def. kernels with feature maps ϕ_1 and ϕ_2.
- New kernel k with feature map ϕ.

$k(x, y)$	features
$k_1(x, y) + k_2(x, y)$	$\phi(x) =$ stack of $\phi_1(x)$ and $\phi_2(x)$
$k_1(f(x), f(x))$	$\phi(x) = \phi_1(f(x))$
$k_1(x, y)k_2(x, y)$	$\phi(x) =$ tensor product of $\phi_1(x), \phi_2(x)$
$\exp(k_1(x, y))$	$\approx$ weighted polynomial features of all orders
$\vdots$	$\vdots$

- $k(x, y) = -k_1(x, y)$ is NOT valid. Why?

New Kernels from Old [Shawe-Taylor and Cristianini, 2004, p. 75]

- Assume k_1, k_2 are pos. def. kernels with feature maps ϕ_1 and ϕ_2.
- New kernel k with feature map ϕ.

$k(x, y)$	features
$k_1(x, y) + k_2(x, y)$	$\phi(x) =$ stack of $\phi_1(x)$ and $\phi_2(x)$
$k_1(f(x), f(x))$	$\phi(x) = \phi_1(f(x))$
$k_1(x, y)k_2(x, y)$	$\phi(x) =$ tensor product of $\phi_1(x), \phi_2(x)$
$\exp(k_1(x, y))$	$\approx$ weighted polynomial features of all orders
$\vdots$	$\vdots$

- $k(x, y) = -k_1(x, y)$ is NOT valid. Why?

New Kernels from Old [Shawe-Taylor and Cristianini, 2004, p. 75]

- Assume k_1, k_2 are pos. def. kernels with feature maps ϕ_1 and ϕ_2.
- New kernel k with feature map ϕ.

$k(x, y)$	features
$k_1(x, y) + k_2(x, y)$	$\phi(x) =$ stack of $\phi_1(x)$ and $\phi_2(x)$
$k_1(f(x), f(x))$	$\phi(x) = \phi_1(f(x))$
$k_1(x, y)k_2(x, y)$	$\phi(x) =$ tensor product of $\phi_1(x), \phi_2(x)$
$\exp(k_1(x, y))$	$\approx$ weighted polynomial features of all orders
$\vdots$	$\vdots$

- $k(x, y) = -k_1(x, y)$ is NOT valid. Why?

New Kernels from Old [Shawe-Taylor and Cristianini, 2004, p. 75]

- Assume k_1, k_2 are pos. def. kernels with feature maps ϕ_1 and ϕ_2.
- New kernel k with feature map ϕ.

$k(x, y)$	features
$k_1(x, y) + k_2(x, y)$	$\phi(x) =$ stack of $\phi_1(x)$ and $\phi_2(x)$
$k_1(f(x), f(x))$	$\phi(x) = \phi_1(f(x))$
$k_1(x, y)k_2(x, y)$	$\phi(x) =$ tensor product of $\phi_1(x), \phi_2(x)$
$\exp(k_1(x, y))$	$\approx$ weighted polynomial features of all orders
$\vdots$	$\vdots$

- $k(x, y) = -k_1(x, y)$ is NOT valid. Why?

New Kernels from Old [Shawe-Taylor and Cristianini, 2004, p. 75]

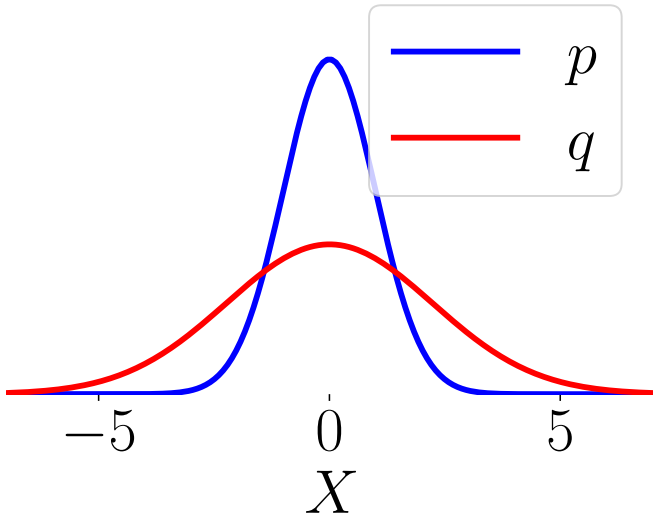
- Assume k_1, k_2 are pos. def. kernels with feature maps ϕ_1 and ϕ_2.
- New kernel k with feature map ϕ.

$k(x, y)$	features
$k_1(x, y) + k_2(x, y)$	$\phi(x) =$ stack of $\phi_1(x)$ and $\phi_2(x)$
$k_1(f(x), f(x))$	$\phi(x) = \phi_1(f(x))$
$k_1(x, y)k_2(x, y)$	$\phi(x) =$ tensor product of $\phi_1(x), \phi_2(x)$
$\exp(k_1(x, y))$	$\approx$ weighted polynomial features of all orders
$\vdots$	$\vdots$

- $k(x, y) = -k_1(x, y)$ is NOT valid. Why?

Non-Injective Mean Embedding

Variance difference revisited . . .



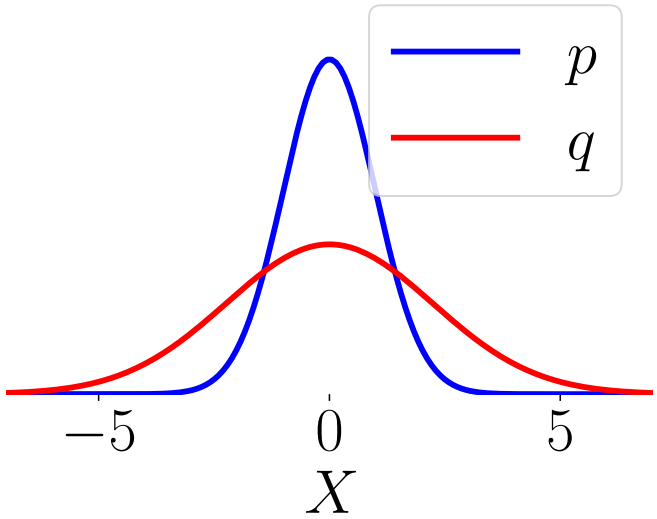
- We used $\phi(x) = x$. So, $k(x, y) = xy$ (linear kernel).
- $\mathrm{MMD}_k^2(p, q) = \left(\mathbb{E}_{X\sim p}[\phi(X)] - \mathbb{E}_{Y\sim q}[\phi(X)]\right)^2 = 0$ but $p \neq q$.

Why?

- k (and thus ϕ) is not powerful enough.
- Mathematically, the map $p \mapsto \mu_p$ is not injective.

Non-Injective Mean Embedding

Variance difference revisited . . .



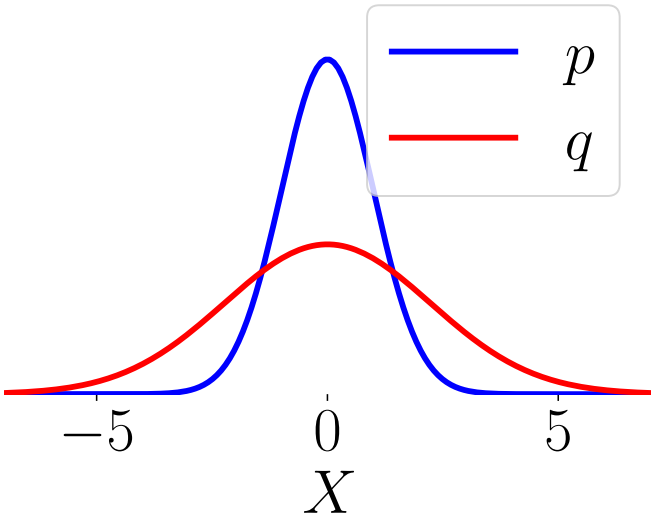
- We used $\phi(x) = x$. So, $k(x, y) = xy$ (linear kernel).
- $\mathrm{MMD}_k^2(p, q) = (\mathbb{E}_{X\sim p}[\phi(X)] - \mathbb{E}_{Y\sim q}[\phi(X)])^2 = 0$ but $p \neq q$.

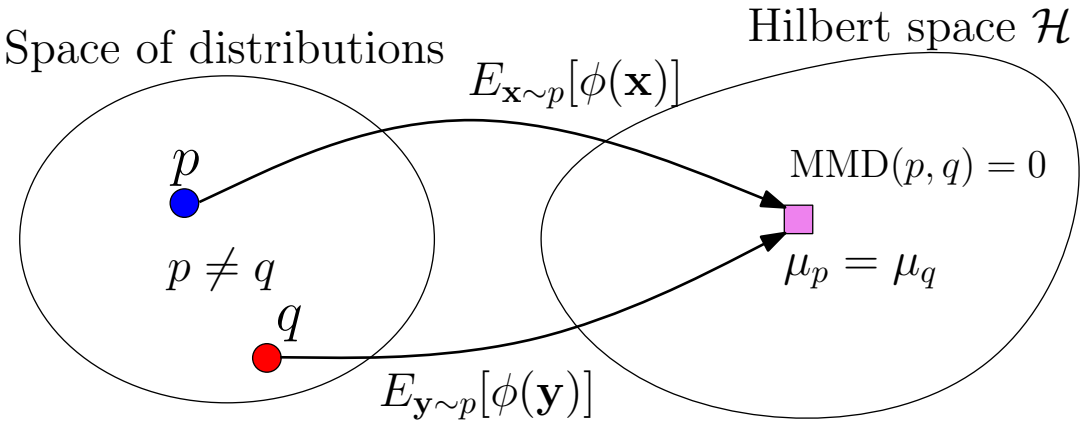
Why?

- k (and thus ϕ) is not powerful enough.
- Mathematically, the map $p \mapsto \mu_p$ is not injective.

Non-Injective Mean Embedding

Variance difference revisited ...





- We used $\phi(x) = x$. So, $k(x, y) = xy$ (linear kernel).
- $\mathrm{MMD}_k^2(p, q) = (\mathbb{E}_{X\sim p}[\phi(X)] - \mathbb{E}_{Y\sim q}[\phi(X)])^2 = 0$ but $p \neq q$.

Why?

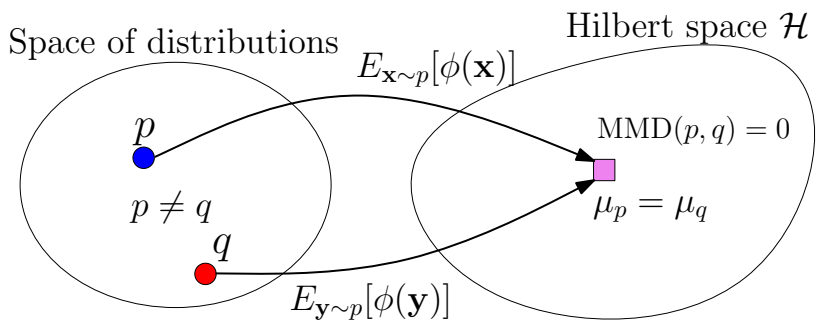
- k (and thus ϕ) is not powerful enough.
- Mathematically, the map $p \mapsto \mu_p$ is not injective.

Characteristic Kernels [Fukumizu et al., 2008]

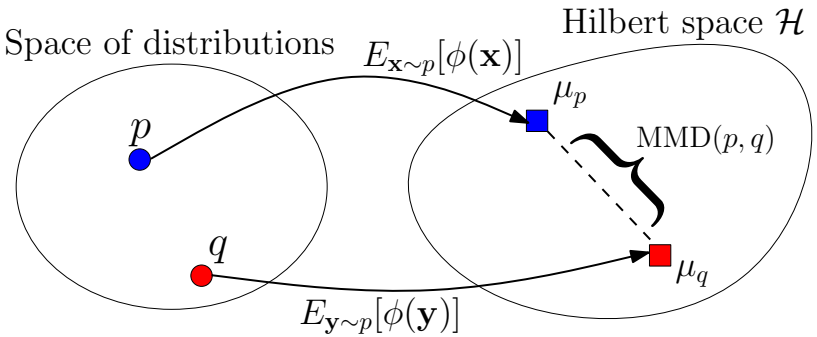
Definition 3.

A pos. def. kernel k is said to be **characteristic** if distinct distributions are embedded to different points in $\mathcal{H}$.

- Mathematically, $p \mapsto \mathbb{E}_{x \sim p}[\phi(x)]$ is injective.



not characteristic



characteristic

If k is characteristic, ...

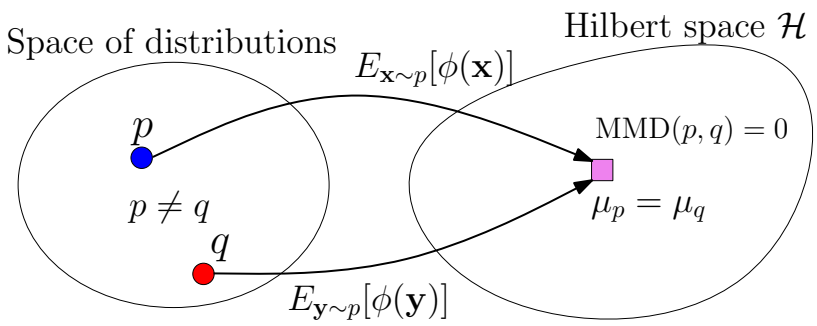
- μ_p contains all information of p,
- $\mathrm{MMD}_k(p,q) = 0$ if and only if $p = q$ [Gretton et al., 2012a]. A proper distance.

Characteristic Kernels [Fukumizu et al., 2008]

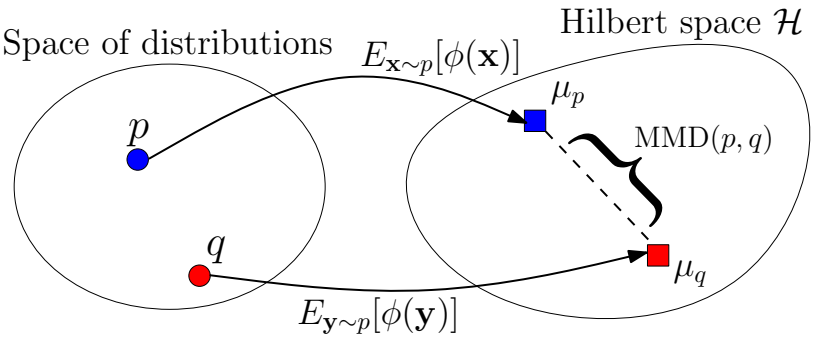
Definition 3.

A pos. def. kernel k is said to be **characteristic** if distinct distributions are embedded to different points in $\mathcal{H}$.

- Mathematically, $p \mapsto \mathbb{E}_{x \sim p}[\phi(x)]$ is injective.



not characteristic



characteristic

If k is characteristic, ...

- μ_p contains all information of p,
- $\mathrm{MMD}_k(p,q) = 0$ if and only if $p = q$ [Gretton et al., 2012a]. A proper distance.

Examples of Characteristic Kernels

Characteristic kernels on $\mathcal{X} = \mathbb{R}^d$:

- **Gaussian kernel**:

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

  for $\sigma > 0$.

- Laplace kernel: $k(x, y) = \exp\left(-\frac{\|x-y\|}{2\sigma}\right)$ for $\sigma > 0$.
- Matérn class of kernels [Rasmussen and Williams, 2006, Sec 4.2.1]
- etc. See [Sriperumbudur et al., 2010].

Not characteristic:

- Polynomial kernel: $k(x, y) = \left(x^\top y + c\right)^d$ for $c \geq 0, d \in \{1, 2, \ldots\}$.

Examples of Characteristic Kernels

Characteristic kernels on $\mathcal{X} = \mathbb{R}^d$:

- **Gaussian kernel**:

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

  for $\sigma > 0$.
- Laplace kernel: $k(x, y) = \exp\left(-\frac{\|x-y\|}{2\sigma}\right)$ for $\sigma > 0$.
- Matérn class of kernels [Rasmussen and Williams, 2006, Sec 4.2.1]
- etc. See [Sriperumbudur et al., 2010].

Not characteristic:

- Polynomial kernel: $k(x, y) = \left(x^\top y + c\right)^d$ for $c \geq 0, d \in \{1, 2, \ldots\}$.

Examples of Characteristic Kernels

Characteristic kernels on $\mathcal{X} = \mathbb{R}^d$:

- **Gaussian kernel**:

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

  for $\sigma > 0$.
- Laplace kernel: $k(x, y) = \exp\left(-\frac{\|x-y\|}{2\sigma}\right)$ for $\sigma > 0$.
- Matérn class of kernels [Rasmussen and Williams, 2006, Sec 4.2.1]
- etc. See [Sriperumbudur et al., 2010].

Not characteristic:

- Polynomial kernel: $k(x, y) = \left(x^\top y + c\right)^d$ for $c \geq 0, d \in \{1, 2, \ldots\}$.

Summary So Far

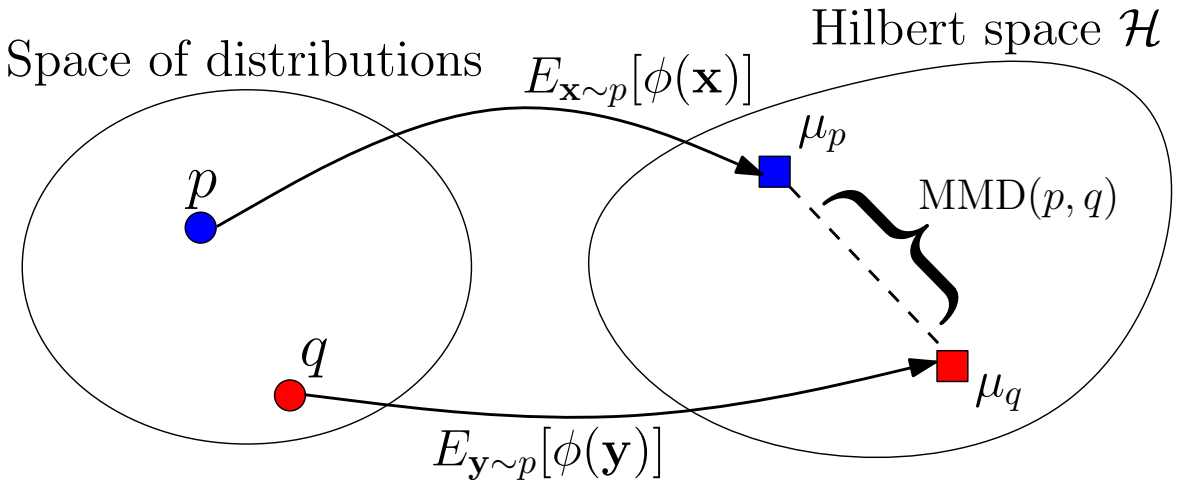
Only population quantities.

1. Cannot compute ∞-dimensional $\phi(\cdot)$. Can still compute $\mathrm{MMD}_k(p, q)$. Kernel trick.
2. Positive definite $k(\cdot, \cdot) \iff \phi(\cdot)$ exists.
3. If k is characteristic, $\mathbb{E}_{x \sim p}[\phi(x)]$ fully characterizes p.
4. Characteristic k implies $\mathrm{MMD}_k(p, q) = \|\mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)]\|$ iff $p = q$.

Summary So Far

Only population quantities.

1. Cannot compute ∞-dimensional $\phi(\cdot)$. Can still compute $\mathrm{MMD}_k(p, q)$. Kernel trick.
2. Positive definite $k(\cdot, \cdot) \iff \phi(\cdot)$ exists.
3. If k is characteristic, $\mathbb{E}_{x \sim p}[\phi(x)]$ fully characterizes p.

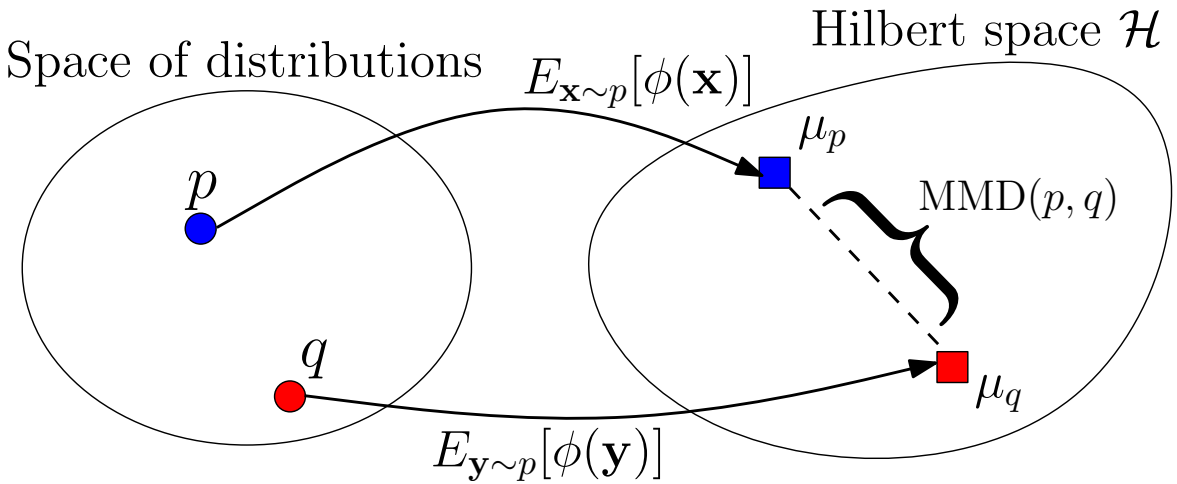


4. Characteristic k implies $\mathrm{MMD}_k(p, q) = \|\mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)]\|$ iff $p = q$.

Summary So Far

Only population quantities.

1. Cannot compute ∞-dimensional $\phi(\cdot)$. Can still compute $\mathrm{MMD}_k(p, q)$. Kernel trick.
2. Positive definite $k(\cdot, \cdot) \iff \phi(\cdot)$ exists.
3. If k is characteristic, $\mathbb{E}_{x \sim p}[\phi(x)]$ fully characterizes p.

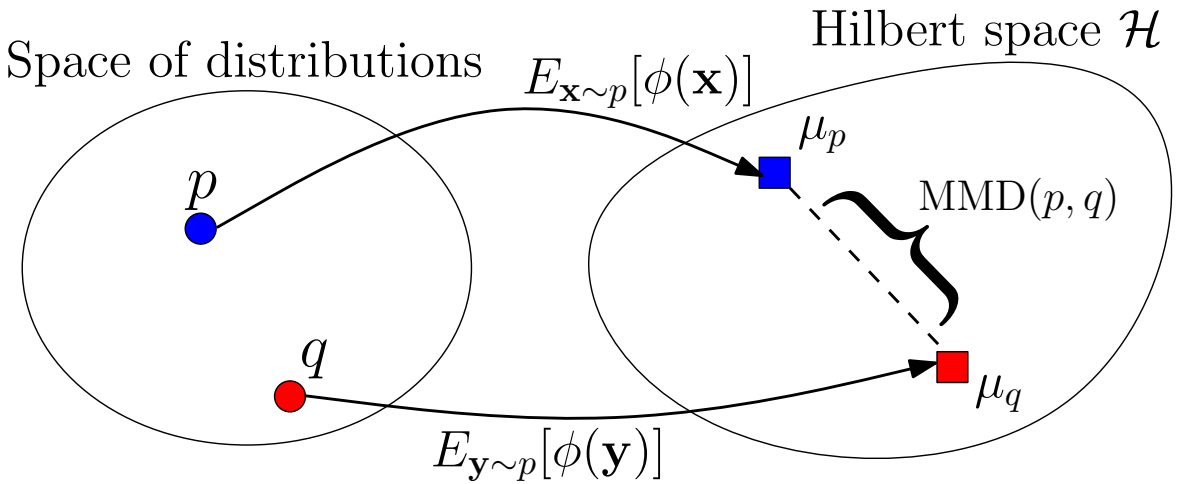


4. Characteristic k implies $\mathrm{MMD}_k(p, q) = \|\mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)]\|$ iff $p = q$.

Summary So Far

Only population quantities.

1. Cannot compute ∞-dimensional $\phi(\cdot)$. Can still compute $\mathrm{MMD}_k(p, q)$. Kernel trick.
2. Positive definite $k(\cdot, \cdot) \iff \phi(\cdot)$ exists.
3. If k is characteristic, $\mathbb{E}_{x \sim p}[\phi(x)]$ fully characterizes p.



4. Characteristic k implies $\mathrm{MMD}_k(p, q) = \|\mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)]\|$ iff $p = q$.

Outline

Two-Sample Testing with MMD

Have: Two collections of samples X, Y from unknown p and q.

Goal: Test $H_0 : p = q$ vs $H_1 : p \neq q$.

1. When $p = q$, $n\widehat{\mathrm{MMD}}_k^2$ (random) is "close to 0."
2. When $p \neq q$, $n\widehat{\mathrm{MMD}}_k^2$ is "far from 0."

Two-Sample Testing with MMD

Have: Two collections of samples X, Y from unknown p and q.

Goal: Test $H_0 : p = q$ vs $H_1 : p \neq q$.

1. When $p = q$, $n\widehat{\mathrm{MMD}}_k^2$ (random) is "close to 0."
2. When $p \neq q$, $n\widehat{\mathrm{MMD}}_k^2$ is "far from 0."

Two-Sample Testing with MMD

Have: Two collections of samples X, Y from unknown p and q.

Goal: Test $H_0 : p = q$ vs $H_1 : p \neq q$.

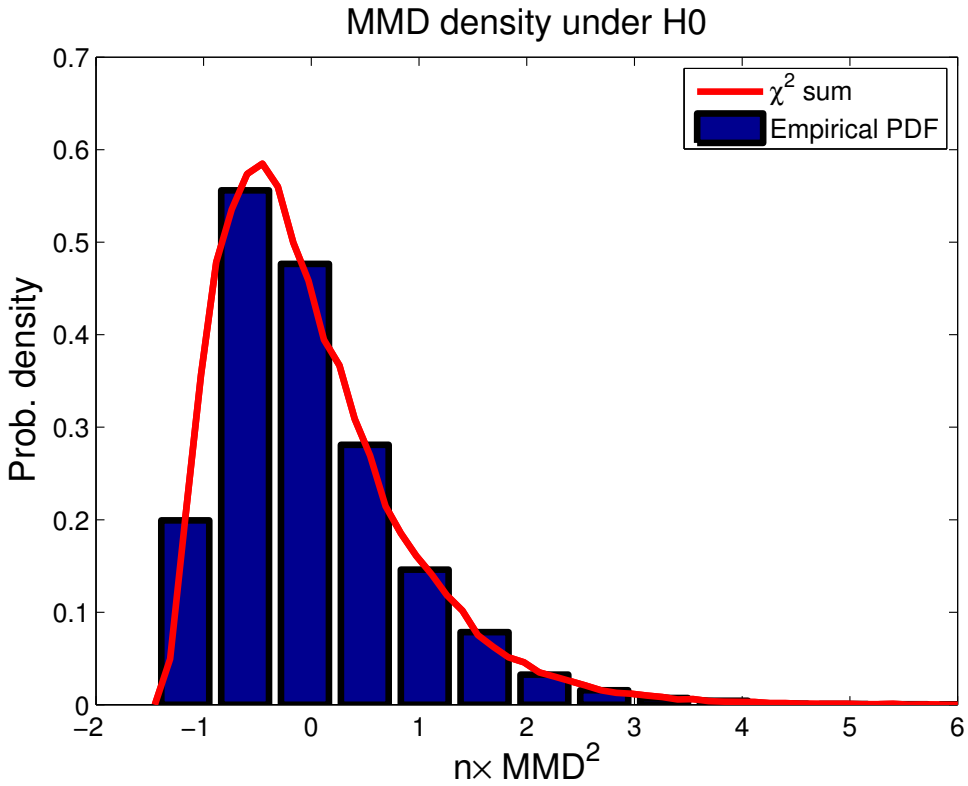
1. When $p = q$, $n\widehat{\mathrm{MMD}}_k^2$ (random) is "close to 0."
2. When $p \neq q$, $n\widehat{\mathrm{MMD}}_k^2$ is "far from 0."

Two-Sample Testing with MMD

Have: Two collections of samples X, Y from unknown p and q.

Goal: Test $H_0 : p = q$ vs $H_1 : p \neq q$.

1. When $p = q$, $n\widehat{\mathrm{MMD}}_k^2$ (random) is "close to 0."
2. When $p \neq q$, $n\widehat{\mathrm{MMD}}_k^2$ is "far from 0."

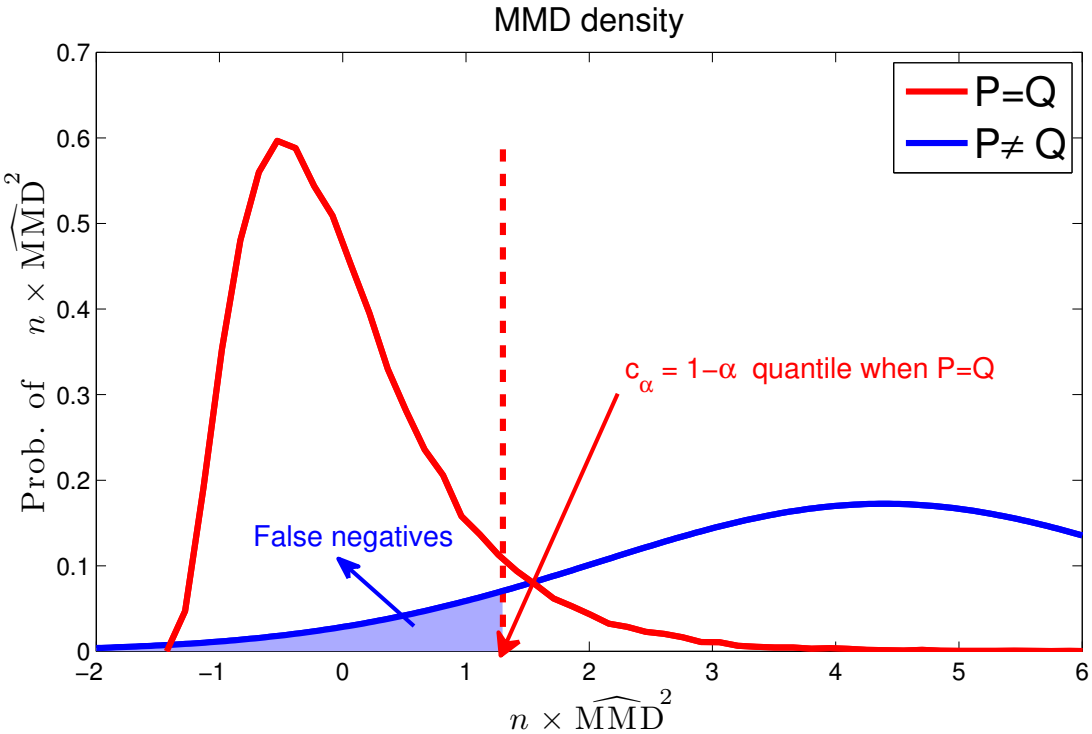


Two-Sample Testing with MMD

Have: Two collections of samples X, Y from unknown p and q.

Goal: Test $H_0 : p = q$ vs $H_1 : p \neq q$.

1. When $p = q$, $n\widehat{\mathrm{MMD}}_k^2$ (random) is "close to 0."
2. When $p \neq q$, $n\widehat{\mathrm{MMD}}_k^2$ is "far from 0."

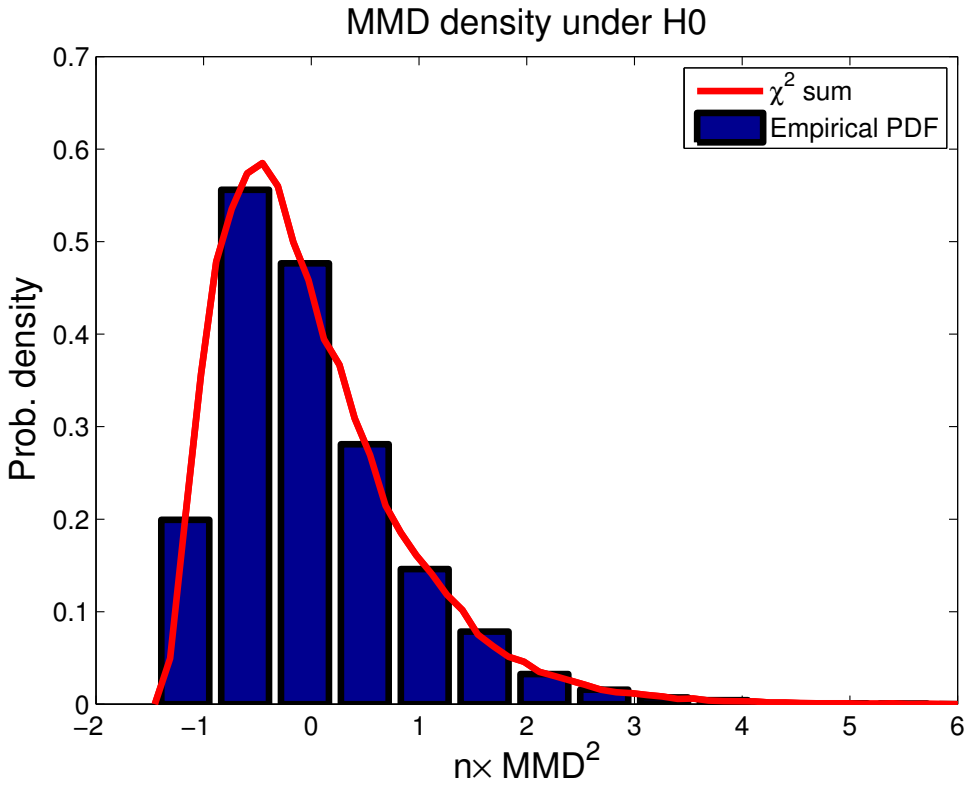


- Reject H_0 if $n\widehat{\mathrm{MMD}}_k^2 > c_\alpha$ (threshold).
- $\alpha =$ significance level.

Asymptotic Null Distribution of $n\widehat{\mathrm{MMD}}_k^2$

When $H_0 : p = q$, statistic has asymptotic distribution

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



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

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

Main point:

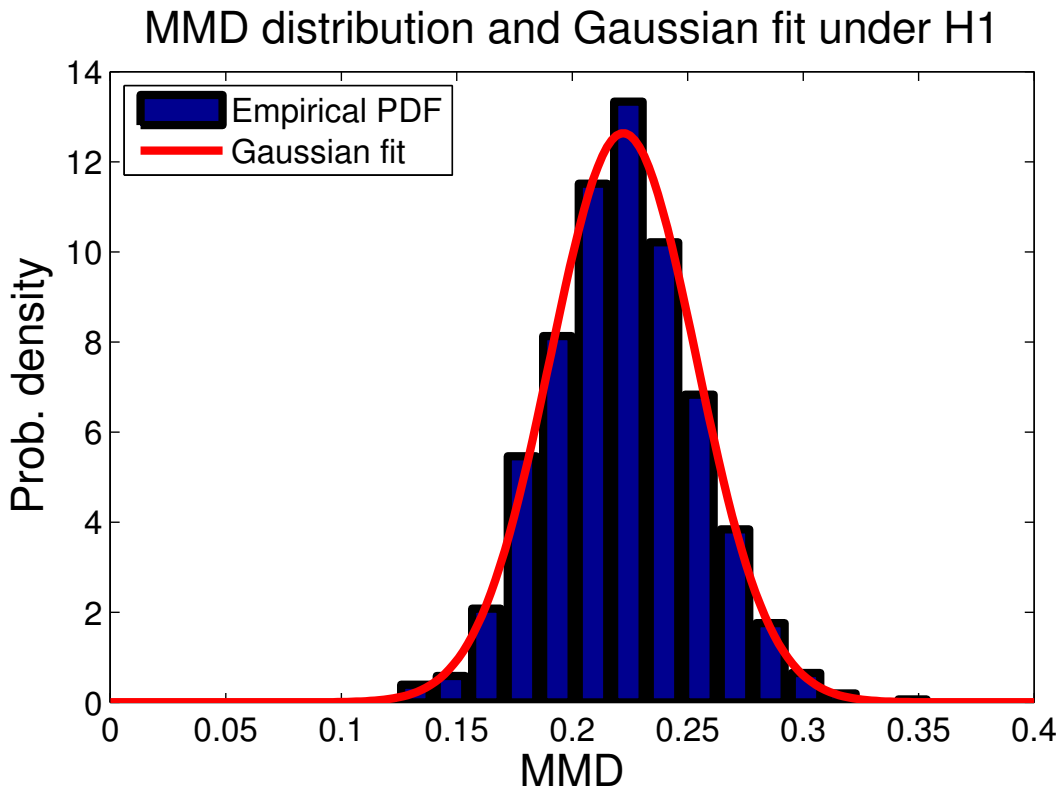
- Can simulate to get threshold c_α.

Asymptotic Distribution Under H_1 [Gretton et al., 2012a]

- When $H_1 : p \neq q$, statistic is asymptotically normal,

$$\sqrt{n}\left(\widehat{\mathrm{MMD}}_k^2 - \mathrm{MMD}_k^2(p,q)\right) \xrightarrow{d} \mathcal{N}\left(0, V_k(p,q)\right),$$

$V_k(p,q) =$ variance term.



Which Kernel to Use?

- Gaussian kernel: $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. Best σ^2?

- Keep false rejection rate at α. Maximize true rejection rate.
- Keep $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) \leq \alpha$. Maximize $\mathbb{P}(\text{reject } H_0 | H_1 \text{ true})$.

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$

Which Kernel to Use?

- Gaussian kernel: $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. Best σ^2?

- Keep false rejection rate at α. Maximize true rejection rate.
- Keep $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) \leq \alpha$. Maximize $\mathbb{P}(\text{reject } H_0 | H_1 \text{ true})$.

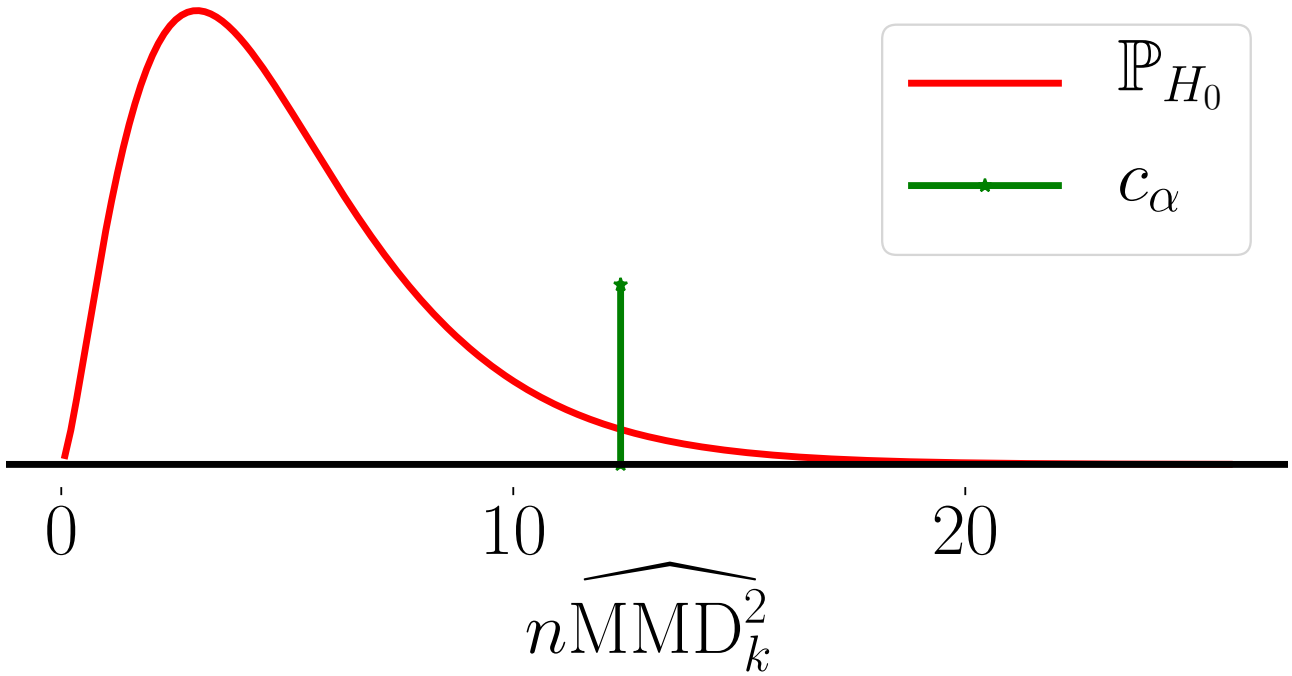
Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$

Which Kernel to Use?

- Gaussian kernel: $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. Best σ^2?

- Keep false rejection rate at α. Maximize true rejection rate.
- Keep $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) \leq \alpha$. Maximize $\mathbb{P}(\text{reject } H_0 | H_1 \text{ true})$.

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$

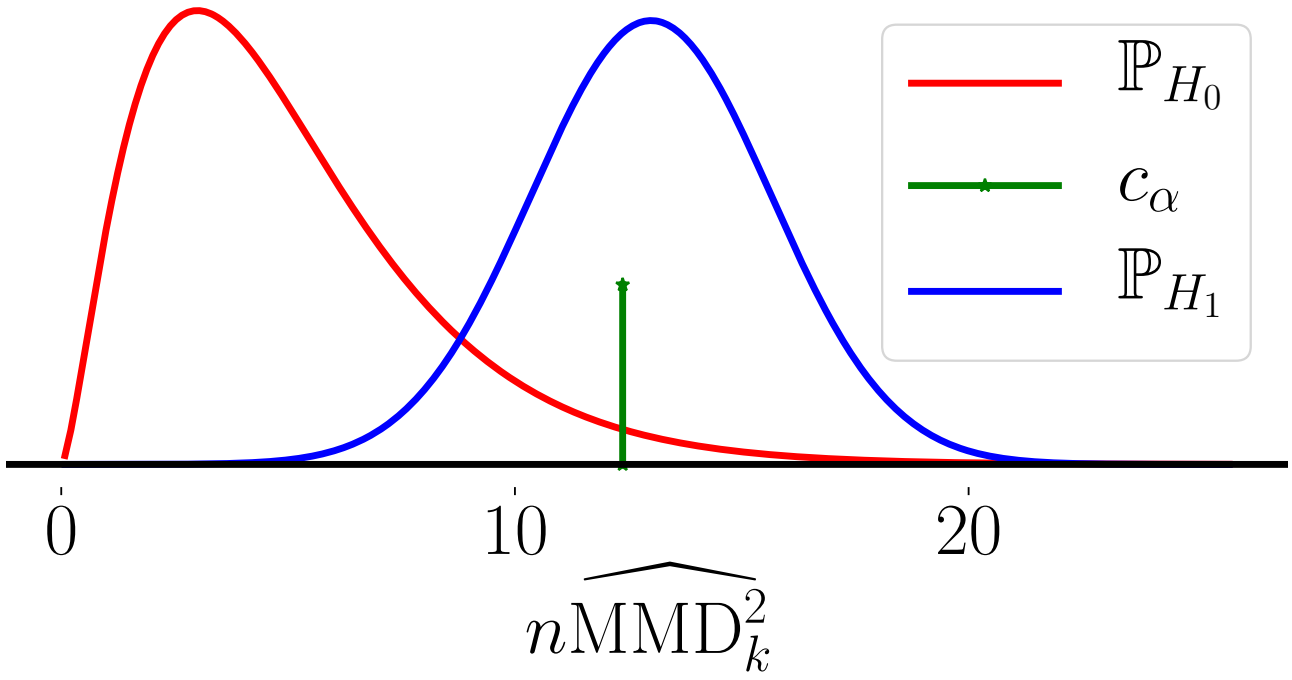


Which Kernel to Use?

- Gaussian kernel: $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. Best σ^2?

- Keep false rejection rate at α. Maximize true rejection rate.
- Keep $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) \leq \alpha$. Maximize $\mathbb{P}(\text{reject } H_0 | H_1 \text{ true})$.

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$

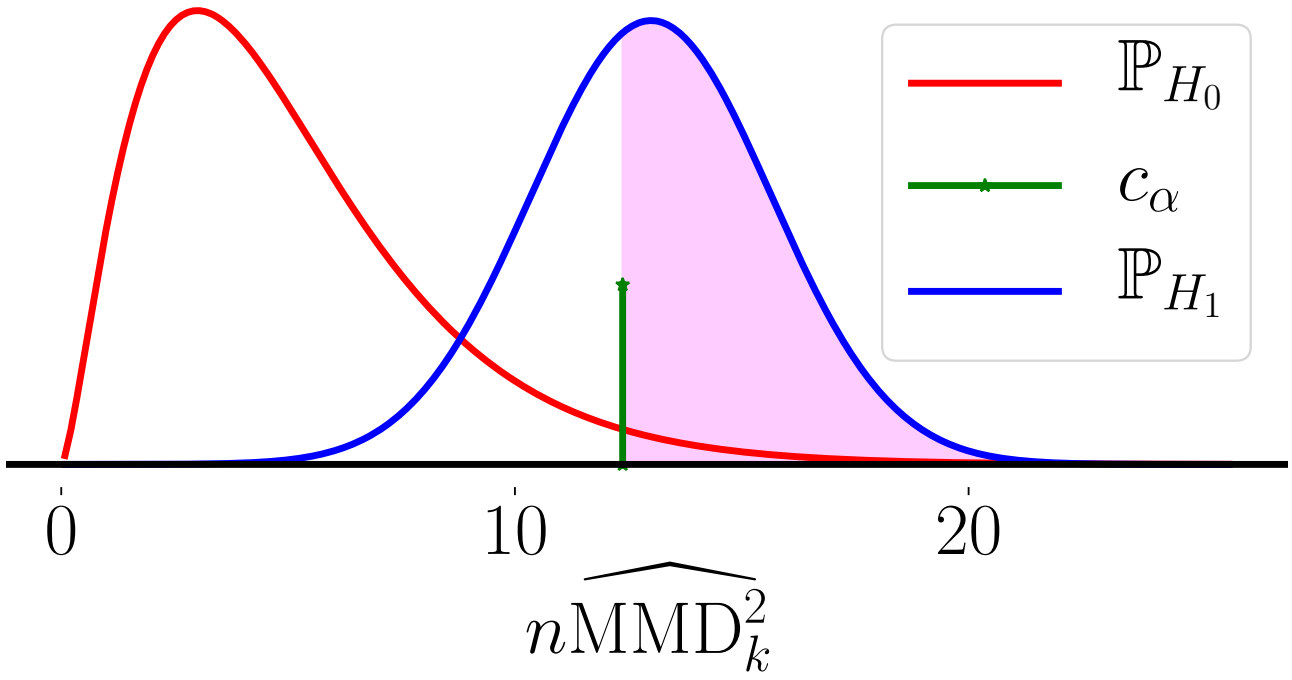


Which Kernel to Use?

- Gaussian kernel: $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. Best σ^2?

- Keep false rejection rate at α. Maximize true rejection rate.
- Keep $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) \leq \alpha$. Maximize $\mathbb{P}(\text{reject } H_0 | H_1 \text{ true})$.

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$

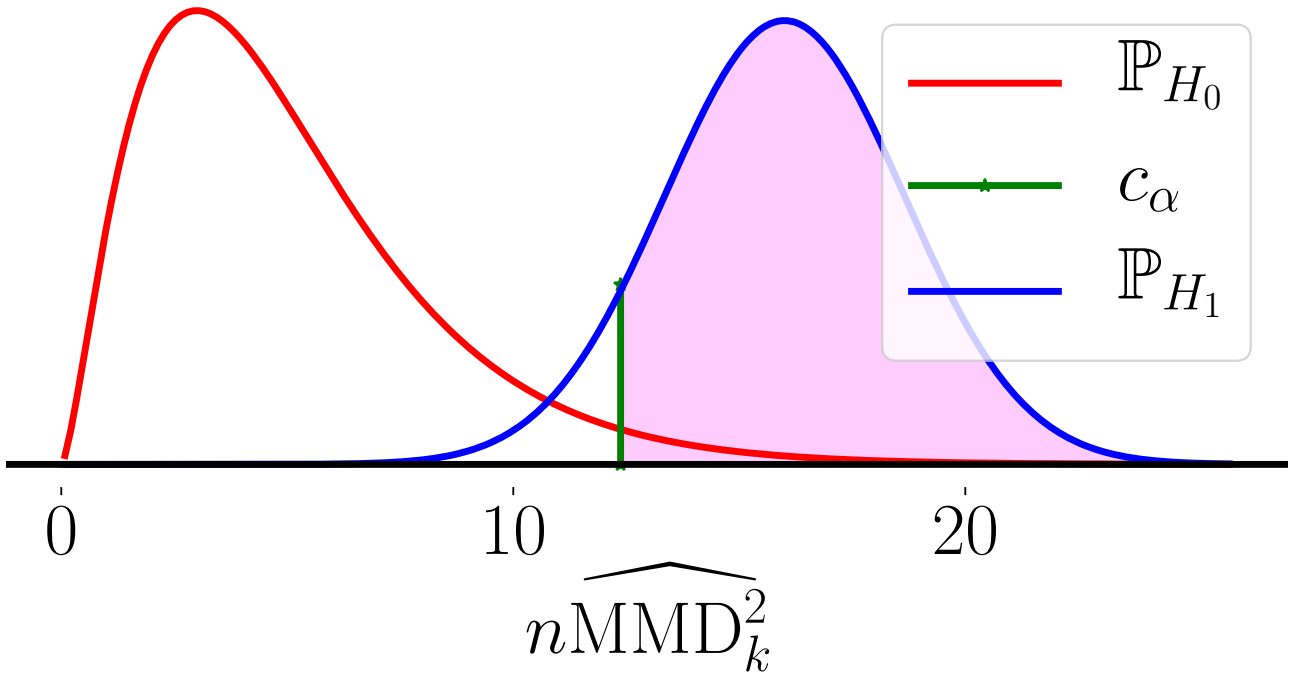


Which Kernel to Use?

- Gaussian kernel: $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. Best σ^2?

- Keep false rejection rate at α. Maximize true rejection rate.
- Keep $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) \leq \alpha$. Maximize $\mathbb{P}(\text{reject } H_0 | H_1 \text{ true})$.

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$



MMD Power Criterion [Sutherland et al., 2016]

- The test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) =$

$$\mathbb{P}_{H_1}\left(n\widehat{\mathrm{MMD}}_k^2 > \hat{c}_\alpha\right) \to \Phi\left(\sqrt{n}\frac{\mathrm{MMD}_k^2}{\sqrt{V_k}} - \frac{\hat{c}_\alpha}{\sqrt{n V_k}}\right),$$

  where

  - Φ is the CDF of the standard normal distribution.
  - $\hat{c}_\alpha$ is an estimate of the $1-\alpha$ quantile c_α of the null distribution.

- Choose the kernel which maximizes

$$\frac{\mathrm{MMD}_k^2}{\sqrt{V_k}}.$$

- Signal-to-noise ratio. Can be estimated with samples.

MMD Power Criterion [Sutherland et al., 2016]

- The test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) =$

$$\mathbb{P}_{H_1}\left(n\widehat{\mathrm{MMD}}_k^2 > \hat{c}_\alpha\right) \to \Phi\left(\sqrt{n}\frac{\mathrm{MMD}_k^2}{\sqrt{V_k}} - \frac{\hat{c}_\alpha}{\sqrt{nV_k}}\right),$$

  where
  - Φ is the CDF of the standard normal distribution.
  - $\hat{c}_\alpha$ is an estimate of the $1-\alpha$ quantile c_α of the null distribution.
- Choose the kernel which maximizes

$$\frac{\mathrm{MMD}_k^2}{\sqrt{V_k}}.$$

- Signal-to-noise ratio. Can be estimated with samples.

Deriving the MMD Power Criterion

- Let $Z \sim \mathcal{N}(0,1)$.

$$\mathbb{P}_{H_1}\left(n\widehat{\mathrm{MMD}}_k^2 > \hat{c}_\alpha\right)$$

$$= \mathbb{P}_{H_1}\left(\frac{n\widehat{\mathrm{MMD}}_k^2 - n\mathrm{MMD}_k^2}{\sqrt{V_k}} > \frac{\hat{c}_\alpha}{\sqrt{V_k}} - \frac{n\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$= \mathbb{P}_{H_1}\left(\frac{\sqrt{n}\widehat{\mathrm{MMD}}_k^2 - \sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}} > \frac{\hat{c}_\alpha}{\sqrt{nV_k}} - \frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$\to \mathbb{P}_{H_1}\left(Z > \frac{\hat{c}_\alpha}{\sqrt{nV_k}} - \frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$\to \Phi\left(\frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}} - \frac{\hat{c}_\alpha}{\sqrt{nV_k}}\right)$$

where

- Φ is the CDF of the standard normal distribution.
  - $\hat{c}_\alpha$ is an estimate of the $1-\alpha$ quantile c_α of the null distribution.

Deriving the MMD Power Criterion

- Let $Z \sim \mathcal{N}(0,1)$.

$$\mathbb{P}_{H_1}\left(n\widehat{\mathrm{MMD}}_k^2 > \hat{c}_\alpha\right)$$

$$= \mathbb{P}_{H_1}\left(\frac{n\widehat{\mathrm{MMD}}_k^2 - n\mathrm{MMD}_k^2}{\sqrt{V_k}} > \frac{\hat{c}_\alpha}{\sqrt{V_k}} - \frac{n\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$= \mathbb{P}_{H_1}\left(\frac{\sqrt{n}\widehat{\mathrm{MMD}}_k^2 - \sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}} > \frac{\hat{c}_\alpha}{\sqrt{nV_k}} - \frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$\to \mathbb{P}_{H_1}\left(Z > \frac{\hat{c}_\alpha}{\sqrt{nV_k}} - \frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$\to \Phi\left(\frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}} - \frac{\hat{c}_\alpha}{\sqrt{nV_k}}\right)$$

where

- Φ is the CDF of the standard normal distribution.
  - $\hat{c}_\alpha$ is an estimate of the $1-\alpha$ quantile c_α of the null distribution.

Deriving the MMD Power Criterion

- Let $Z \sim \mathcal{N}(0,1)$.

$$\mathbb{P}_{H_1}\left(n\widehat{\mathrm{MMD}}_k^2 > \hat{c}_\alpha\right)$$

$$= \mathbb{P}_{H_1}\left(\frac{n\widehat{\mathrm{MMD}}_k^2 - n\mathrm{MMD}_k^2}{\sqrt{V_k}} > \frac{\hat{c}_\alpha}{\sqrt{V_k}} - \frac{n\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$= \mathbb{P}_{H_1}\left(\frac{\sqrt{n}\widehat{\mathrm{MMD}}_k^2 - \sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}} > \frac{\hat{c}_\alpha}{\sqrt{nV_k}} - \frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$\to \mathbb{P}_{H_1}\left(Z > \frac{\hat{c}_\alpha}{\sqrt{nV_k}} - \frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$\to \Phi\left(\frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}} - \frac{\hat{c}_\alpha}{\sqrt{nV_k}}\right)$$

where

- Φ is the CDF of the standard normal distribution.
  - $\hat{c}_\alpha$ is an estimate of the $1-\alpha$ quantile c_α of the null distribution.

Deriving the MMD Power Criterion

- Let $Z \sim \mathcal{N}(0,1)$.

$$\mathbb{P}_{H_1}\left(n\widehat{\mathrm{MMD}}_k^2 > \hat{c}_\alpha\right)$$

$$= \mathbb{P}_{H_1}\left(\frac{n\widehat{\mathrm{MMD}}_k^2 - n\mathrm{MMD}_k^2}{\sqrt{V_k}} > \frac{\hat{c}_\alpha}{\sqrt{V_k}} - \frac{n\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$= \mathbb{P}_{H_1}\left(\frac{\sqrt{n}\widehat{\mathrm{MMD}}_k^2 - \sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}} > \frac{\hat{c}_\alpha}{\sqrt{nV_k}} - \frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$\to \mathbb{P}_{H_1}\left(Z > \frac{\hat{c}_\alpha}{\sqrt{nV_k}} - \frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}}\right)$$

$$\to \Phi\left(\frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}} - \frac{\hat{c}_\alpha}{\sqrt{nV_k}}\right)$$

where

- Φ is the CDF of the standard normal distribution.
  - $\hat{c}_\alpha$ is an estimate of the $1-\alpha$ quantile c_α of the null distribution.

Deriving the MMD Power Criterion

- Let $Z \sim \mathcal{N}(0, 1)$.

$$
\begin{aligned}
&\mathbb{P}_{H_1}\left(n\widehat{\mathrm{MMD}}_k^2 > \hat{c}_\alpha\right) \\
&= \mathbb{P}_{H_1}\left(\frac{n\widehat{\mathrm{MMD}}_k^2 - n\mathrm{MMD}_k^2}{\sqrt{V_k}} > \frac{\hat{c}_\alpha}{\sqrt{V_k}} - \frac{n\mathrm{MMD}_k^2}{\sqrt{V_k}}\right) \\
&= \mathbb{P}_{H_1}\left(\frac{\sqrt{n}\widehat{\mathrm{MMD}}_k^2 - \sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}} > \frac{\hat{c}_\alpha}{\sqrt{nV_k}} - \frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}}\right) \\
&\to \mathbb{P}_{H_1}\left(Z > \frac{\hat{c}_\alpha}{\sqrt{nV_k}} - \frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}}\right) \\
&\to \Phi\left(\frac{\sqrt{n}\mathrm{MMD}_k^2}{\sqrt{V_k}} - \frac{\hat{c}_\alpha}{\sqrt{nV_k}}\right)
\end{aligned}
$$

where

- Φ is the CDF of the standard normal distribution.
  - $\hat{c}_\alpha$ is an estimate of the $1-\alpha$ quantile c_α of the null distribution.

Properties of the MMD Test

As sample size $n \to \infty$,

1. If $H_0 : p = q$, then $\mathbb{P}(\text{reject } H_0) \leq \alpha$.
2. If k is characteristic and $H_1 : p \neq q$, then $\mathbb{P}(\text{reject } H_0) \to 1$.

- (1) and (2) $\implies$ a **consistent test**.
- $\widehat{\text{MMD}}_k^2$ can be estimated in $\mathcal{O}(n^2)$ time.
  - But, linear-time versions ($\mathcal{O}(n)$) exist [Gretton et al., 2012a, Zaremba et al., 2013].

Properties of the MMD Test

As sample size $n \to \infty$,

1. If $H_0 : p = q$, then $\mathbb{P}(\text{reject } H_0) \leq \alpha$.
2. If k is characteristic and $H_1 : p \neq q$, then $\mathbb{P}(\text{reject } H_0) \to 1$.

- (1) and (2) $\implies$ a **consistent test**.
- $\widehat{\text{MMD}}_k^2$ can be estimated in $\mathcal{O}(n^2)$ time.
  - But, linear-time versions ($\mathcal{O}(n)$) exist [Gretton et al., 2012a, Zaremba et al., 2013].

Properties of the MMD Test

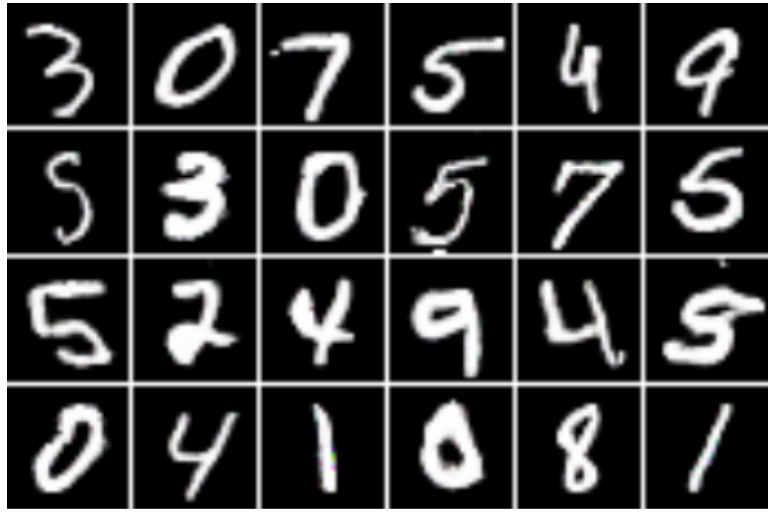
As sample size $n \to \infty$,

1. If $H_0 : p = q$, then $\mathbb{P}(\text{reject } H_0) \leq \alpha$.
2. If k is characteristic and $H_1 : p \neq q$, then $\mathbb{P}(\text{reject } H_0) \to 1$.

- (1) and (2) $\implies$ a **consistent test**.
- $\widehat{\text{MMD}}_k^2$ can be estimated in $\mathcal{O}(n^2)$ time.
  - But, linear-time versions $(\mathcal{O}(n))$ exist [Gretton et al., 2012a, Zaremba et al., 2013].

Generate MNIST Handwritten Digits

Observed MNIST handwritten digits. X.



Generated images from a model. Y.

- **Goal**: Learn a function which transforms noise into a handwritten digit.

Generative Moment Matching Networks [Li et al., 2015]

Generative Moment Matching Networks

Yujia Li[1] YUJIALI@CS.TORONTO.EDU
Kevin Swersky[1] KSWERSKY@CS.TORONTO.EDU
Richard Zemel[1,2] ZEMEL@CS.TORONTO.EDU
[1]Department of Computer Science, University of Toronto, Toronto, ON, CANADA
[2]Canadian Institute for Advanced Research, Toronto, ON, CANADA

- ICML 2015.
- Code: https://github.com/yujiali/gmmn
- One of the first to use MMD to train a generative network.

More Recent Works on MMD Based Generative Nets

MMD GAN: Towards Deeper Understanding of Moment Matching Network
Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, Barnabás Póczos
https://arxiv.org/abs/1705.08584

Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy
Dougal J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, Arthur Gretton
ICLR 2017
https://arxiv.org/abs/1611.04488

Demystifying MMD GANs
Mikolaj Binkowski, Dougal J. Sutherland, Michael Arbel, Arthur Gretton
ICLR 2018
https://openreview.net/pdf?id=r1lUOzWCW

Generative Moment Matching Networks [Li et al., 2015]

$$\arg\min_{\theta} \widehat{\mathrm{MMD}}_k^2 \left(\mathrm{X}, \{g_\theta(z_i)\}_{i=1}^n\right)$$

- $\mathrm{X} =$ training sets. $x_i =$ one digit (an image with $28 \times 28 = 784$ pixels). 60000 images.
- $\mathrm{Z} = \{z_i\}_{i=1}^n$ random noise vectors. Drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
- $g_\theta(z)$ a deep net transforming noise z into an image.
- Kernel k: sum of 5 Gaussian kernels of different bandwidths

Network architecture (my own, not [Li et al., 2015]):

- 4 hidden layers. Total parameters 60,608 (in θ).
- Training for 15 epochs. ≈ 7 minutes. My laptop without GPU.

Generative Moment Matching Networks [Li et al., 2015]

$$\arg\min_{\theta} \widehat{\mathrm{MMD}}_k^2 \left(\mathrm{X}, \{g_\theta(z_i)\}_{i=1}^n\right)$$

- $\mathrm{X} =$ training sets. $x_i =$ one digit (an image with $28 \times 28 = 784$ pixels). 60000 images.
- $\mathrm{Z} = \{z_i\}_{i=1}^n$ random noise vectors. Drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
- $g_\theta(z)$ a deep net transforming noise z into an image.
- Kernel k: sum of 5 Gaussian kernels of different bandwidths

Network architecture (my own, not [Li et al., 2015]):

- 4 hidden layers. Total parameters 60,608 (in θ).
- Training for 15 epochs. $\approx$ 7 minutes. My laptop without GPU.

Generative Moment Matching Networks [Li et al., 2015]

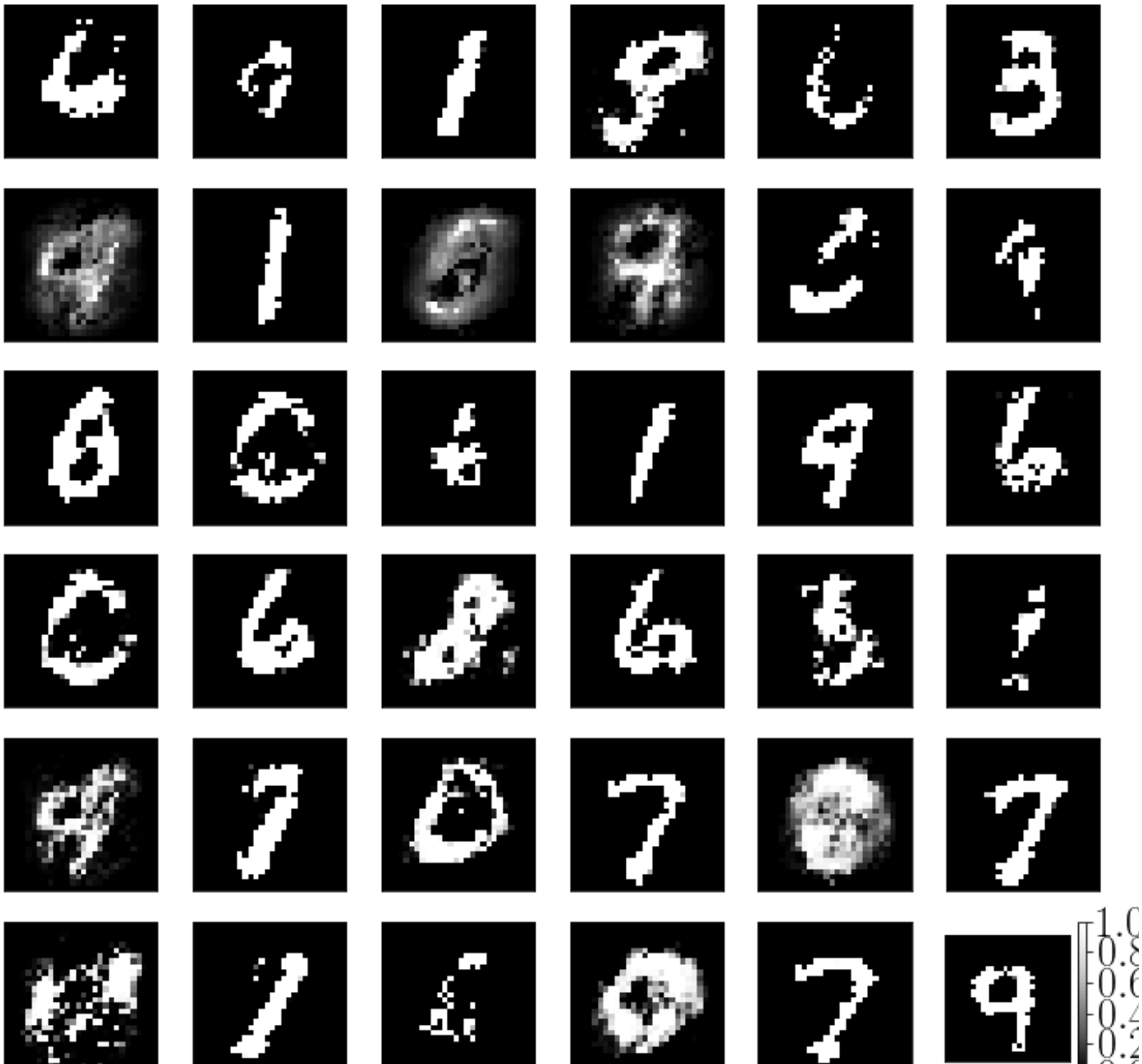
$$\arg\min_{\theta} \widehat{\mathrm{MMD}}_k^2 \left(\mathrm{X}, \{g_\theta(z_i)\}_{i=1}^n\right)$$

- $\mathrm{X} =$ training sets. $x_i =$ one digit (an image with $28 \times 28 = 784$ pixels). 60000 images.
- $\mathrm{Z} = \{z_i\}_{i=1}^n$ random noise vectors. Drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
- $g_\theta(z)$ a deep net transforming noise z into an image.
- Kernel k: sum of 5 Gaussian kernels of different bandwidths

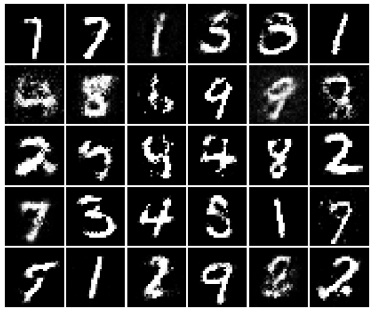
Network architecture (my own, not [Li et al., 2015]):

- 4 hidden layers. Total parameters 60,608 (in θ).
- Training for 15 epochs. ≈ 7 minutes. My laptop without GPU.

My Results



Quick Comments

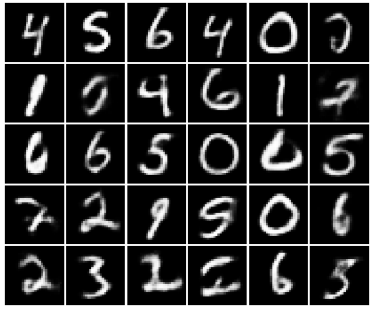


(a) GMMN MNIST samples

(b) GMMN TFD samples

(e) GMMN nearest neighbors for MNIST samples

(f) GMMN+AE nearest neighbors for MNIST samples



(c) GMMN+AE MNIST samples

(d) GMMN+AE TFD samples

(g) GMMN nearest neighbors for TFD samples

(h) GMMN+AE nearest neighbors for TFD samples

- I could have done better. Just had to wait + bigger network.

Key points:

- Easy to train. Simple implementation.
- Stable training.
- Image quality depends on kernel k.

Outline

Further Topics I

"Dual view": Reproducing Kernel Hilbert Spaces (RKHSs)

- Each point in $\mathcal{H}$ can be seen as a function:
  $f \in \mathcal{H} \iff f(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i)$ for some $\{\alpha_i\}_{i=1}^{m}$, $\{x_i\}_{i=1}^{m}$.

MMD

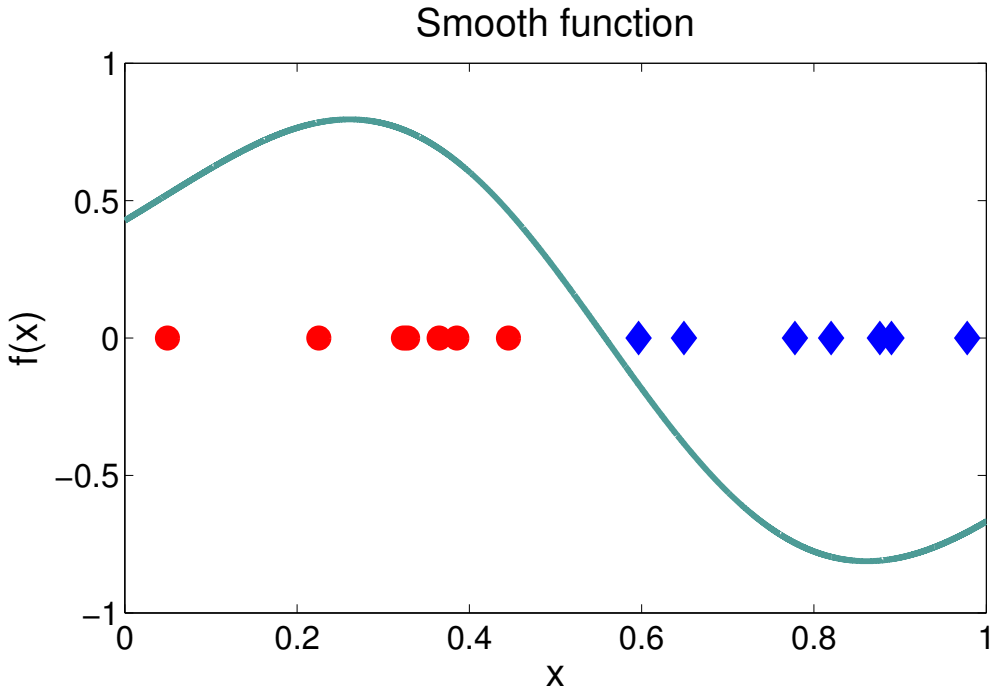
- Associated with $\mathrm{MMD}(p, q)$ is the witness function.
- Unit-norm function in $\mathcal{H}$ that best distinguishes p and q.

Further Topics I

"Dual view": Reproducing Kernel Hilbert Spaces (RKHSs)

- Each point in $\mathcal{H}$ can be seen as a function:
  $f \in \mathcal{H} \iff f(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i)$ for some $\{\alpha_i\}_{i=1}^{m}$, $\{x_i\}_{i=1}^{m}$.

MMD



- Associated with $\mathrm{MMD}(p, q)$ is the **witness function**.
- Unit-norm function in $\mathcal{H}$ that best distinguishes p and q.

Further Topics II

Dependence measure

- Recall X independent of Y iff $p_{xy}(X, Y) = p_x(X)p_y(Y)$.
- $\mathrm{MMD}(p_{xy}, p_x p_y)$ can be used to measure dependence [Gretton et al., 2005].
- Applications: Feature selection, clustering etc.

Others

- Linear-time versions of MMD [Gretton et al., 2012b, Chwialkowski et al., 2015, Jitkrittum et al., 2016].
- Goodness-of-fit test by distance(model , data) [Liu et al., 2016, Chwialkowski et al., 2016, Jitkrittum et al., 2017].
- Gaussian process regression/classification [Rasmussen and Williams, 2006] .

Further Topics II

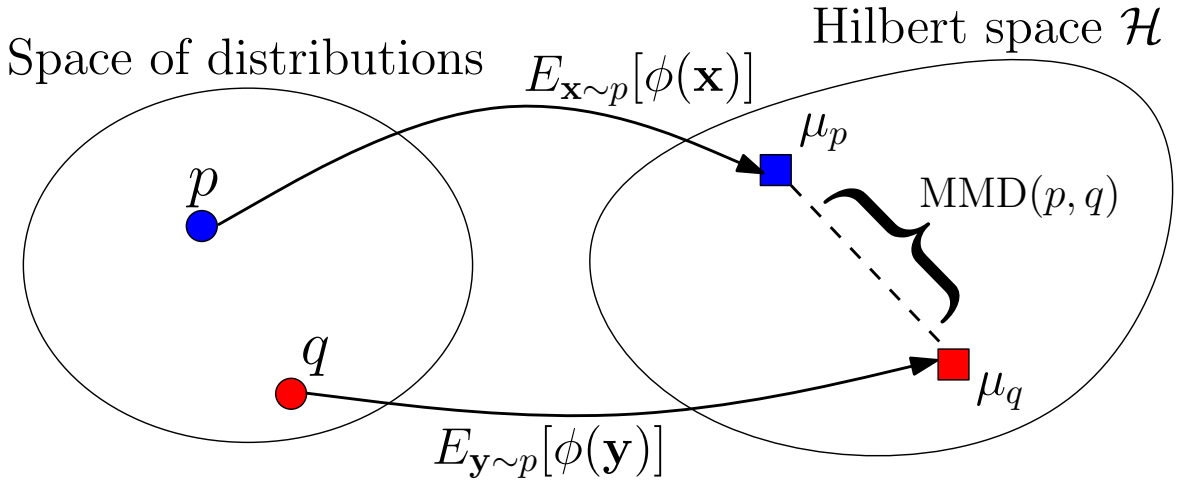
Dependence measure

- Recall X independent of Y iff $p_{xy}(X, Y) = p_x(X)p_y(Y)$.
- MMD$(p_{xy}, p_x p_y)$ can be used to measure dependence [Gretton et al., 2005].
- Applications: Feature selection, clustering etc.

Others

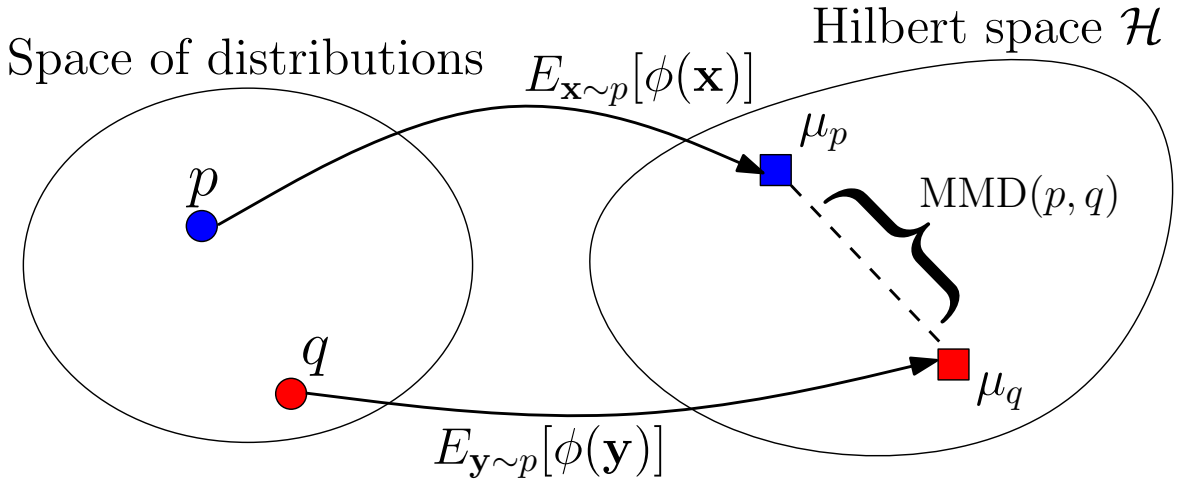
- Linear-time versions of MMD [Gretton et al., 2012b, Chwialkowski et al., 2015, Jitkrittum et al., 2016].
- Goodness-of-fit test by distance(model , data) [Liu et al., 2016, Chwialkowski et al., 2016, Jitkrittum et al., 2017].
- Gaussian process regression/classification [Rasmussen and Williams, 2006] .

Conclusion



- Maximum Mean Discrepancy (MMD) = distance between two distributions
  - "Mean embed" distributions to a high-dimensional space $\mathcal{H}$.
  - Measure the distance in $\mathcal{H}$.
- Characteristic kernel (e.g., Gaussian kernel)
  $\Longrightarrow$ $\mathrm{MMD}(p,q) = 0$ iff $p = q$.
- Two-sample testing with MMD. Consistent test.

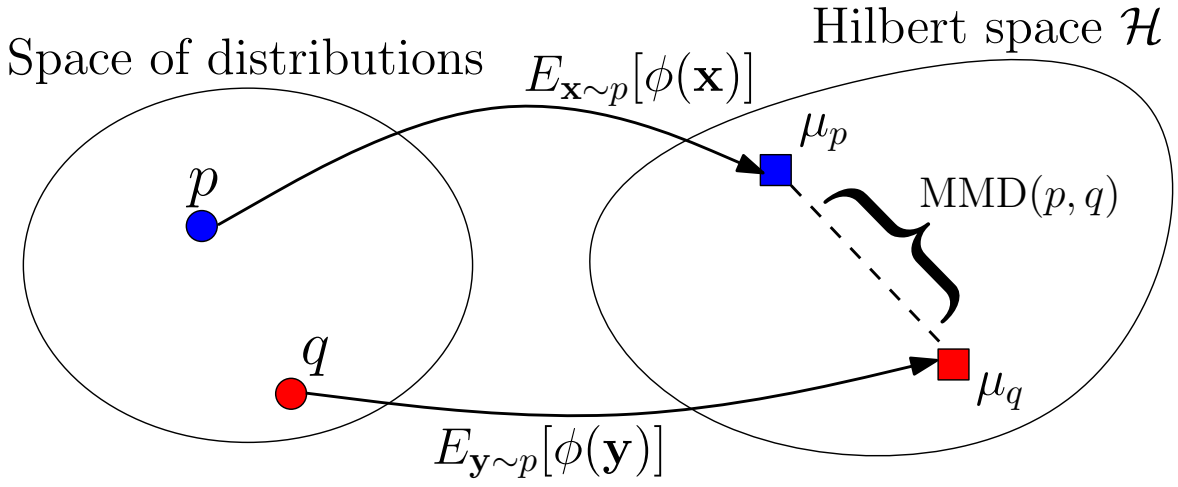
Conclusion



- Maximum Mean Discrepancy (MMD) = distance between two distributions
  - "Mean embed" distributions to a high-dimensional space $\mathcal{H}$.
  - Measure the distance in $\mathcal{H}$.
- Characteristic kernel (e.g., Gaussian kernel)

  $\Longrightarrow$ $\mathrm{MMD}(p,q) = 0$ iff $p = q$.
- Two-sample testing with MMD. Consistent test.

Conclusion



- Maximum Mean Discrepancy (MMD) = distance between two distributions
  - "Mean embed" distributions to a high-dimensional space $\mathcal{H}$.
  - Measure the distance in $\mathcal{H}$.
- Characteristic kernel (e.g., Gaussian kernel)

  $\Longrightarrow$ $\mathrm{MMD}(p,q) = 0$ iff $p = q$.
- Two-sample testing with MMD. Consistent test.

Questions?

Thank you

Wittawat Jitkrittum

wittawat.com

wittawatj@gmail.com

References I


Chwialkowski, K., Ramdas, A., Sejdinovic, D., and Gretton, A. (2015).
Fast two-sample testing with analytic representations of probability measures.
In *NIPS*, pages 1972–1980.


Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A kernel test of goodness of fit.
In *ICML*, pages 2606–2615.


Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008).
Kernel measures of conditional dependence.
In *NIPS*, pages 489–496.

References II


Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2012a).
A kernel two-sample test.
Journal of Machine Learning Research, 13:723–773.


Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005).
Kernel methods for measuring independence.
Journal of Machine Learning Research, 6:2075–2129.


Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b).
Optimal kernel choice for large-scale two-sample tests.
In *NIPS*, pages 1205–1213.

References III


Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. (2016).
Interpretable Distribution Features with Maximum Testing Power.
In *NIPS*, pages 181–189.


Jitkrittum, W., Xu, W., Szabo, Z., Fukumizu, K., and Gretton, A. (2017).
A linear-time kernel goodness-of-fit test.


Li, Y., Swersky, K., and Zemel, R. (2015).
Generative moment matching networks.
In *ICML*, pages 1718–1727.

References IV


Liu, Q., Lee, J., and Jordan, M. (2016).
A kernelized Stein discrepancy for goodness-of-fit tests.
In *ICML*, pages 276–284.


Rasmussen, C. E. and Williams, C. K. I. (2006).
Gaussian Processes for Machine Learning.
MIT Press, Cambridge, MA.


Shawe-Taylor, J. and Cristianini, N. (2004).
Kernel methods for pattern analysis.
Cambridge university press.


Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007).
A Hilbert space embedding for distributions.
In *International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31.

References V


Sriperumbudur, B., Gretton, A., Fukumizu, K., Schoelkopf, B., and Lanckriet, G. (2010).
Hilbert space embeddings and metrics on probability measures.
Journal of Machine Learning Research, 11:1517–1561.


Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2016).
Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy.
arXiv: 1611.04488.


Zaremba, W., Gretton, A., and Blaschko, M. (2013).
B-test: A non-parametric, low variance kernel two-sample test.
In *NIPS*, pages 755–763.