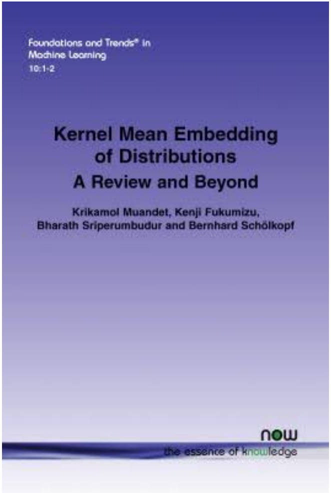
Recent Advances in Hilbert Space Representation of Probability Distributions

Krikamol Muandet
Max Planck Institute for Intelligent Systems
Tübingen, Germany

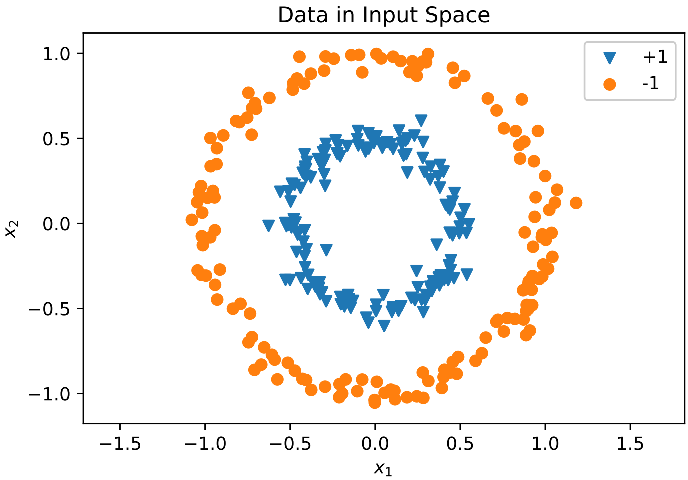
RegML 2020, Genova, Italy
July 1, 2020

Reference



Kernel Mean Embedding of Distributions: A Review and Beyond
KM, *K. Fukumizu, B. Sriperumbudur, and B. Schölkopf*. FnT ML, 2017.

Classification Problem

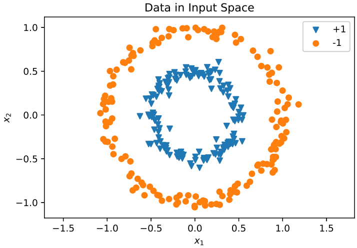


Feature Map

$$\phi \ : \ (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1 x_2)$$

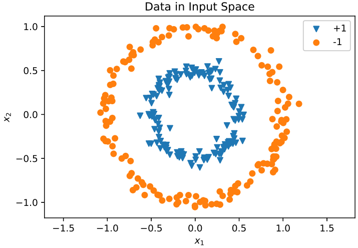
Feature Map

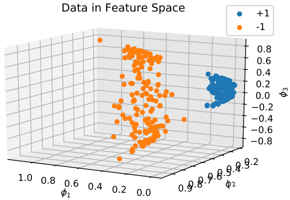
$$\phi \;:\; (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1 x_2)$$



Feature Map

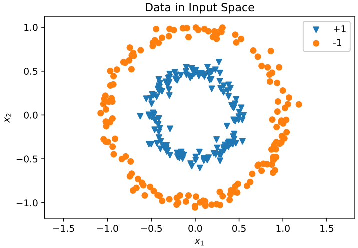
$$\phi \;:\; (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

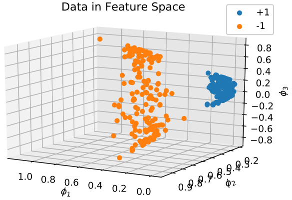




Feature Map

$$\phi \ : \ (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1 x_2)$$

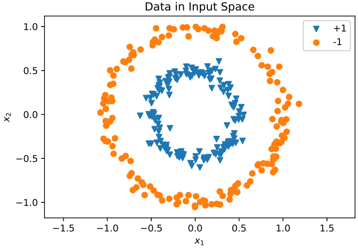


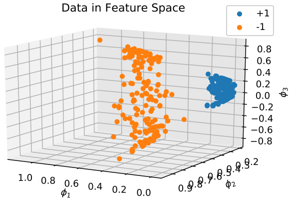


$$\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle_{\mathbb{R}^3} = x_1^2 z_1^2 + x_2^2 z_2^2 + 2(x_1 x_2)(z_1 z_2) = (x_1 z_1 + x_2 z_2)^2 = (\mathbf{x} \cdot \mathbf{z})^2$$

Feature Map

$$\phi \ : \ (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

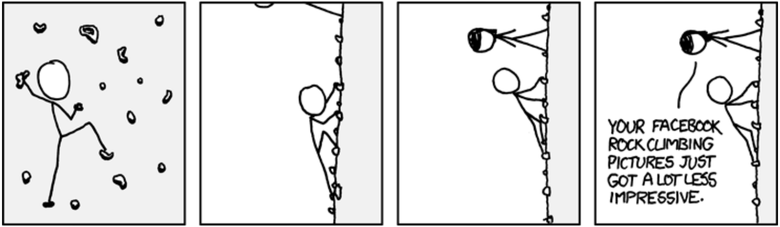




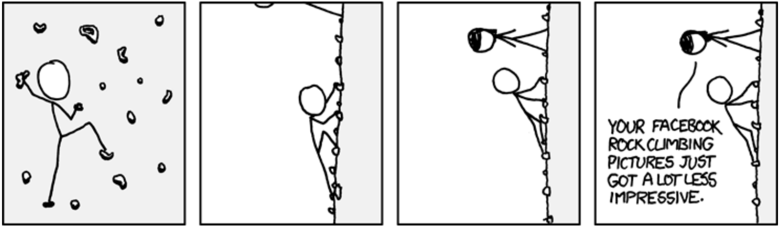
$$\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle_{\mathbb{R}^3} = x_1^2 z_1^2 + x_2^2 z_2^2 + 2(x_1x_2)(z_1z_2) = (x_1z_1 + x_2z_2)^2 = (\mathbf{x} \cdot \mathbf{z})^2$$

Question: How to generalize the idea of **implicit** feature map?

https://xkcd.com/655/



https://xkcd.com/655/



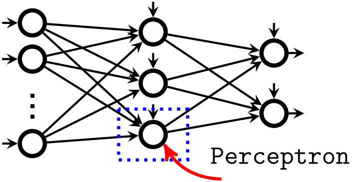
Recipe for ML Problems

1. Collect a data set $D = \{x_1, x_2, \ldots, x_n\}$.
2. Specify or learn a feature map $\phi : \mathcal{X} \to \mathcal{H}$.
3. Apply the feature map $D_\phi = \{\phi(x_1), \phi(x_2), \ldots, \phi(x_n)\}$.
4. Solve the (easier) problem in the feature space $\mathcal{H}$ using D_ϕ.

Representation Learning

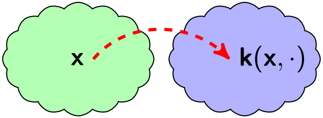
Perceptron[1]**:** $f(\mathbf{x}) = \boldsymbol{w}^\top \mathbf{x} + b$

Explicit Representation



$$f(\mathbf{x}) = \boldsymbol{w}_2^\top \sigma(\boldsymbol{w}_1^\top \mathbf{x} + \boldsymbol{b}_1) + b_2$$

Implicit Representation



$$f(\mathbf{x}) = \boldsymbol{w}^\top \phi(\mathbf{x}) + b$$

[1] Rosenblatt 1958; Minsky and Papert 1969

Kernels

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ we have

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

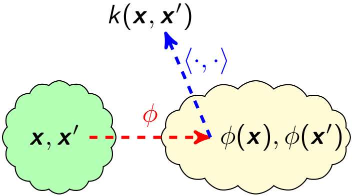
We call ϕ a **feature map** and $\mathcal{H}$ a **feature space** associated with k.

Kernels

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ we have

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

We call ϕ a **feature map** and $\mathcal{H}$ a **feature space** associated with k.



Kernels

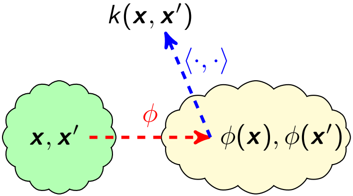
A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ we have

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

We call ϕ a **feature map** and $\mathcal{H}$ a **feature space** associated with k.

Example

1. $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}')^2$ for $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^2$
   - $\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$
   - $\mathcal{H} = \mathbb{R}^3$
2. $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + c)^m$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$
   - $\dim(\mathcal{H}) = \binom{d+m}{m}$
3. $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma\|\mathbf{x} - \mathbf{x}'\|_2^2\right)$
   - $\mathcal{H} = \mathbb{R}^\infty$



Positive Definite Kernels

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called **positive definite** if, for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and all $x_1, \ldots, x_n \in \mathcal{X}$, we have

$$\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) \geq 0, \quad \mathbf{K} := \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$

Equivalently, the **Gram** matrix $\mathbf{K}$ is positive definite.

Positive Definite Kernels

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called **positive definite** if, for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and all $x_1, \ldots, x_n \in \mathcal{X}$, we have

$$\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) \geq 0, \quad \mathbf{K} := \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$

Equivalently, the **Gram** matrix $\mathbf{K}$ is positive definite.

Any **explicit** kernel is positive definite

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

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) = \left\langle \sum_{i=1}^{n} \alpha_i \phi(x_i), \sum_{j=1}^{n} \alpha_j \phi(x_j) \right\rangle_{\mathcal{H}} \geq 0.$$

Positive Definite Kernels

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called **positive definite** if, for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and all $x_1, \ldots, x_n \in \mathcal{X}$, we have

$$\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) \geq 0, \quad \mathbf{K} := \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$

Equivalently, the **Gram** matrix **K** is positive definite.

Any **explicit** kernel is positive definite

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

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) = \left\langle \sum_{i=1}^{n} \alpha_i \phi(x_i), \sum_{j=1}^{n} \alpha_j \phi(x_j) \right\rangle_{\mathcal{H}} \geq 0.$$

Positive definiteness is a **necessary** (and **sufficient**) condition.

Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$.

[2] N. Aronszajn. Theory of reproducing kernels. Transactions of the American Mathematical Society, 68(3):337–404, 1950.

Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$.

1. The space $\mathcal{H}$ is called a **reproducing kernel Hilbert space (RKHS)** over $\mathcal{X}$ if for all $x \in \mathcal{X}$ the Dirac functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by

$$\delta_x(f) := f(x), \qquad f \in \mathcal{H},$$

is continuous.

[2] N. Aronszajn. Theory of reproducing kernels. Transactions of the American Mathematical Society, 68(3):337–404, 1950.

Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$.

1. The space $\mathcal{H}$ is called a **reproducing kernel Hilbert space (RKHS)** over $\mathcal{X}$ if for all $x \in \mathcal{X}$ the Dirac functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by

$$\delta_x(f) := f(x), \qquad f \in \mathcal{H},$$

   is continuous.

2. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **reproducing kernel** of $\mathcal{H}$ if $k(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$ and the **reproducing property**

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

   holds for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$.

[2] N. Aronszajn. Theory of reproducing kernels. Transactions of the American Mathematical Society, 68(3):337–404, 1950.

Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$.

1. The space $\mathcal{H}$ is called a **reproducing kernel Hilbert space (RKHS)** over $\mathcal{X}$ if for all $x \in \mathcal{X}$ the Dirac functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by

$$\delta_x(f) := f(x), \qquad f \in \mathcal{H},$$

   is continuous.

2. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **reproducing kernel** of $\mathcal{H}$ if $k(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$ and the **reproducing property**

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

   holds for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$.

Aronszajn (1950)[2]**:** *"There is a one-to-one correspondance between the reproducing kernel k and the RKHS $\mathcal{H}$"*.

[2] N. Aronszajn. Theory of reproducing kernels. Transactions of the American Mathematical Society, 68(3):337–404, 1950.

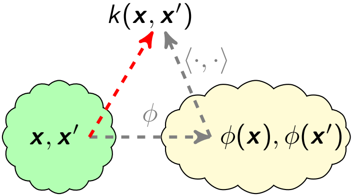
RKHS as Feature Space

Reproducing kernels are kernels

Let $\mathcal{H}$ be a Hilbert space on $\mathcal{X}$ with a **reproducing kernel** k. Then, $\mathcal{H}$ is an RKHS and is also a feature space of k, where the feature map $\phi : \mathcal{X} \to \mathcal{H}$ is given by

$$\phi(x) = k(\cdot, x).$$

We call ϕ the **canonical feature map**.



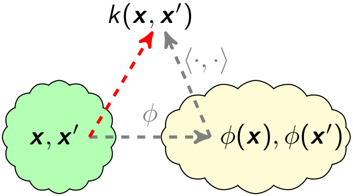
RKHS as Feature Space

Reproducing kernels are kernels

Let $\mathcal{H}$ be a Hilbert space on $\mathcal{X}$ with a **reproducing kernel** k. Then, $\mathcal{H}$ is an RKHS and is also a feature space of k, where the feature map $\phi : \mathcal{X} \to \mathcal{H}$ is given by

$$\phi(x) = k(\cdot, x).$$

We call ϕ the **canonical feature map**.



Proof

We fix an $\mathbf{x}' \in \mathcal{X}$ and write $f := k(\cdot, \mathbf{x}')$. Then, for $\mathbf{x} \in \mathcal{X}$, the reproducing property implies

$$\langle \phi(\mathbf{x}'), \phi(\mathbf{x}) \rangle = \langle k(\cdot, \mathbf{x}'), k(\cdot, \mathbf{x}) \rangle = \langle f, k(\cdot, \mathbf{x}) \rangle = f(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}').$$

RKHS as Feature Space

Universal kernels (Steinwart 2002)

A continuous kernel k on a compact metric space $\mathcal{X}$ is called **universal** if the RKHS $\mathcal{H}$ of k is dense in $C(\mathcal{X})$, i.e., for every function $g \in C(\mathcal{X})$ and all $\varepsilon > 0$ there exist an $f \in \mathcal{H}$ such that

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

RKHS as Feature Space

Universal kernels (Steinwart 2002)

A continuous kernel k on a compact metric space $\mathcal{X}$ is called **universal** if the RKHS $\mathcal{H}$ of k is dense in $C(\mathcal{X})$, i.e., for every function $g \in C(\mathcal{X})$ and all $\varepsilon > 0$ there exist an $f \in \mathcal{H}$ such that

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

Universal approximation theorem (Cybenko 1989)

Given any $\varepsilon > 0$ and $f \in C(\mathcal{X})$, there exist

$$h(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \boldsymbol{\varphi}(\boldsymbol{w}_i^\top \mathbf{x} + \boldsymbol{b}_i)$$

such that $|f(\mathbf{x}) - h(\mathbf{x})| < \varepsilon$ for all $x \in \mathcal{X}$.

Quick Summary

- A **positive definite** kernel $k(x, x')$ defines an **implicit** feature map:

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

Quick Summary

- A **positive definite** kernel $k(x, x')$ defines an **implicit** feature map:

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

- There exists a *unique* **reproducing kernel Hilbert space** (RKHS) $\mathcal{H}$ of functions on $\mathcal{X}$ for which k is a **reproducing kernel**:

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

Quick Summary

- A **positive definite** kernel $k(x, x')$ defines an **implicit** feature map:

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

- There exists a *unique* **reproducing kernel Hilbert space** (RKHS) $\mathcal{H}$ of functions on $\mathcal{X}$ for which k is a **reproducing kernel**:

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

- Implicit representation of **data points**:
  - Support vector machine (SVM)
  - Gaussian process (GP)
  - Neural tangent kernel (NTK)

Quick Summary

- A **positive definite** kernel $k(x, x')$ defines an **implicit** feature map:

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

- There exists a *unique* **reproducing kernel Hilbert space** (RKHS) $\mathcal{H}$ of functions on $\mathcal{X}$ for which k is a **reproducing kernel**:

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

- Implicit representation of **data points**:
  - Support vector machine (SVM)
  - Gaussian process (GP)
  - Neural tangent kernel (NTK)
- Good references on kernel methods.
  - *Support vector machine* (2008), Christmann and Steinwart.
  - *Gaussian process for ML* (2005), Rasmussen and Williams.
  - *Learning with kernels* (1998), Schölkopf and Smola.

Kernel Methods

From Points to Probability Measures

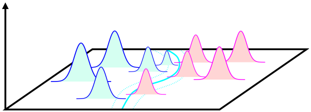
Embedding of Marginal Distributions

Embedding of Conditional Distributions

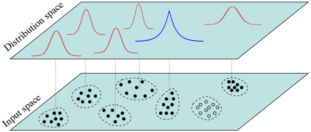
Recent Development

Probability Measures

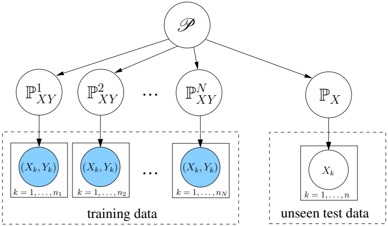
Learning on Distributions/Point Clouds



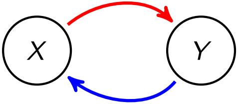
Group Anomaly/OOD Detection

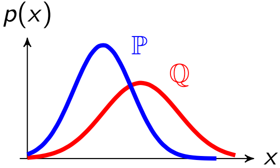


Generalization across Environments

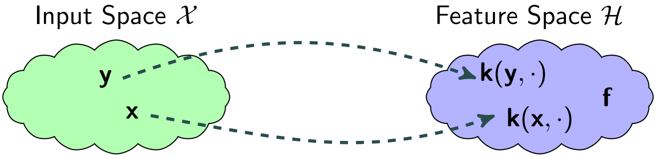


Statistical and Causal Inference

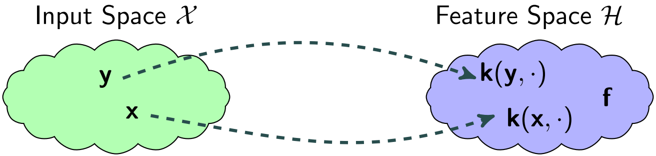


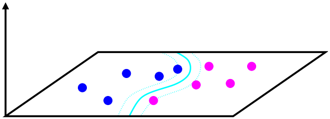


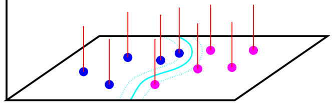
Embedding of Dirac Measures



Embedding of Dirac Measures



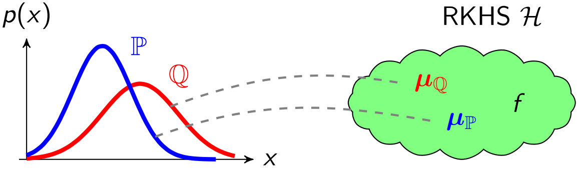




$x \mapsto k(\cdot, x)$

$\delta_x \mapsto \int k(\cdot, z)\, d\delta_x(z) = k(\cdot, x)$

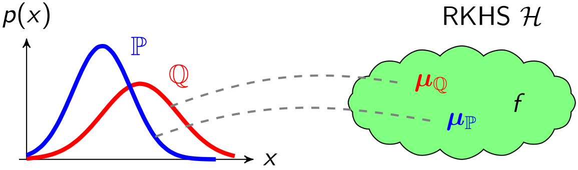
Embedding of Marginal Distributions



Probability measure

Let $\mathbb{P}$ be a probability measure defined on a measurable space $(\mathcal{X}, \Sigma)$ with a σ-algebra Σ.

Embedding of Marginal Distributions



Probability measure

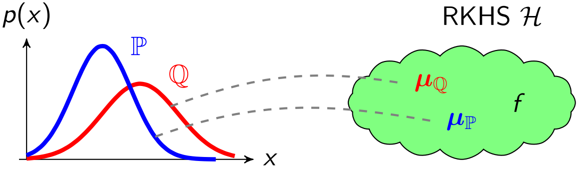
Let $\mathbb{P}$ be a probability measure defined on a measurable space $(\mathcal{X}, \Sigma)$ with a σ-algebra Σ.

Kernel mean embedding

Let $\mathscr{P}$ be a space of all probability measures $\mathbb{P}$. A **kernel mean embedding** is defined by

$$\mu : \mathscr{P} \to \mathcal{H}, \quad \mathbb{P} \mapsto \int k(\cdot, \mathbf{x}) \, \mathrm{d}\mathbb{P}(\mathbf{x}).$$

Embedding of Marginal Distributions



Probability measure

Let $\mathbb{P}$ be a probability measure defined on a measurable space $(\mathcal{X}, \Sigma)$ with a σ-algebra Σ.

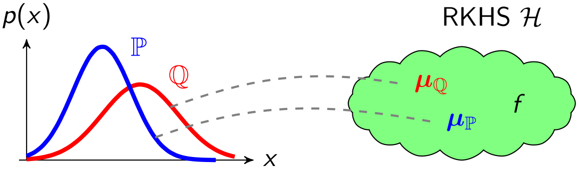
Kernel mean embedding

Let $\mathscr{P}$ be a space of all probability measures $\mathbb{P}$. A **kernel mean embedding** is defined by

$$\mu : \mathscr{P} \to \mathcal{H}, \quad \mathbb{P} \mapsto \int k(\cdot, \mathbf{x}) \, \mathrm{d}\mathbb{P}(\mathbf{x}).$$

Remark: The kernel k is Bochner integrable if it is **bounded**.

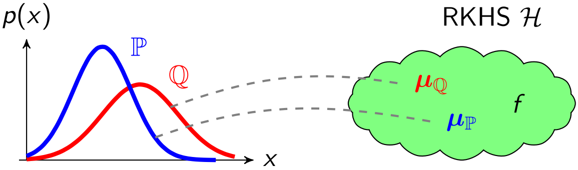
Embedding of Marginal Distributions



- If $\mathbb{E}_{X\sim\mathbb{P}}[\sqrt{k(X,X)}] < \infty$, then for $\mu_{\mathbb{P}} \in \mathcal{H}$ and $f \in \mathcal{H}$,

$$\langle f, \mu_{\mathbb{P}}\rangle = \langle f, \mathbb{E}_{X\sim\mathbb{P}}[k(\cdot, X)]\rangle = \mathbb{E}_{X\sim\mathbb{P}}[\langle f, k(\cdot, X)\rangle] = \mathbb{E}_{X\sim\mathbb{P}}[f(X)].$$

Embedding of Marginal Distributions



- If $\mathbb{E}_{X\sim\mathbb{P}}[\sqrt{k(X,X)}] < \infty$, then for $\mu_{\mathbb{P}} \in \mathcal{H}$ and $f \in \mathcal{H}$,

$$\langle f, \mu_{\mathbb{P}} \rangle = \langle f, \mathbb{E}_{X\sim\mathbb{P}}[k(\cdot, X)] \rangle = \mathbb{E}_{X\sim\mathbb{P}}[\langle f, k(\cdot, X) \rangle] = \mathbb{E}_{X\sim\mathbb{P}}[f(X)].$$

- The kernel k is said to be **characteristic** if the map

$$\mathbb{P} \mapsto \boldsymbol{\mu}_{\mathbb{P}}$$

  is **injective**, i.e., $\|\boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{Q}}\|_{\mathcal{H}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.

Interpretation of Kernel Mean Representation

What properties are captured by $\mu_{\mathbb{P}}$?

- $k(x, x') = \langle x, x' \rangle$ **the first moment of $\mathbb{P}$**
- $k(x, x') = (\langle x, x' \rangle + 1)^p$ **moments of $\mathbb{P}$ up to order $p \in \mathbb{N}$**
- $k(x, x')$ is *universal/characteristic* **all information of $\mathbb{P}$**

Interpretation of Kernel Mean Representation

What properties are captured by $\mu_{\mathbb{P}}$?

- $k(x, x') = \langle x, x' \rangle$ **the first moment of $\mathbb{P}$**
- $k(x, x') = (\langle x, x' \rangle + 1)^p$ **moments of $\mathbb{P}$ up to order $p \in \mathbb{N}$**
- $k(x, x')$ is *universal/characteristic* **all information of $\mathbb{P}$**

Moment-generating function

Consider $k(x, x') = \exp(\langle x, x' \rangle)$. Then, $\boldsymbol{\mu}_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}[e^{\langle X, \cdot \rangle}]$.

Interpretation of Kernel Mean Representation

What properties are captured by $\mu_{\mathbb{P}}$?

- $k(x, x') = \langle x, x' \rangle$ **the first moment of $\mathbb{P}$**
- $k(x, x') = (\langle x, x' \rangle + 1)^p$ **moments of $\mathbb{P}$ up to order $p \in \mathbb{N}$**
- $k(x, x')$ is *universal/characteristic* **all information of $\mathbb{P}$**

Moment-generating function

Consider $k(x, x') = \exp(\langle x, x' \rangle)$. Then, $\boldsymbol{\mu}_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}[e^{\langle X, \cdot \rangle}]$.

Characteristic function

If $k(x, y) = \psi(x - y)$ where ψ is a positive definite function, then

$$\boldsymbol{\mu}_{\mathbb{P}}(y) = \int \psi(x - y) \, \mathrm{d}\mathbb{P}(x) = \Lambda_k \cdot \varphi_{\mathbb{P}}$$

for positive finite measure Λ_k.

Characteristic Kernels

- **All universal kernels are characteristic**, but characteristic kernels may not be universal.

Characteristic Kernels

- **All universal kernels are characteristic**, but characteristic kernels may not be universal.
- Important characterizations:
  - Discrete kernel on discrete space
  - Shift-invariant kernels on $\mathbb{R}^d$ whose Fourier transform has full support.
  - Integrally strictly positive definite (ISPD) kernels
  - Characteristic kernels on groups

Characteristic Kernels

- **All universal kernels are characteristic**, but characteristic kernels may not be universal.
- Important characterizations:
  - Discrete kernel on discrete space
  - Shift-invariant kernels on $\mathbb{R}^d$ whose Fourier transform has full support.
  - Integrally strictly positive definite (ISPD) kernels
  - Characteristic kernels on groups
- Examples of characteristic kernels:

Gaussian RBF kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\sigma^2}\right)$$

Laplacian kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_1}{\sigma}\right)$$

Characteristic Kernels

- **All universal kernels are characteristic**, but characteristic kernels may not be universal.
- Important characterizations:
  - Discrete kernel on discrete space
  - Shift-invariant kernels on $\mathbb{R}^d$ whose Fourier transform has full support.
  - Integrally strictly positive definite (ISPD) kernels
  - Characteristic kernels on groups
- Examples of characteristic kernels:

Gaussian RBF kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\sigma^2}\right)$$

Laplacian kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_1}{\sigma}\right)$$

- **Kernel choice** vs **parametric assumption**
  - Parametric assumption is susceptible to **model misspecification**.
  - But the choice of kernel matters in practice.
  - We can optimize the kernel to maximize the performance of the downstream tasks.

Kernel Mean Estimation

- Given an i.i.d. sample $x_1, x_2, \ldots, x_n$ from $\mathbb{P}$, we can estimate $\boldsymbol{\mu}_{\mathbb{P}}$ by

$$\hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot) \in \mathcal{H}, \qquad \widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}.$$

[3] Tolstikhin et al. *Minimax Estimation of Kernel Mean Embeddings*. JMLR, 2017.
[4] Muandet et al. *Kernel Mean Shrinkage Estimators*. JMLR, 2016.

Kernel Mean Estimation

- Given an i.i.d. sample $x_1, x_2, \ldots, x_n$ from $\mathbb{P}$, we can estimate $\boldsymbol{\mu}_{\mathbb{P}}$ by

$$\hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot) \in \mathcal{H}, \qquad \widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}.$$

- For each $f \in \mathcal{H}$, we have $\mathbb{E}_{X\sim\widehat{\mathbb{P}}}[f(X)] = \langle f, \hat{\boldsymbol{\mu}}_{\mathbb{P}} \rangle$.

[3] Tolstikhin et al. *Minimax Estimation of Kernel Mean Embeddings*. JMLR, 2017.
[4] Muandet et al. *Kernel Mean Shrinkage Estimators*. JMLR, 2016.

Kernel Mean Estimation

- Given an i.i.d. sample $x_1, x_2, \ldots, x_n$ from $\mathbb{P}$, we can estimate $\boldsymbol{\mu}_{\mathbb{P}}$ by

$$\hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot) \in \mathcal{H}, \qquad \widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}.$$

- For each $f \in \mathcal{H}$, we have $\mathbb{E}_{X\sim\widehat{\mathbb{P}}}[f(X)] = \langle f, \hat{\boldsymbol{\mu}}_{\mathbb{P}}\rangle$.
- **Consistency**: with probability at least $1-\delta$,

$$\|\hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}}\|_{\mathcal{H}} \leq 2\sqrt{\frac{\mathbb{E}_{X\sim\mathbb{P}}[k(X,X)]}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

[3] Tolstikhin et al. *Minimax Estimation of Kernel Mean Embeddings*. JMLR, 2017.
[4] Muandet et al. *Kernel Mean Shrinkage Estimators*. JMLR, 2016.

Kernel Mean Estimation

- Given an i.i.d. sample $x_1, x_2, \dots, x_n$ from $\mathbb{P}$, we can estimate $\boldsymbol{\mu}_{\mathbb{P}}$ by

$$\hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot) \in \mathcal{H}, \qquad \widehat{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

- For each $f \in \mathcal{H}$, we have $\mathbb{E}_{X \sim \widehat{\mathbb{P}}}[f(X)] = \langle f, \hat{\boldsymbol{\mu}}_{\mathbb{P}} \rangle$.
- **Consistency**: with probability at least $1 - \delta$,

$$\|\hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}}\|_{\mathcal{H}} \leq 2\sqrt{\frac{\mathbb{E}_{X \sim \mathbb{P}}[k(X, X)]}{n}} + \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}.$$

- The rate $O_p(n^{-1/2})$ was shown to be **minimax optimal.**[3]

[3] Tolstikhin et al. *Minimax Estimation of Kernel Mean Embeddings*. JMLR, 2017.
[4] Muandet et al. *Kernel Mean Shrinkage Estimators*. JMLR, 2016.

Kernel Mean Estimation

- Given an i.i.d. sample $x_1, x_2, \dots, x_n$ from $\mathbb{P}$, we can estimate $\boldsymbol{\mu}_{\mathbb{P}}$ by

$$\hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^n k(x_i, \cdot) \in \mathcal{H}, \qquad \widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}.$$

- For each $f \in \mathcal{H}$, we have $\mathbb{E}_{X\sim\widehat{\mathbb{P}}}[f(X)] = \langle f, \hat{\boldsymbol{\mu}}_{\mathbb{P}}\rangle$.
- **Consistency**: with probability at least $1-\delta$,

$$\|\hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}}\|_{\mathcal{H}} \leq 2\sqrt{\frac{\mathbb{E}_{X\sim\mathbb{P}}[k(X,X)]}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

- The rate $O_p(n^{-1/2})$ was shown to be **minimax optimal**.[3]
- Similar to James-Stein estimators, we can improve an estimation by shrinkage estimators:[4]

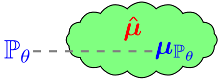
$$\hat{\boldsymbol{\mu}}_{\alpha} := \alpha f^* + (1-\alpha)\hat{\boldsymbol{\mu}}_{\mathbb{P}}, \quad f^* \in \mathcal{H}.$$

[3] Tolstikhin et al. *Minimax Estimation of Kernel Mean Embeddings*. JMLR, 2017.
[4] Muandet et al. *Kernel Mean Shrinkage Estimators*. JMLR, 2016.

Recovering Samples/Distributions

- An approximate pre-image problem

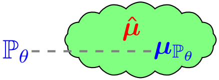
$$\theta^* = \arg\min_{\theta} \; \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\mathbb{P}_\theta}\|_{\mathcal{H}}^2.$$



Recovering Samples/Distributions

- An approximate pre-image problem

$$\theta^* = \arg\min_{\theta} \; \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\mathbb{P}_\theta}\|^2_{\mathcal{H}}.$$



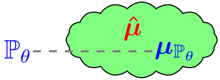
- The distribution $\mathbb{P}_\theta$ is assumed to be in a certain class

$$\mathbb{P}_\theta(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x : \mu_k, \Sigma_k), \quad \sum_{k=1}^{K} \pi_k = 1.$$

Recovering Samples/Distributions

- An approximate pre-image problem

$$\theta^* = \arg\min_{\theta} \ \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\mathbb{P}_\theta}\|^2_{\mathcal{H}}.$$



- The distribution $\mathbb{P}_\theta$ is assumed to be in a certain class

$$\mathbb{P}_\theta(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x : \mu_k, \Sigma_k), \quad \sum_{k=1}^{K} \pi_k = 1.$$

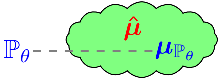
- **Kernel herding** generates deterministic *pseudo-samples* by greedily minimizing the squared error

$$\mathcal{E}_T^2 = \left\| \boldsymbol{\mu}_{\mathbb{P}} - \frac{1}{T} \sum_{t=1}^{T} k(\cdot, \mathbf{x}_t) \right\|^2_{\mathcal{H}}.$$

Recovering Samples/Distributions

- An approximate pre-image problem

$$\theta^* = \arg\min_{\theta} \ \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\mathbb{P}_\theta}\|^2_{\mathcal{H}}.$$



- The distribution $\mathbb{P}_\theta$ is assumed to be in a certain class

$$\mathbb{P}_\theta(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x : \mu_k, \Sigma_k), \quad \sum_{k=1}^{K} \pi_k = 1.$$

- **Kernel herding** generates deterministic *pseudo-samples* by greedily minimizing the squared error

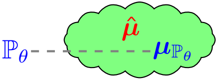
$$\mathcal{E}_T^2 = \left\| \boldsymbol{\mu}_{\mathbb{P}} - \frac{1}{T} \sum_{t=1}^{T} k(\cdot, \mathbf{x}_t) \right\|^2_{\mathcal{H}}.$$

- **Negative autocorrelation**: $O(1/T)$ rate of convergence.

Recovering Samples/Distributions

- An approximate pre-image problem

$$\theta^* = \arg\min_{\theta} \ \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\mathbb{P}_\theta}\|^2_{\mathcal{H}}.$$



- The distribution $\mathbb{P}_\theta$ is assumed to be in a certain class

$$\mathbb{P}_\theta(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x : \mu_k, \Sigma_k), \quad \sum_{k=1}^{K} \pi_k = 1.$$

- **Kernel herding** generates deterministic *pseudo-samples* by greedily minimizing the squared error

$$\mathcal{E}_T^2 = \left\| \boldsymbol{\mu}_{\mathbb{P}} - \frac{1}{T} \sum_{t=1}^{T} k(\cdot, \mathbf{x}_t) \right\|^2_{\mathcal{H}}.$$

- **Negative autocorrelation**: $O(1/T)$ rate of convergence.
- Deep generative models (see the following slides).

Quick Summary

- A kernel mean embedding of distribution $\mathbb{P}$

$$\boldsymbol{\mu}_{\mathbb{P}} := \int k(\cdot, \mathbf{x}) \, \mathrm{d}\mathbb{P}(\mathbf{x}), \qquad \hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot).$$

Quick Summary

- A kernel mean embedding of distribution $\mathbb{P}$

$$\boldsymbol{\mu}_{\mathbb{P}} := \int k(\cdot, \mathbf{x})\,\mathrm{d}\mathbb{P}(\mathbf{x}), \qquad \hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot).$$

- If k is **characteristic**, $\boldsymbol{\mu}_{\mathbb{P}}$ captures all information about $\mathbb{P}$.

Quick Summary

- A kernel mean embedding of distribution $\mathbb{P}$

$$\boldsymbol{\mu}_{\mathbb{P}} := \int k(\cdot, \mathbf{x}) \, \mathrm{d}\mathbb{P}(\mathbf{x}), \qquad \hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot).$$

- If k is **characteristic**, $\boldsymbol{\mu}_{\mathbb{P}}$ captures all information about $\mathbb{P}$.
- All **universal** kernels are characteristic, but not vice versa.

Quick Summary

- A kernel mean embedding of distribution $\mathbb{P}$

$$\boldsymbol{\mu}_{\mathbb{P}} := \int k(\cdot, \mathbf{x}) \,\mathrm{d}\mathbb{P}(\mathbf{x}), \qquad \hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot).$$

- If k is **characteristic**, $\boldsymbol{\mu}_{\mathbb{P}}$ captures all information about $\mathbb{P}$.
- All **universal** kernels are characteristic, but not vice versa.
- The empirical $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ requires **no parametric assumption** about $\mathbb{P}$.

Quick Summary

- A kernel mean embedding of distribution $\mathbb{P}$

$$\boldsymbol{\mu}_{\mathbb{P}} := \int k(\cdot, \mathbf{x})\, \mathrm{d}\mathbb{P}(\mathbf{x}), \qquad \hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot).$$

- If k is **characteristic**, $\boldsymbol{\mu}_{\mathbb{P}}$ captures all information about $\mathbb{P}$.
- All **universal** kernels are characteristic, but not vice versa.
- The empirical $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ requires **no parametric assumption** about $\mathbb{P}$.
- It can be estimated consistently, i.e., with probability at least $1-\delta$,

$$\|\hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}}\|_{\mathcal{H}} \leq 2\sqrt{\frac{\mathbb{E}_{X\sim\mathbb{P}}[k(X,X)]}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

Quick Summary

- A kernel mean embedding of distribution $\mathbb{P}$

$$\boldsymbol{\mu}_{\mathbb{P}} := \int k(\cdot, \mathbf{x}) \, \mathrm{d}\mathbb{P}(\mathbf{x}), \qquad \hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot).$$

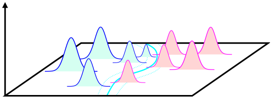
- If k is **characteristic**, $\boldsymbol{\mu}_{\mathbb{P}}$ captures all information about $\mathbb{P}$.
- All **universal** kernels are characteristic, but not vice versa.
- The empirical $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ requires **no parametric assumption** about $\mathbb{P}$.
- It can be estimated consistently, i.e., with probability at least $1 - \delta$,

$$\|\hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}}\|_{\mathcal{H}} \leq 2\sqrt{\frac{\mathbb{E}_{X\sim\mathbb{P}}[k(X,X)]}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

- Given the embedding $\hat{\boldsymbol{\mu}}$, it is possible to reconstruct the distribution or generate samples from it.

Application: High-Level Generalization

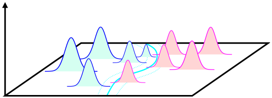
Learning from Distributions



KM., Fukumizu, Dinuzzo, Schölkopf. NIPS 2012.

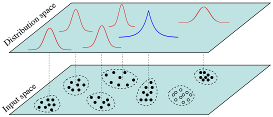
Application: High-Level Generalization

Learning from Distributions



KM., Fukumizu, Dinuzzo, Schölkopf. NIPS 2012.

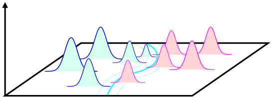
Group Anomaly Detection



KM. and Schölkopf, UAI 2013.

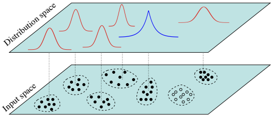
Application: High-Level Generalization

Learning from Distributions



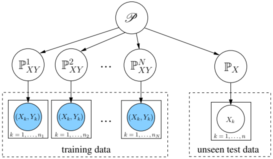
KM., Fukumizu, Dinuzzo, Schölkopf. NIPS 2012.

Group Anomaly Detection



KM. and Schölkopf, UAI 2013.

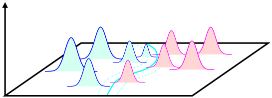
Domain Generalization



KM. et al. ICML 2013; Zhang, **KM.** et al. ICML 2013

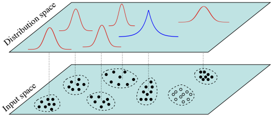
Application: High-Level Generalization

Learning from Distributions



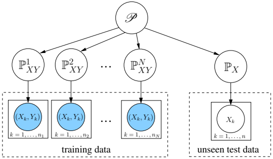
KM., Fukumizu, Dinuzzo, Schölkopf. NIPS 2012.

Group Anomaly Detection



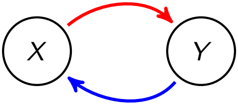
KM. and Schölkopf, UAI 2013.

Domain Generalization



KM. et al. ICML 2013; Zhang, KM. et al. ICML 2013

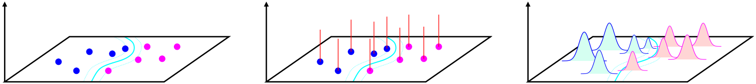
Cause-Effect Inference



Lopez-Paz, KM. et al. JMLR 2015, ICML 2015.

Support Measure Machine (SMM)

KM, K. Fukumizu, F. Dinuzzo, and B. Schölkopf (NeurIPS2012)

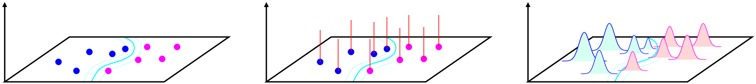


$x \mapsto k(\cdot, x)$ $\qquad$ $\delta_x \mapsto \int k(\cdot, z) d\delta_x(z)$ $\qquad$ $\mathbb{P} \mapsto \int k(\cdot, z) d\mathbb{P}(z)$

Training data: $(\mathbb{P}_1, y_1), (\mathbb{P}_2, y_2), \ldots, (\mathbb{P}_n, y_n) \sim \mathscr{P} \times \mathcal{Y}$

Support Measure Machine (SMM)

KM, K. Fukumizu, F. Dinuzzo, and B. Schölkopf (NeurIPS2012)



$x \mapsto k(\cdot, x)$ $\qquad$ $\delta_x \mapsto \int k(\cdot, z) d\delta_x(z)$ $\qquad$ $\mathbb{P} \mapsto \int k(\cdot, z) d\mathbb{P}(z)$

Training data: $(\mathbb{P}_1, y_1), (\mathbb{P}_2, y_2), \ldots, (\mathbb{P}_n, y_n) \sim \mathscr{P} \times \mathcal{Y}$

Theorem (Distributional representer theorem)

Under technical assumptions on $\Omega : [0, +\infty) \to \mathbb{R}$, *and a loss function* $\ell : (\mathcal{P} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{+\infty\}$, *any* $f \in \mathcal{H}$ *minimizing*

$$\ell\left(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], \ldots, \mathbb{P}_m, y_m, \mathbb{E}_{\mathbb{P}_m}[f]\right) + \Omega\left(\|f\|_{\mathcal{H}}\right)$$

admits a representation of the form

$$f = \sum_{i=1}^{m} \alpha_i \mathbb{E}_{x \sim \mathbb{P}_i}[k(x, \cdot)] = \sum_{i=1}^{m} \alpha_i \mu_{\mathbb{P}_i}.$$

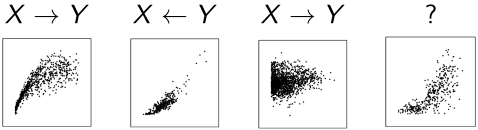
Supervised Learning on Point Clouds

Training set $(S_1, y_1), \ldots, (S_n, y_n)$ with $S_i = \{x_j^{(i)}\} \sim \mathbb{P}_i(X)$.

Supervised Learning on Point Clouds

Training set $(S_1, y_1), \ldots, (S_n, y_n)$ with $S_i = \{x_j^{(i)}\} \sim \mathbb{P}_i(X)$.

Causal Prediction



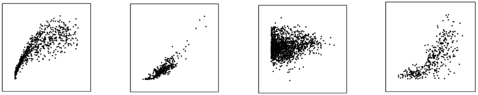
Lopez-Paz, **KM.**, B. Schölkopf, I. Tolstikhin. JMLR 2015, ICML 2015.

Supervised Learning on Point Clouds

Training set $(S_1, y_1), \ldots, (S_n, y_n)$ with $S_i = \{x_j^{(i)}\} \sim \mathbb{P}_i(X)$.

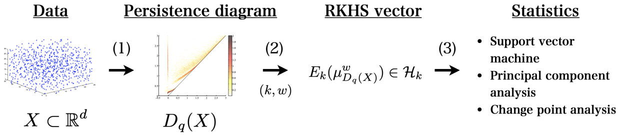
Causal Prediction

$X \to Y$ $\quad X \leftarrow Y$ $\quad X \to Y$ $\quad ?$



Lopez-Paz, **KM.**, B. Schölkopf, I. Tolstikhin. JMLR 2015, ICML 2015.

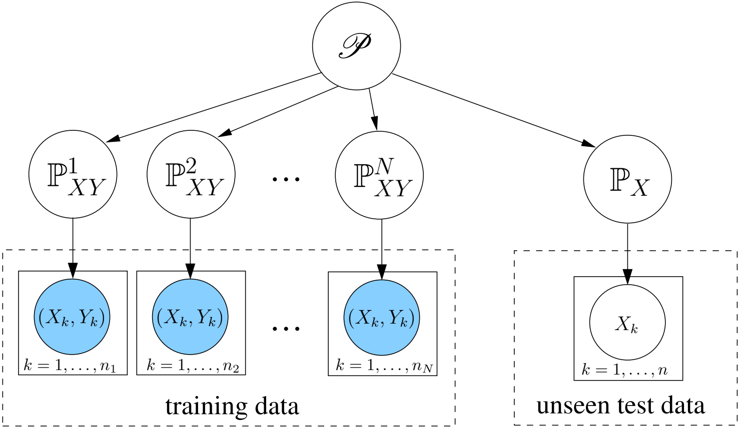
Topological Data Analysis



G. Kusano, K. Fukumizu, and Y. Hiraoka. JMLR2018

Domain Generalization

Blanchard et al., NeurIPS2012; **KM**, D. Balduzzi, B. Schölkopf, ICML2013



$$K((\mathbb{P}_i, x), (\mathbb{P}_j, \tilde{x})) = k_1(\mathbb{P}_i, \mathbb{P}_j)k_2(x, \tilde{x}) = k_1(\boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j})k_2(x, \tilde{x})$$

Comparing Distributions

- **Maximum mean discrepancy (MMD)** corresponds to the RKHS distance between mean embeddings:

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q},\mathcal{H}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}^2 = \|\mu_{\mathbb{P}}\|_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}} + \|\mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

Comparing Distributions

- **Maximum mean discrepancy (MMD)** corresponds to the RKHS distance between mean embeddings:

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}, \mathcal{H}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}} = \|\mu_{\mathbb{P}}\|_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} + \|\mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

- MMD is an **integral probability metric (IPM)**:

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}, \mathcal{H}) := \sup_{h \in \mathcal{H}, \|h\| \leq 1} \left| \int h(x)\, \mathrm{d}\mathbb{P}(x) - \int h(x)\, \mathrm{d}\mathbb{Q}(x) \right|.$$

Comparing Distributions

- **Maximum mean discrepancy (MMD)** corresponds to the RKHS distance between mean embeddings:

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}, \mathcal{H}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}} = \|\mu_{\mathbb{P}}\|_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} + \|\mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

- MMD is an **integral probability metric (IPM)**:

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}, \mathcal{H}) := \sup_{h \in \mathcal{H}, \|h\| \leq 1} \left| \int h(x)\, \mathrm{d}\mathbb{P}(x) - \int h(x)\, \mathrm{d}\mathbb{Q}(x) \right|.$$

- If k is **characteristic**, then $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.

Comparing Distributions

- **Maximum mean discrepancy (MMD)** corresponds to the RKHS distance between mean embeddings:

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q},\mathcal{H}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}} = \|\mu_{\mathbb{P}}\|_{\mathcal{H}} - 2\langle\mu_{\mathbb{P}},\mu_{\mathbb{Q}}\rangle_{\mathcal{H}} + \|\mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

- MMD is an **integral probability metric (IPM)**:

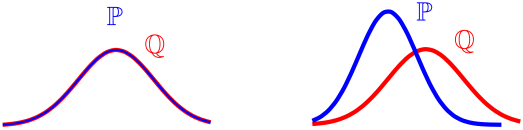
$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q},\mathcal{H}) := \sup_{h\in\mathcal{H},\|h\|\leq 1}\left|\int h(x)\,\mathrm{d}\mathbb{P}(x) - \int h(x)\,\mathrm{d}\mathbb{Q}(x)\right|.$$

- If k is **characteristic**, then $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.
- Given $\{\mathbf{x}_i\}_{i=1}^n \sim \mathbb{P}$ and $\{\mathbf{y}_j\}_{j=1}^m \sim \mathbb{Q}$, the empirical MMD is

$$\widehat{\mathrm{MMD}}^2_u(\mathbb{P},\mathbb{Q},\mathcal{H}) = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n k(\mathbf{x}_i,\mathbf{x}_j) + \frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j\neq i}^m k(\mathbf{y}_i,\mathbf{y}_j) - \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m k(\mathbf{x}_i,\mathbf{y}_j).$$

Kernel Two-Sample Testing

Gretton et al., JMLR2012

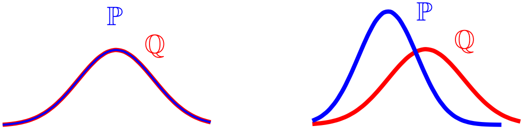


Question: Given $\{\boldsymbol{x}_i\}_{i=1}^n \sim \mathbb{P}$ and $\{\boldsymbol{y}_j\}_{j=1}^n \sim \mathbb{Q}$, check if $\mathbb{P} = \mathbb{Q}$.

$$H_0 : \mathbb{P} = \mathbb{Q}, \quad H_1 : \mathbb{P} \neq \mathbb{Q}$$

Kernel Two-Sample Testing

Gretton et al., JMLR2012



Question: Given $\{\boldsymbol{x}_i\}_{i=1}^n \sim \mathbb{P}$ and $\{\boldsymbol{y}_j\}_{j=1}^n \sim \mathbb{Q}$, check if $\mathbb{P} = \mathbb{Q}$.

$$H_0 : \mathbb{P} = \mathbb{Q}, \quad H_1 : \mathbb{P} \neq \mathbb{Q}$$

- MMD test statistic:

$$\begin{aligned} t^2 &= \widehat{\mathrm{MMD}}_u^2(\mathbb{P}, \mathbb{Q}, \mathcal{H}) \\ &= \frac{1}{n(n-1)} \sum_{1 \le i \neq j \le n} h((x_i, y_i), (x_j, y_j)) \end{aligned}$$

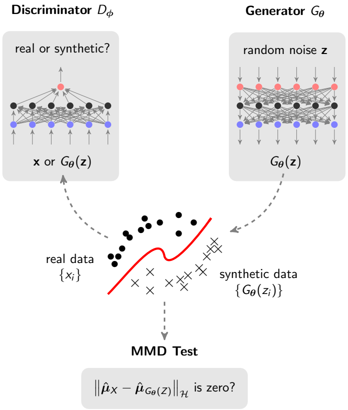
where $h((x_i, y_i), (x_j, y_j)) = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$.

Generative Adversarial Networks

Learn a deep generative model G via a minimax optimization

$$\min_G \max_D \ \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))]$$

where D is a discriminator and $z \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.



Generative Moment Matching Network

- The GAN aims to match two distributions $\mathbb{P}(X)$ and $\mathbb{G}_{\boldsymbol{\theta}}$.

Generative Moment Matching Network

- The GAN aims to match two distributions $\mathbb{P}(X)$ and $\mathbb{G}_{\boldsymbol{\theta}}$.
- Generative moment matching network (GMMN) proposed by Dziugaite et al. (2015) and Li et al. (2015) considers

$$\begin{aligned}
\min_{\boldsymbol{\theta}} \left\|\boldsymbol{\mu}_X - \boldsymbol{\mu}_{G_{\boldsymbol{\theta}}(Z)}\right\|_{\mathcal{H}}^2 &= \min_{\boldsymbol{\theta}} \left\|\int \phi(X)\,\mathrm{d}\mathbb{P}(X) - \int \phi(\tilde{X})\,\mathrm{d}\mathbb{G}_{\boldsymbol{\theta}}(\tilde{X})\right\|_{\mathcal{H}}^2 \\
&= \min_{\boldsymbol{\theta}} \left\{ \sup_{h \in \mathcal{H}, \|h\| \leq 1} \left| \int h\,\mathrm{d}\mathbb{P} - \int h\,\mathrm{d}\mathbb{G}_{\boldsymbol{\theta}} \right| \right\}
\end{aligned}$$

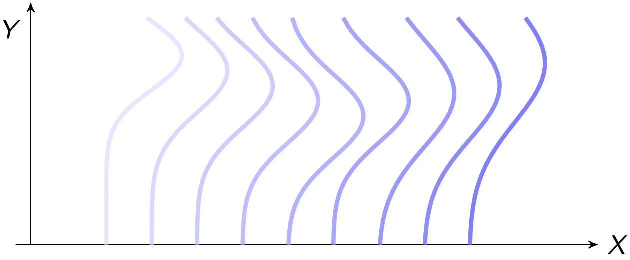
Generative Moment Matching Network

- The GAN aims to match two distributions $\mathbb{P}(X)$ and $\mathbb{G}_{\boldsymbol{\theta}}$.
- Generative moment matching network (GMMN) proposed by Dziugaite et al. (2015) and Li et al. (2015) considers

$$\begin{aligned}\min_{\boldsymbol{\theta}} \left\|\boldsymbol{\mu}_X - \boldsymbol{\mu}_{G_{\boldsymbol{\theta}}(Z)}\right\|_{\mathcal{H}}^2 &= \min_{\boldsymbol{\theta}} \left\|\int \phi(X)\,\mathrm{d}\mathbb{P}(X) - \int \phi(\tilde{X})\,\mathrm{d}\mathbb{G}_{\boldsymbol{\theta}}(\tilde{X})\right\|_{\mathcal{H}}^2 \\ &= \min_{\boldsymbol{\theta}} \left\{ \sup_{h\in\mathcal{H},\|h\|\leq 1} \left| \int h\,\mathrm{d}\mathbb{P} - \int h\,\mathrm{d}\mathbb{G}_{\boldsymbol{\theta}} \right| \right\}\end{aligned}$$

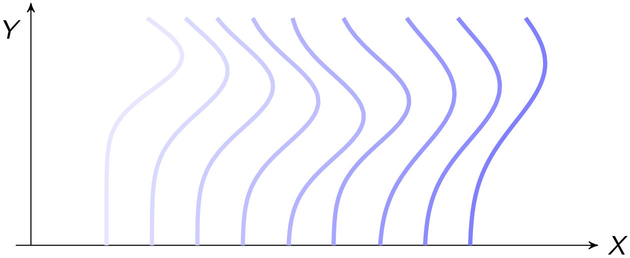
- Many tricks have been proposed to improve the GMMN:
  - Optimized kernels and feature extractors (Sutherland et al., 2017; Li et al., 2017a)
  - Gradient regularization (Binkowski et al., 2018; Arbel et al., 2018)
  - Repulsive loss (Wang et al., 2019)
  - Optimized witness points (Mehrjou et al., 2019)
  - Etc.

Conditional Distribution $\mathbb{P}(Y|X)$



A collection of distributions $\mathscr{P}_Y := \{\mathbb{P}(Y|X = x) : x \in \mathcal{X}\}$.

Conditional Distribution $\mathbb{P}(Y|X)$



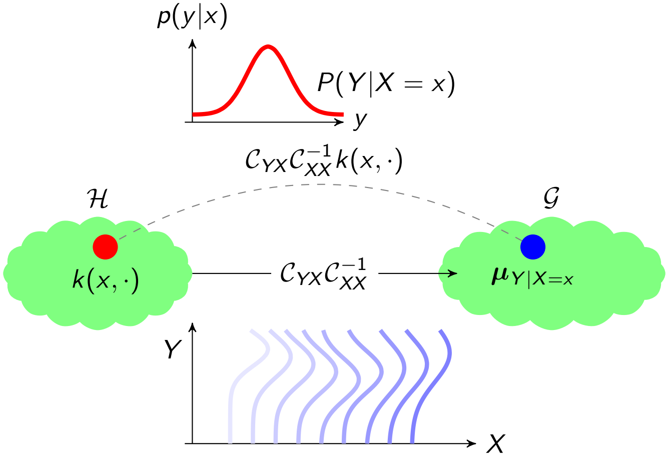
A collection of distributions $\mathscr{P}_Y := \{\mathbb{P}(Y|X=x) : x \in \mathcal{X}\}$.

- For each $x \in \mathcal{X}$, we can define an embedding of $\mathbb{P}(Y|X=x)$ as

$$\boldsymbol{\mu}_{Y|x} := \int_Y \varphi(Y)\, \mathrm{d}\mathbb{P}(Y|X=x) = \mathbb{E}_{Y|x}[\varphi(Y)]$$

  where $\varphi : \mathcal{Y} \to \mathcal{G}$ is a feature map of Y.

Embedding of Conditional Distributions



The conditional mean embedding of $\mathbb{P}(Y \mid X)$ can be defined as

$$\mathcal{U}_{Y|X} : \mathcal{H} \to \mathcal{G}, \qquad \mathcal{U}_{Y|X} := \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$$

Conditional Mean Embedding

- To fully represent $\mathbb{P}(Y|X)$, we need to perform **conditioning** and **conditional expectation**.

Conditional Mean Embedding

- To fully represent $\mathbb{P}(Y|X)$, we need to perform **conditioning** and **conditional expectation**.
- To represent $\mathbb{P}(Y|X=x)$ for $x \in \mathcal{X}$, it follows that

$$\mathbb{E}_{Y|x}[\varphi(Y) \,|\, X = x] = \mathcal{U}_{Y|X} k(x,\cdot) = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} k(x,\cdot) =: \boldsymbol{\mu}_{Y|x}.$$

Conditional Mean Embedding

- To fully represent $\mathbb{P}(Y|X)$, we need to perform **conditioning** and **conditional expectation**.
- To represent $\mathbb{P}(Y|X=x)$ for $x \in \mathcal{X}$, it follows that

$$\mathbb{E}_{Y|x}[\varphi(Y) \,|\, X = x] = \mathcal{U}_{Y|X} k(x,\cdot) = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} k(x,\cdot) =: \boldsymbol{\mu}_{Y|x}.$$

- It follows from the reproducing property of $\mathcal{G}$ that

$$\mathbb{E}_{Y|x}[g(Y) \,|\, X = x] = \langle \boldsymbol{\mu}_{Y|x}, g \rangle_{\mathcal{G}}, \quad \forall g \in \mathcal{G}.$$

Conditional Mean Embedding

- To fully represent $\mathbb{P}(Y|X)$, we need to perform **conditioning** and **conditional expectation**.
- To represent $\mathbb{P}(Y|X = x)$ for $x \in \mathcal{X}$, it follows that

$$\mathbb{E}_{Y|x}[\varphi(Y) \,|\, X = x] = \mathcal{U}_{Y|X} k(x, \cdot) = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} k(x, \cdot) =: \boldsymbol{\mu}_{Y|x}.$$

- It follows from the reproducing property of $\mathcal{G}$ that

$$\mathbb{E}_{Y|x}[g(Y) \,|\, X = x] = \langle \boldsymbol{\mu}_{Y|x}, g \rangle_{\mathcal{G}}, \quad \forall g \in \mathcal{G}.$$

- In an infinite RKHS, $\mathcal{C}_{XX}^{-1}$ does not exists. Hence, we often use

$$\mathcal{U}_{Y|X} := \mathcal{C}_{YX}(\mathcal{C}_{XX} + \varepsilon\mathcal{I})^{-1}.$$

Conditional Mean Embedding

- To fully represent $\mathbb{P}(Y|X)$, we need to perform **conditioning** and **conditional expectation**.
- To represent $\mathbb{P}(Y|X=x)$ for $x \in \mathcal{X}$, it follows that

$$\mathbb{E}_{Y|x}[\varphi(Y) \,|\, X=x] = \mathcal{U}_{Y|X} k(x,\cdot) = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} k(x,\cdot) =: \boldsymbol{\mu}_{Y|x}.$$

- It follows from the reproducing property of $\mathcal{G}$ that

$$\mathbb{E}_{Y|x}[g(Y) \,|\, X=x] = \langle \boldsymbol{\mu}_{Y|x}, g \rangle_{\mathcal{G}}, \quad \forall g \in \mathcal{G}.$$

- In an infinite RKHS, $\mathcal{C}_{XX}^{-1}$ does not exists. Hence, we often use

$$\mathcal{U}_{Y|X} := \mathcal{C}_{YX}(\mathcal{C}_{XX} + \varepsilon\mathcal{I})^{-1}.$$

- Conditional mean estimator

$$\hat{\boldsymbol{\mu}}_{Y|x} = \sum_{i=1}^{n} \beta_i(x)\varphi(y_i), \quad \boldsymbol{\beta}(x) := (\mathbf{K} + n\varepsilon I)^{-1}\mathbf{k}_x.$$

Counterfactual Mean Embedding

KM, Kanagawa, Saengkyongam, Marukatat, JMLR2020 (Accepted)

In economics, social science, and public policy, we need to evaluate the **distributional treatment effect (DTE)**

$$\mathbb{P}_{Y_0^*}(\cdot) - \mathbb{P}_{Y_1^*}(\cdot)$$

where Y_0^* and Y_1^* are **potential outcomes** of a treatment policy T.

Counterfactual Mean Embedding

KM, Kanagawa, Saengkyongam, Marukatat, JMLR2020 (Accepted)

In economics, social science, and public policy, we need to evaluate the **distributional treatment effect (DTE)**

$$\mathbb{P}_{Y_0^*}(\cdot) - \mathbb{P}_{Y_1^*}(\cdot)$$

where Y_0^* and Y_1^* are **potential outcomes** of a treatment policy T.

- We can only observe either $\mathbb{P}_{Y_0^*}$ or $\mathbb{P}_{Y_1^*}$.

Counterfactual Mean Embedding

KM, Kanagawa, Saengkyongam, Marukatat, JMLR2020 (Accepted)

In economics, social science, and public policy, we need to evaluate the **distributional treatment effect (DTE)**

$$\mathbb{P}_{Y_0^*}(\cdot) - \mathbb{P}_{Y_1^*}(\cdot)$$

where Y_0^* and Y_1^* are **potential outcomes** of a treatment policy T.

- We can only observe either $\mathbb{P}_{Y_0^*}$ or $\mathbb{P}_{Y_1^*}$.
- **Counterfactual distribution**

$$\mathbb{P}_{Y\langle 0|1\rangle}(y) = \int \mathbb{P}_{Y_0|X_0}(y|x) \, \mathrm{d}\mathbb{P}_{X_1}(x).$$

Counterfactual Mean Embedding

KM, Kanagawa, Saengkyongam, Marukatat, JMLR2020 (Accepted)

In economics, social science, and public policy, we need to evaluate the **distributional treatment effect (DTE)**

$$\mathbb{P}_{Y_0^*}(\cdot) - \mathbb{P}_{Y_1^*}(\cdot)$$

where Y_0^* and Y_1^* are **potential outcomes** of a treatment policy T.

- We can only observe either $\mathbb{P}_{Y_0^*}$ or $\mathbb{P}_{Y_1^*}$.
- **Counterfactual distribution**

$$\mathbb{P}_{Y\langle 0|1\rangle}(y) = \int \mathbb{P}_{Y_0|X_0}(y|x) \, \mathrm{d}\mathbb{P}_{X_1}(x).$$

- The counterfactual distribution $\mathbb{P}_{Y\langle 0|1\rangle}(y)$ can be estimated using the kernel mean embedding.

Quantum Mean Embedding

PHYSICAL REVIEW RESEARCH **1**, 033159 (2019)

Quantum mean embedding of probability distributions

Jonas M. Kübler,* Krikamol Muandet,† and Bernhard Schölkopf‡
Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany

(Received 15 June 2019; published 9 December 2019)

The kernel mean embedding of probability distributions is commonly used in machine learning as an injective mapping from distributions to functions in an infinite-dimensional Hilbert space. It allows us, for example, to define a distance measure between probability distributions, called the maximum mean discrepancy. In this work, we propose to represent probability distributions in a pure quantum state of a system that is described by an infinite-dimensional Hilbert space and prove that the representation is unique if the corresponding kernel function is c_0 universal. This enables us to work with an explicit representation of the mean embedding, whereas classically one can only work implicitly with an infinite-dimensional Hilbert space through the use of the kernel trick. We show how this explicit representation can speed up methods that rely on inner products of mean embeddings and discuss the theoretical and experimental challenges that need to be solved in order to achieve these speedups.

Quick Summary

- Many applications requires information in $\mathbb{P}(Y|X)$.

Quick Summary

- Many applications requires information in $\mathbb{P}(Y|X)$.
- Hilbert space embedding of $\mathbb{P}(Y|X)$ is **not a single element**, but an **operator** $\mathcal{U}_{Y|X}$ mapping from $\mathcal{H}$ to $\mathcal{G}$:

$$\begin{aligned} \boldsymbol{\mu}_{Y|x} &= \mathcal{U}_{Y|X} k(x,\cdot) = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} k(x,\cdot) \\ \langle \boldsymbol{\mu}_{Y|x}, g \rangle_{\mathcal{G}} &= \mathbb{E}_{Y|x}[g(Y) \,|\, X = x] \end{aligned}$$

Quick Summary

- Many applications requires information in $\mathbb{P}(Y|X)$.
- Hilbert space embedding of $\mathbb{P}(Y|X)$ is **not a single element**, but an **operator** $\mathcal{U}_{Y|X}$ mapping from $\mathcal{H}$ to $\mathcal{G}$:

$$\begin{aligned} \boldsymbol{\mu}_{Y|x} &= \mathcal{U}_{Y|X} k(x,\cdot) = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} k(x,\cdot) \\ \langle \boldsymbol{\mu}_{Y|x}, g\rangle_{\mathcal{G}} &= \mathbb{E}_{Y|x}[g(Y) \,|\, X = x] \end{aligned}$$

- The conditional mean operator

$$\mathcal{U}_{Y|X} := \mathcal{C}_{YX}(\mathcal{C}_{XX} + \varepsilon\mathcal{I})^{-1}, \qquad \widehat{\mathcal{U}}_{Y|X} = \widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \varepsilon\mathcal{I})^{-1}$$

Quick Summary

- Many applications requires information in $\mathbb{P}(Y|X)$.
- Hilbert space embedding of $\mathbb{P}(Y|X)$ is **not a single element**, but an **operator** $\mathcal{U}_{Y|X}$ mapping from $\mathcal{H}$ to $\mathcal{G}$:

$$\begin{aligned} \boldsymbol{\mu}_{Y|x} &= \mathcal{U}_{Y|X} k(x,\cdot) = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} k(x,\cdot) \\ \langle \boldsymbol{\mu}_{Y|x}, g\rangle_{\mathcal{G}} &= \mathbb{E}_{Y|x}[g(Y) \,|\, X = x] \end{aligned}$$

- The conditional mean operator

$$\mathcal{U}_{Y|X} := \mathcal{C}_{YX}(\mathcal{C}_{XX} + \varepsilon\mathcal{I})^{-1}, \qquad \widehat{\mathcal{U}}_{Y|X} = \widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \varepsilon\mathcal{I})^{-1}$$

- Probabilistic inference such as **sum, product, and Bayes rules**, can be performed via the embeddings.

Kernel Methods

From Points to Probability Measures

Embedding of Marginal Distributions

Embedding of Conditional Distributions

Recent Development

Machine Learning in Economics

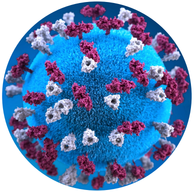
Recommendation

Autonomous Car

Healthcare

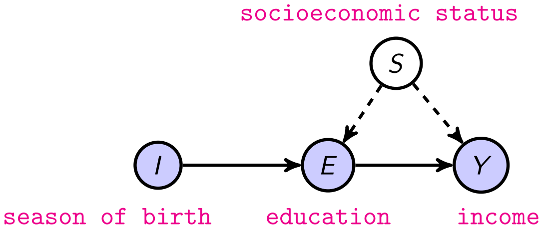
Finance

Law Enforcement

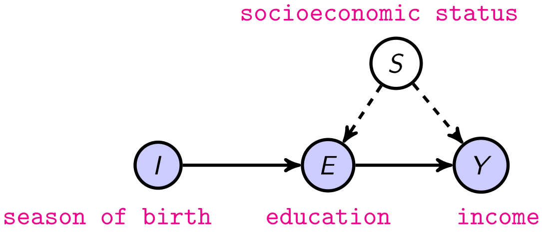


Public Policy

Instrumental Variable Regression



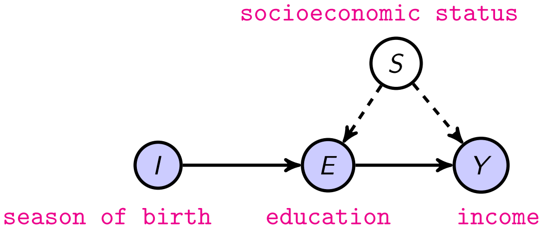
Instrumental Variable Regression



- We aim to estimate a function f from a structural equation model

$$Y = f(E) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid E] \neq 0.$$

Instrumental Variable Regression



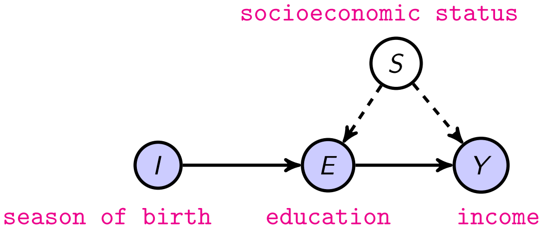
- We aim to estimate a function f from a structural equation model

$$Y = f(E) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid E] \neq 0.$$

- We have an **instrumental** variable I with property $\mathbb{E}[\varepsilon \mid I] = 0$, i.e.,

$$\boxed{\mathbb{E}[Y - f(E) \mid I] = 0}$$

Instrumental Variable Regression



- We aim to estimate a function f from a structural equation model

$$Y = f(E) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid E] \neq 0.$$

- We have an **instrumental** variable I with property $\mathbb{E}[\varepsilon \mid I] = 0$, i.e.,

$$\boxed{\mathbb{E}[Y - f(E) \mid I] = 0}$$

- **Conditional moment restriction (CMR)**: $\mathbb{E}[\psi(Z, \theta) \mid X] = 0$.

$$Z = (E, Y), \quad X = I, \quad \theta = f, \quad \psi(Z; \theta) = Y - f(E).$$

Conditional Moment Restriction (CMR)

Newey (1993), Ai and Chen (2003)

There exists a true parameter $\theta_0 \in \Theta$ that satisfies

$$\mathbb{E}[\boldsymbol{\psi}(Z;\theta_0) \mid X] = \mathbf{0}, \quad P_X - \text{a.s.},$$

where $\boldsymbol{\psi} : \mathcal{Z} \times \Theta \to \mathbb{R}^q$ is a **generalized residual function**.

Conditional Moment Restriction (CMR)

Newey (1993), Ai and Chen (2003)

There exists a true parameter $\theta_0 \in \Theta$ that satisfies

$$\mathbb{E}[\boldsymbol{\psi}(Z;\theta_0) \,|\, X] = \mathbf{0}, \quad P_X - \text{a.s.},$$

where $\boldsymbol{\psi} : \mathcal{Z} \times \Theta \to \mathbb{R}^q$ is a **generalized residual function**.

- The function $\boldsymbol{\psi}$ is known and is problem-dependent, e.g.,

$$\boldsymbol{\psi}(Z;\theta) = Y - f(E), \quad Z = (Y, E), X = I, \theta = f.$$

Conditional Moment Restriction (CMR)

Newey (1993), Ai and Chen (2003)

There exists a true parameter $\theta_0 \in \Theta$ that satisfies

$$\mathbb{E}[\boldsymbol{\psi}(Z;\theta_0) \,|\, X] = \mathbf{0}, \quad P_X - \text{a.s.},$$

where $\boldsymbol{\psi} : \mathcal{Z} \times \Theta \to \mathbb{R}^q$ is a **generalized residual function**.

- The function $\boldsymbol{\psi}$ is known and is problem-dependent, e.g.,

$$\boldsymbol{\psi}(Z;\theta) = Y - f(E), \quad Z = (Y, E), X = I, \theta = f.$$

- The CMR implies **unconditional moment restriction (UMR)**:

$$\mathbb{E}[\boldsymbol{\psi}(Z;\theta_0)^\top f(X)] = 0$$

for any measurable vector-valued function $f : \mathcal{X} \to \mathbb{R}^q$. The function $f(X)$ is often called an **instrument**.

Conditional Moment Restriction (CMR)

Newey (1993), Ai and Chen (2003)

There exists a true parameter $\theta_0 \in \Theta$ that satisfies

$$\mathbb{E}[\boldsymbol{\psi}(Z;\theta_0) \mid X] = \mathbf{0}, \quad P_X - \text{a.s.},$$

where $\boldsymbol{\psi} : \mathcal{Z} \times \Theta \to \mathbb{R}^q$ is a **generalized residual function**.

- The function $\boldsymbol{\psi}$ is known and is problem-dependent, e.g.,

$$\boldsymbol{\psi}(Z;\theta) = Y - f(E), \quad Z = (Y, E), X = I, \theta = f.$$

- The CMR implies **unconditional moment restriction (UMR)**:

$$\mathbb{E}[\boldsymbol{\psi}(Z;\theta_0)^\top f(X)] = 0$$

  for any measurable vector-valued function $f : \mathcal{X} \to \mathbb{R}^q$. The function $f(X)$ is often called an **instrument**.
- Given the instruments $f_1, \dots, f_m$, one can use the **generalized method of moment (GMM)** to learn the parameter θ.

Maximum Moment Restriction (MMR)

KM, W. Jitkrittum, J. Kübler, UAI2020

Let $\mathscr{F}$ be a space of instruments $f(x)$.

$$\underbrace{\mathbb{E}[\psi(Z;\theta_0)\,|\,X] = \mathbf{0}}_{\text{CMR}} \quad \Leftrightarrow \quad \underbrace{\sup_{f\in\mathscr{F}} \left|\mathbb{E}[\psi(Z;\theta_0)^\top f(X)]\right| = 0}_{\text{MMR}(\mathscr{F},\theta_0)}$$

Maximum Moment Restriction (MMR)

KM, W. Jitkrittum, J. Kübler, UAI2020

Let $\mathscr{F}$ be a space of instruments $f(x)$.

$$\underbrace{\mathbb{E}[\psi(Z;\theta_0)\,|\,X] = \mathbf{0}}_{\text{CMR}} \quad \Leftrightarrow \quad \underbrace{\sup_{f\in\mathscr{F}} \left|\mathbb{E}[\psi(Z;\theta_0)^\top f(X)]\right| = 0}_{\text{MMR}(\mathscr{F},\theta_0)}$$

- The equivalence above holds if $\mathscr{F}$ is a **universal** vector-valued RKHS.

Maximum Moment Restriction (MMR)

KM, W. Jitkrittum, J. Kübler, UAI2020

Let $\mathscr{F}$ be a space of instruments $f(x)$.

$$\underbrace{\mathbb{E}[\psi(Z;\theta_0)\,|\,X] = \mathbf{0}}_{\text{CMR}} \quad \Leftrightarrow \quad \underbrace{\sup_{f\in\mathscr{F}} \left|\mathbb{E}[\psi(Z;\theta_0)^\top f(X)]\right| = 0}_{\text{MMR}(\mathscr{F},\theta_0)}$$

- The equivalence above holds if $\mathscr{F}$ is a **universal** vector-valued RKHS.
- Let $\boldsymbol{\mu}_\theta := K_X\psi(Z;\theta)$.

$$\begin{aligned}\text{MMR}(\mathscr{F},\theta) \quad &:= \sup_{f\in\mathscr{F},\|f\|\le 1} \left|\mathbb{E}\left[\psi(Z;\theta)^\top f(X)\right]\right| \\ &= \|\mathbb{E}[K_X\psi(Z;\theta)]\|_{\mathscr{F}} \\ &= \|\boldsymbol{\mu}_\theta\|_{\mathscr{F}}.\end{aligned}$$

Maximum Moment Restriction (MMR)

KM, W. Jitkrittum, J. Kübler, UAI2020

Let $\mathscr{F}$ be a space of instruments $f(x)$.

$$\underbrace{\mathbb{E}[\psi(Z;\theta_0)\,|\,X] = \mathbf{0}}_{\text{CMR}} \quad \Leftrightarrow \quad \underbrace{\sup_{f\in\mathscr{F}} \left|\mathbb{E}[\psi(Z;\theta_0)^\top f(X)]\right| = 0}_{\text{MMR}(\mathscr{F},\theta_0)}$$

- The equivalence above holds if $\mathscr{F}$ is a **universal** vector-valued RKHS.
- Let $\boldsymbol{\mu}_\theta := K_X \psi(Z;\theta)$.

$$\begin{aligned} \text{MMR}(\mathscr{F},\theta) &:= \sup_{f\in\mathscr{F},\|f\|\le 1} \left|\mathbb{E}\left[\psi(Z;\theta)^\top f(X)\right]\right| \\ &= \|\mathbb{E}[K_X\psi(Z;\theta)]\|_{\mathscr{F}} \\ &= \|\boldsymbol{\mu}_\theta\|_{\mathscr{F}}. \end{aligned}$$

- $\text{MMR}^2(\mathscr{F},\theta) = \mathbb{E}[\psi(Z;\theta)^\top K(X,X')\psi(Z';\theta)]$.

Maximum Moment Restriction (MMR)

KM, W. Jitkrittum, J. Kübler, UAI2020

Parameter Estimation

Given observations $(x_i, z_i)_{i=1}^n$ from $\mathbb{P}(X, Z)$, we aim to estimate θ_0 by

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta\in\Theta} \widehat{\mathrm{MMR}}^2(\mathscr{F}, \theta) \\ &= \arg\min_{\theta\in\Theta} \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} \boldsymbol{\psi}(z_i;\theta)^\top K(x_i, x_j)\boldsymbol{\psi}(z_j;\theta).\end{aligned}$$

Maximum Moment Restriction (MMR)

KM, W. Jitkrittum, J. Kübler, UAI2020

Parameter Estimation

Given observations $(x_i, z_i)_{i=1}^n$ from $\mathbb{P}(X, Z)$, we aim to estimate θ_0 by

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta\in\Theta} \widehat{\mathrm{MMR}}^2(\mathscr{F}, \theta) \\ &= \arg\min_{\theta\in\Theta} \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} \psi(z_i;\theta)^\top K(x_i, x_j)\psi(z_j;\theta).\end{aligned}$$

Hypothesis Testing

Given observations $(x_i, z_i)_{i=1}^n$ from $\mathbb{P}(X, Z)$ and the parameter estimate $\hat{\theta}$, we aim to test

$$H_0 : \widehat{\mathrm{MMR}}^2(\mathscr{F}, \hat{\theta}) = 0, \qquad H_1 : \widehat{\mathrm{MMR}}^2(\mathscr{F}, \hat{\theta}) \neq 0.$$

Conditional Moment Embedding

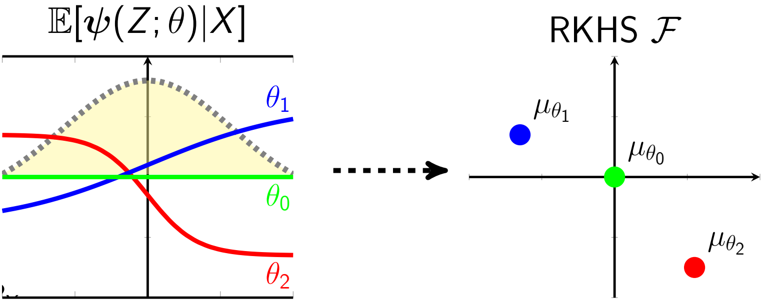


Figure: The conditional moments $\mathbb{E}[\psi(Z;\theta)|X]$ for different parameters θ are *uniquely* (P_X-almost surely) embedded into the RKHS.

Kernel Conditional Moment Test via Maximum Moment Restriction (UAI2020)
Paper: `https://arxiv.org/abs/2002.09225`
Code: `https://github.com/krikamol/kcm-test`

Future Direction

Contact:
krikamol@tuebingen.mpg.de

Website:
http://krikamol.org

Covariance Operators

- Let $\mathcal{H}, \mathcal{G}$ be RKHSes on $\mathcal{X}, \mathcal{Y}$ with feature maps

$$\phi(x) = k(x, \cdot), \qquad \varphi(y) = \ell(y, \cdot).$$

Covariance Operators

- Let $\mathcal{H}, \mathcal{G}$ be RKHSes on $\mathcal{X}, \mathcal{Y}$ with feature maps

$$\phi(x) = k(x, \cdot), \qquad \varphi(y) = \ell(y, \cdot).$$

- Let $\mathcal{C}_{XX}$ and $\mathcal{C}_{YX}$ be the **covariance operator** on X and **cross-covariance operator** from X to Y, i.e.,

$$\begin{aligned} \mathcal{C}_{XX} &= \int \phi(X) \otimes \phi(X) \, \mathrm{d}\mathbb{P}(X), \\ \mathcal{C}_{YX} &= \int \varphi(Y) \otimes \phi(X) \, \mathrm{d}\mathbb{P}(Y, X) \end{aligned}$$

Covariance Operators

- Let $\mathcal{H}, \mathcal{G}$ be RKHSes on $\mathcal{X}, \mathcal{Y}$ with feature maps

$$\phi(x) = k(x, \cdot), \qquad \varphi(y) = \ell(y, \cdot).$$

- Let $\mathcal{C}_{XX}$ and $\mathcal{C}_{YX}$ be the **covariance operator** on X and **cross-covariance operator** from X to Y, i.e.,

$$\begin{aligned}
\mathcal{C}_{XX} &= \int \phi(X) \otimes \phi(X) \, \mathrm{d}\mathbb{P}(X), \\
\mathcal{C}_{YX} &= \int \varphi(Y) \otimes \phi(X) \, \mathrm{d}\mathbb{P}(Y, X)
\end{aligned}$$

- Alternatively, $\mathcal{C}_{YX}$ is a unique bounded operator satisfying

$$\langle g, \mathcal{C}_{YX} f \rangle_{\mathcal{G}} = \mathrm{Cov}[g(Y), f(X)].$$

Covariance Operators

- Let $\mathcal{H}, \mathcal{G}$ be RKHSes on $\mathcal{X}, \mathcal{Y}$ with feature maps

$$\phi(x) = k(x, \cdot), \qquad \varphi(y) = \ell(y, \cdot).$$

- Let $\mathcal{C}_{XX}$ and $\mathcal{C}_{YX}$ be the **covariance operator** on X and **cross-covariance operator** from X to Y, i.e.,

$$\begin{aligned} \mathcal{C}_{XX} &= \int \phi(X) \otimes \phi(X) \, \mathrm{d}\mathbb{P}(X), \\ \mathcal{C}_{YX} &= \int \varphi(Y) \otimes \phi(X) \, \mathrm{d}\mathbb{P}(Y, X) \end{aligned}$$

- Alternatively, $\mathcal{C}_{YX}$ is a unique bounded operator satisfying

$$\langle g, \mathcal{C}_{YX} f \rangle_{\mathcal{G}} = \mathrm{Cov}[g(Y), f(X)].$$

- If $\mathbb{E}_{YX}[g(Y)|X = \cdot] \in \mathcal{H}$ for $g \in \mathcal{G}$, then

$$\mathcal{C}_{XX} \mathbb{E}_{YX}[g(Y)|X = \cdot] = \mathcal{C}_{XY} g.$$

Conditional Mean Estimation

- Given a joint sample $(x_1, y_1), \ldots, (x_n, y_n)$ from $\mathbb{P}(X, Y)$, we have

$$\widehat{\mathcal{C}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i), \qquad \widehat{\mathcal{C}}_{YX} = \frac{1}{n}\sum_{i=1}^{n} \varphi(y_i) \otimes \phi(x_i).$$

Conditional Mean Estimation

- Given a joint sample $(x_1, y_1), \ldots, (x_n, y_n)$ from $\mathbb{P}(X, Y)$, we have

$$\widehat{\mathcal{C}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i), \qquad \widehat{\mathcal{C}}_{YX} = \frac{1}{n}\sum_{i=1}^{n} \varphi(y_i) \otimes \phi(x_i).$$

- Then, $\boldsymbol{\mu}_{Y|x}$ for some $x \in \mathcal{X}$ can be estimated as

$$\hat{\boldsymbol{\mu}}_{Y|x} = \widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \varepsilon\mathcal{I})^{-1}k(x,\cdot) = \Phi(\mathbf{K} + n\varepsilon\mathbf{I}_n)^{-1}\mathbf{k}_x = \sum_{i=1}^{n} \beta_i \varphi(y_i),$$

where $\varepsilon > 0$ is a regularization parameter and

$$\Phi = [\varphi(y_1), .., \varphi(y_n)], \quad \mathbf{K}_{ij} = k(x_i, x_j), \quad \mathbf{k_x} = [k(x_1, x), .., k(x_n, x)].$$

Conditional Mean Estimation

- Given a joint sample $(x_1, y_1), \ldots, (x_n, y_n)$ from $\mathbb{P}(X, Y)$, we have

$$\widehat{\mathcal{C}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i), \qquad \widehat{\mathcal{C}}_{YX} = \frac{1}{n}\sum_{i=1}^{n} \varphi(y_i) \otimes \phi(x_i).$$

- Then, $\boldsymbol{\mu}_{Y|x}$ for some $x \in \mathcal{X}$ can be estimated as

$$\hat{\boldsymbol{\mu}}_{Y|x} = \widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \varepsilon\mathcal{I})^{-1}k(x, \cdot) = \Phi(\mathbf{K} + n\varepsilon\mathbf{I}_n)^{-1}\mathbf{k}_x = \sum_{i=1}^{n} \beta_i \varphi(y_i),$$

  where $\varepsilon > 0$ is a regularization parameter and

$$\Phi = [\varphi(y_1), .., \varphi(y_n)], \quad \mathbf{K}_{ij} = k(x_i, x_j), \quad \mathbf{k_x} = [k(x_1, x), .., k(x_n, x)].$$

- Under some technical assumptions, $\hat{\boldsymbol{\mu}}_{Y|x} \to \boldsymbol{\mu}_{Y|x}$ as $n \to \infty$.

Kernel Sum Rule: $\mathbb{P}(X) = \sum_Y \mathbb{P}(X, Y)$

- By the law of total expectation,

$$
\begin{aligned}
\mu_X &= \mathbb{E}_X[\phi(X)] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[\phi(X)|Y]] \\
&= \mathbb{E}_Y[\mathcal{U}_{X|Y}\varphi(Y)] = \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y)] \\
&= \mathcal{U}_{X|Y}\mu_Y
\end{aligned}
$$

Kernel Sum Rule: $\mathbb{P}(X) = \sum_Y \mathbb{P}(X, Y)$

- By the law of total expectation,

$$\begin{aligned}\mu_X &= \mathbb{E}_X[\phi(X)] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[\phi(X)|Y]] \\ &= \mathbb{E}_Y[\mathcal{U}_{X|Y}\varphi(Y)] = \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y)] \\ &= \mathcal{U}_{X|Y}\mu_Y\end{aligned}$$

- Let $\hat{\boldsymbol{\mu}}_Y = \sum_{i=1}^m \alpha_i \varphi(\tilde{y}_i)$ and $\widehat{\mathcal{U}}_{X|Y} = \widehat{\mathcal{C}}_{XY}\widehat{\mathcal{C}}_{YY}^{-1}$. Then,

$$\hat{\boldsymbol{\mu}}_X = \widehat{\mathcal{U}}_{X|Y}\hat{\boldsymbol{\mu}}_Y = \widehat{\mathcal{C}}_{XY}\widehat{\mathcal{C}}_{YY}^{-1}\hat{\boldsymbol{\mu}}_Y = \Upsilon(\mathbf{L} + n\lambda I)^{-1}\tilde{\mathbf{L}}\boldsymbol{\alpha}.$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^\top$, $\mathbf{L}_{ij} = l(y_i, y_j)$, and $\tilde{\mathbf{L}}_{ij} = l(y_i, \tilde{y}_j)$.

Kernel Sum Rule: $\mathbb{P}(X) = \sum_Y \mathbb{P}(X, Y)$

- By the law of total expectation,

$$\begin{aligned} \mu_X &= \mathbb{E}_X[\phi(X)] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[\phi(X)|Y]] \\ &= \mathbb{E}_Y[\mathcal{U}_{X|Y}\varphi(Y)] = \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y)] \\ &= \mathcal{U}_{X|Y}\mu_Y \end{aligned}$$

- Let $\hat{\boldsymbol{\mu}}_Y = \sum_{i=1}^m \alpha_i \varphi(\tilde{y}_i)$ and $\widehat{\mathcal{U}}_{X|Y} = \widehat{\mathcal{C}}_{XY}\widehat{\mathcal{C}}_{YY}^{-1}$. Then,

$$\hat{\boldsymbol{\mu}}_X = \widehat{\mathcal{U}}_{X|Y}\hat{\boldsymbol{\mu}}_Y = \widehat{\mathcal{C}}_{XY}\widehat{\mathcal{C}}_{YY}^{-1}\hat{\boldsymbol{\mu}}_Y = \Upsilon(\mathbf{L} + n\lambda I)^{-1}\tilde{\mathbf{L}}\boldsymbol{\alpha}.$$

  where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$, $\mathbf{L}_{ij} = l(y_i, y_j)$, and $\tilde{\mathbf{L}}_{ij} = l(y_i, \tilde{y}_j)$.
- That is, we have

$$\boxed{\hat{\boldsymbol{\mu}}_X = \textstyle\sum_{j=1}^n \beta_j \phi(x_j)}$$

  with $\boldsymbol{\beta} = (\mathbf{L} + n\lambda\mathbf{I})^{-1}\tilde{\mathbf{L}}\boldsymbol{\alpha}$.

Kernel Product Rule: $\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$

- We can factorize $\mu_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \varphi(Y)]$ as

$$
\begin{aligned}
\mathbb{E}_Y[\mathbb{E}_{X|Y}[\phi(X)|Y] \otimes \varphi(Y)] &= \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)] \\
\mathbb{E}_X[\mathbb{E}_{Y|X}[\varphi(Y)|X] \otimes \phi(X)] &= \mathcal{U}_{Y|X}\mathbb{E}_X[\phi(X) \otimes \phi(X)]
\end{aligned}
$$

[5] Fukumizu et al. *Kernel Bayes' Rule*. JMLR. 2013

Kernel Product Rule: $\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$

- We can factorize $\mu_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \varphi(Y)]$ as

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}_{X|Y}[\phi(X)|Y] \otimes \varphi(Y)] &= \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)] \\
\mathbb{E}_X[\mathbb{E}_{Y|X}[\varphi(Y)|X] \otimes \phi(X)] &= \mathcal{U}_{Y|X}\mathbb{E}_X[\phi(X) \otimes \phi(X)]
\end{aligned}$$

- Let $\mu_X^{\otimes} = \mathbb{E}_X[\phi(X) \otimes \phi(X)]$ and $\mu_Y^{\otimes} = \mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)]$.

[5] Fukumizu et al. *Kernel Bayes' Rule*. JMLR. 2013

Kernel Product Rule: $\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$

- We can factorize $\mu_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \varphi(Y)]$ as

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}_{X|Y}[\phi(X)|Y] \otimes \varphi(Y)] &= \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)] \\
\mathbb{E}_X[\mathbb{E}_{Y|X}[\varphi(Y)|X] \otimes \phi(X)] &= \mathcal{U}_{Y|X}\mathbb{E}_X[\phi(X) \otimes \phi(X)]
\end{aligned}$$

- Let $\mu_X^{\otimes} = \mathbb{E}_X[\phi(X) \otimes \phi(X)]$ and $\mu_Y^{\otimes} = \mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)]$.
- Then, the product rule becomes

$$\mu_{XY} = \mathcal{U}_{X|Y}\mu_Y^{\otimes} = \mathcal{U}_{Y|X}\mu_X^{\otimes}.$$

[5] Fukumizu et al. *Kernel Bayes' Rule*. JMLR. 2013

Kernel Product Rule: $\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$

- We can factorize $\mu_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \varphi(Y)]$ as

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}_{X|Y}[\phi(X)|Y] \otimes \varphi(Y)] &= \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)] \\
\mathbb{E}_X[\mathbb{E}_{Y|X}[\varphi(Y)|X] \otimes \phi(X)] &= \mathcal{U}_{Y|X}\mathbb{E}_X[\phi(X) \otimes \phi(X)]
\end{aligned}$$

- Let $\mu_X^{\otimes} = \mathbb{E}_X[\phi(X) \otimes \phi(X)]$ and $\mu_Y^{\otimes} = \mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)]$.
- Then, the product rule becomes

$$\mu_{XY} = \mathcal{U}_{X|Y}\mu_Y^{\otimes} = \mathcal{U}_{Y|X}\mu_X^{\otimes}.$$

- Alternatively, we may write the above formulation as

$$\mathcal{C}_{XY} = \mathcal{U}_{X|Y}\mathcal{C}_{YY} \quad \text{and} \quad \mathcal{C}_{YX} = \mathcal{U}_{Y|X}\mathcal{C}_{XX}$$

[5] Fukumizu et al. *Kernel Bayes' Rule*. JMLR. 2013

Kernel Product Rule: $\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$

- We can factorize $\mu_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \varphi(Y)]$ as

$$
\begin{aligned}
\mathbb{E}_Y[\mathbb{E}_{X|Y}[\phi(X)|Y] \otimes \varphi(Y)] &= \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)] \\
\mathbb{E}_X[\mathbb{E}_{Y|X}[\varphi(Y)|X] \otimes \phi(X)] &= \mathcal{U}_{Y|X}\mathbb{E}_X[\phi(X) \otimes \phi(X)]
\end{aligned}
$$

- Let $\mu_X^{\otimes} = \mathbb{E}_X[\phi(X) \otimes \phi(X)]$ and $\mu_Y^{\otimes} = \mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)]$.
- Then, the product rule becomes

$$\mu_{XY} = \mathcal{U}_{X|Y}\mu_Y^{\otimes} = \mathcal{U}_{Y|X}\mu_X^{\otimes}.$$

- Alternatively, we may write the above formulation as

$$\mathcal{C}_{XY} = \mathcal{U}_{X|Y}\mathcal{C}_{YY} \quad \text{and} \quad \mathcal{C}_{YX} = \mathcal{U}_{Y|X}\mathcal{C}_{XX}$$

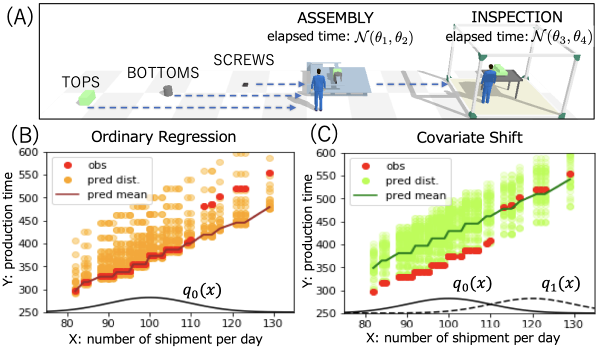
- The kernel sum and product rules can be combined to obtain the **kernel Bayes' rule**.[5]

[5] Fukumizu et al. *Kernel Bayes' Rule*. JMLR. 2013

Calibration of Computer Simulation

Kennedy and O'Hagan (2002); Kisamori et al., (AISTATS 2020)

Figure taken from Kisamori et al., (2020)



The computer simulator: $r(x, \theta)$, $\theta \in \Theta$.

The posterior embedding: $\mu_{\Theta|r^*} := \int k_\Theta(\cdot, \theta)\, dP_\pi(\theta|r^*)$