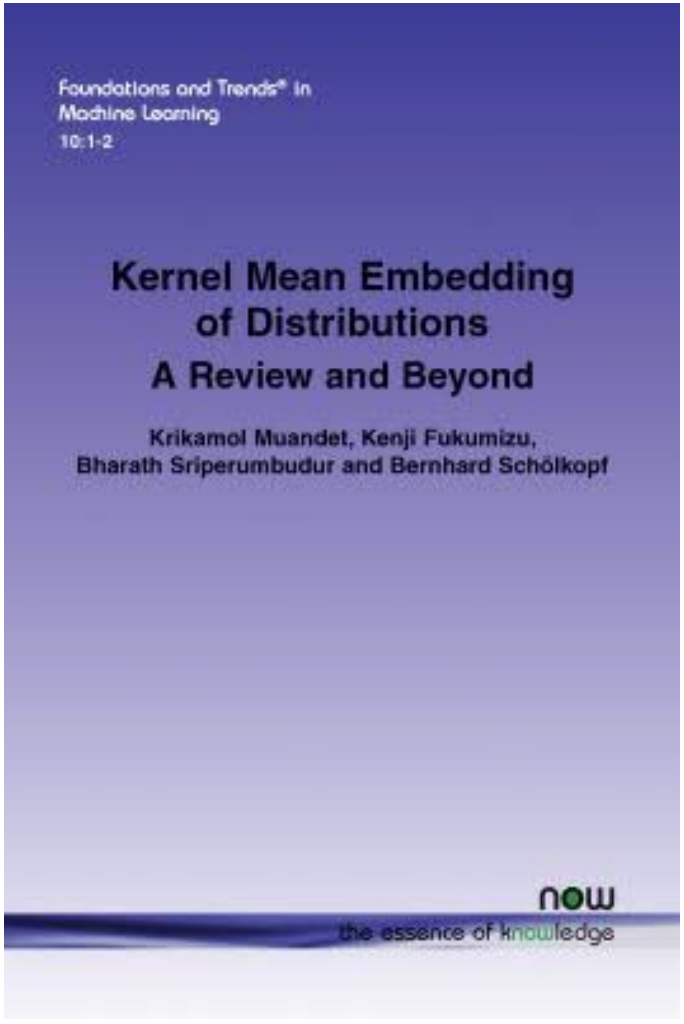
Kernel Mean Embedding with Applications in Deconfounded Causal Learning

Krikamol Muandet
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Lifting Inference with Kernel Embeddings (LIKE22)
January 11, 2022

Reference



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

Recap on Kernel Methods

Kernel Mean Embedding of Distributions

Conditional Mean Embedding

Deconfounded Causal Learning

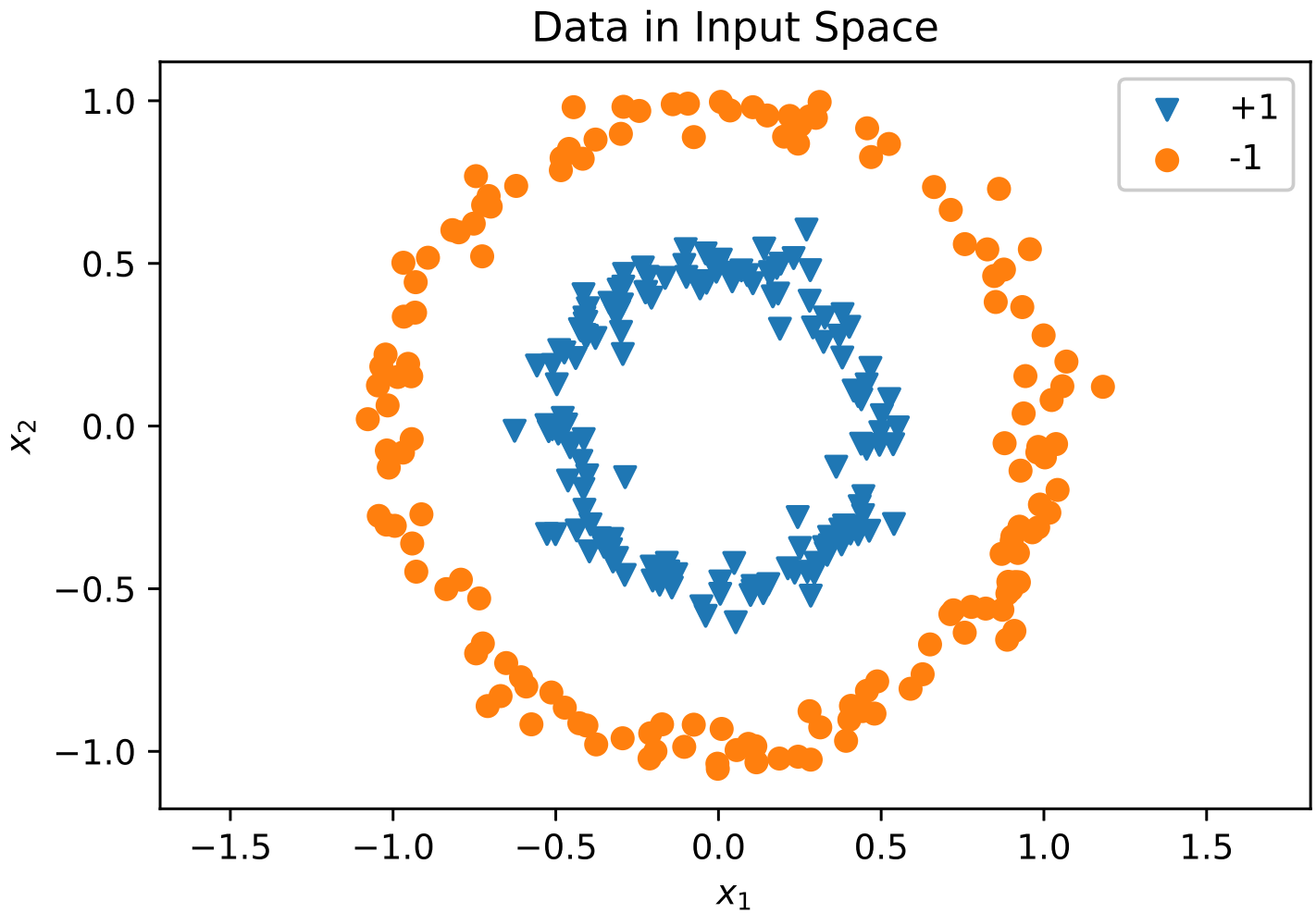
Recap on Kernel Methods

Kernel Mean Embedding of Distributions

Conditional Mean Embedding

Deconfounded Causal Learning

Classification Problem

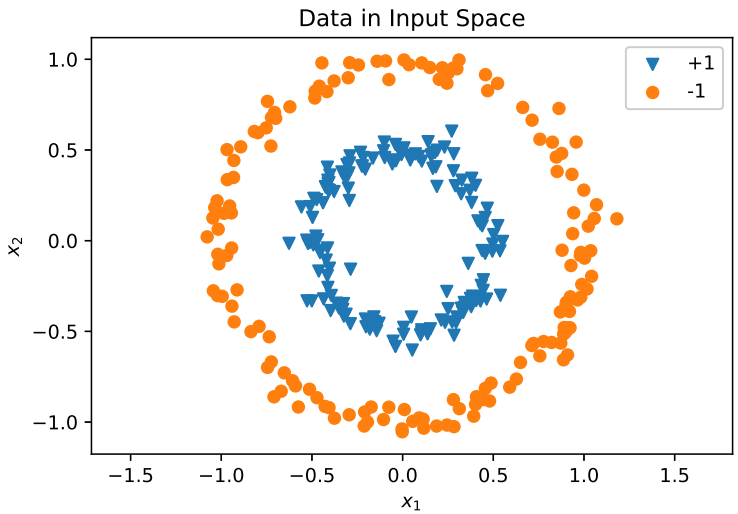


Feature Map

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

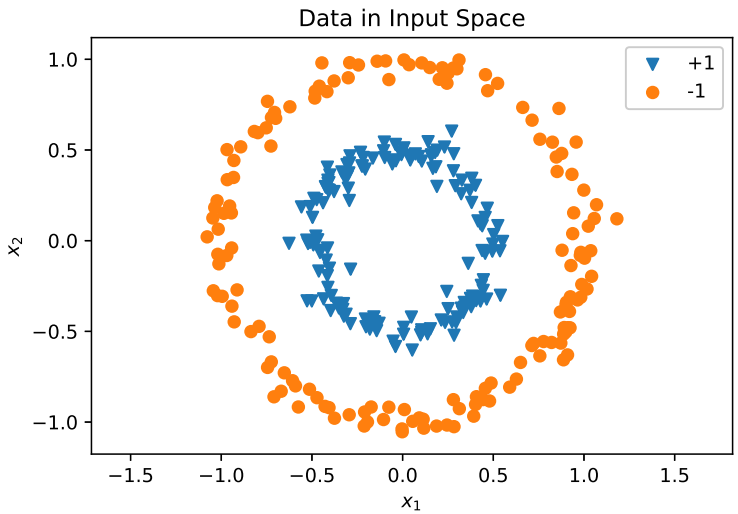
Feature Map

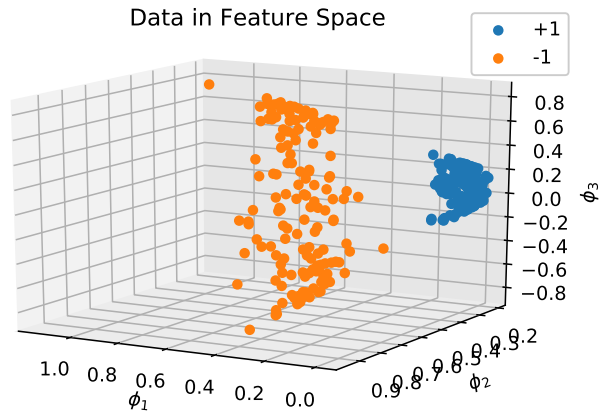
$$\phi \ : \ (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$



Feature Map

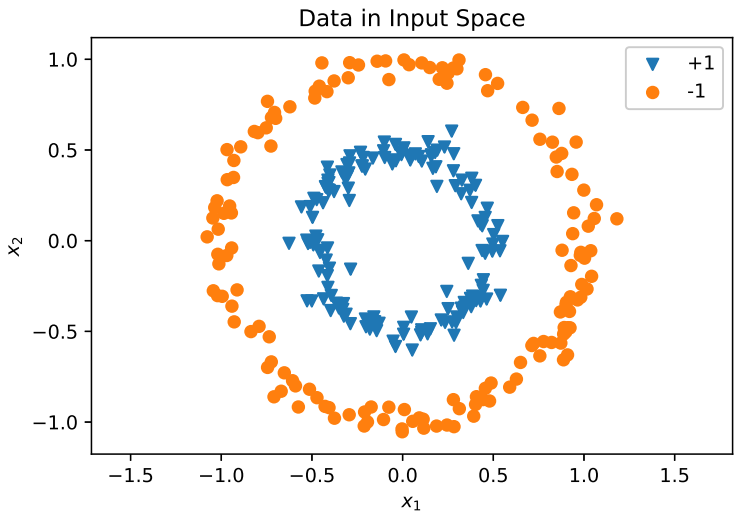
$$\phi \ : \ (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

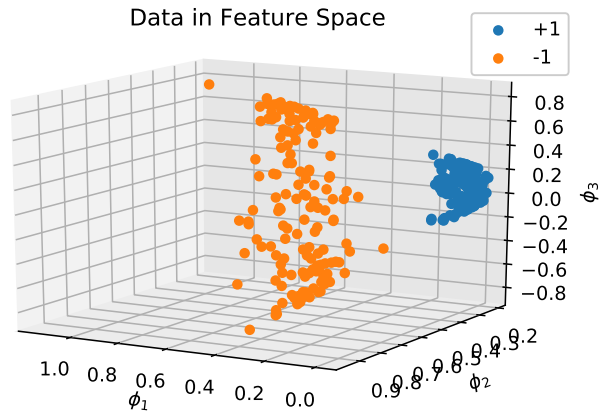




Feature Map

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

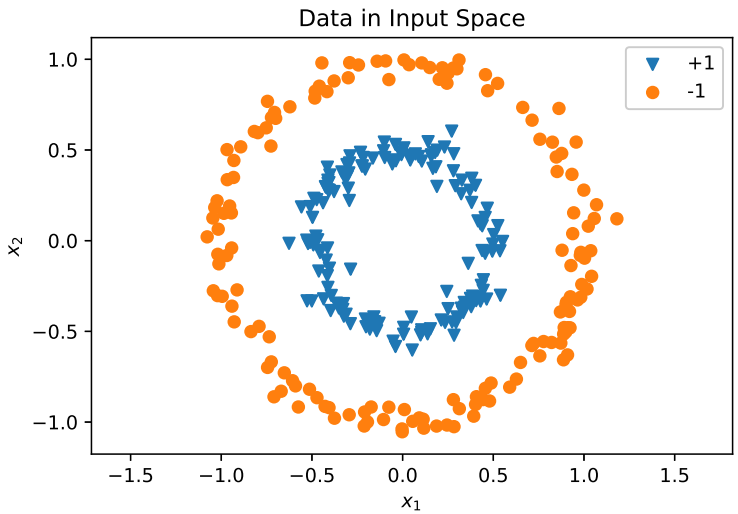


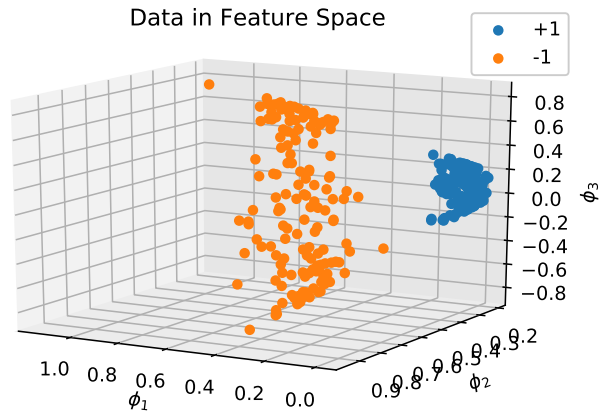


$$\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle_{\mathbb{R}^3} = x_1^2 z_1^2 + x_2^2 z_2^2 + 2(x_1x_2)(z_1z_2) = (x_1z_1 + x_2z_2)^2 = (\mathbf{x} \cdot \mathbf{z})^2$$

Feature Map

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

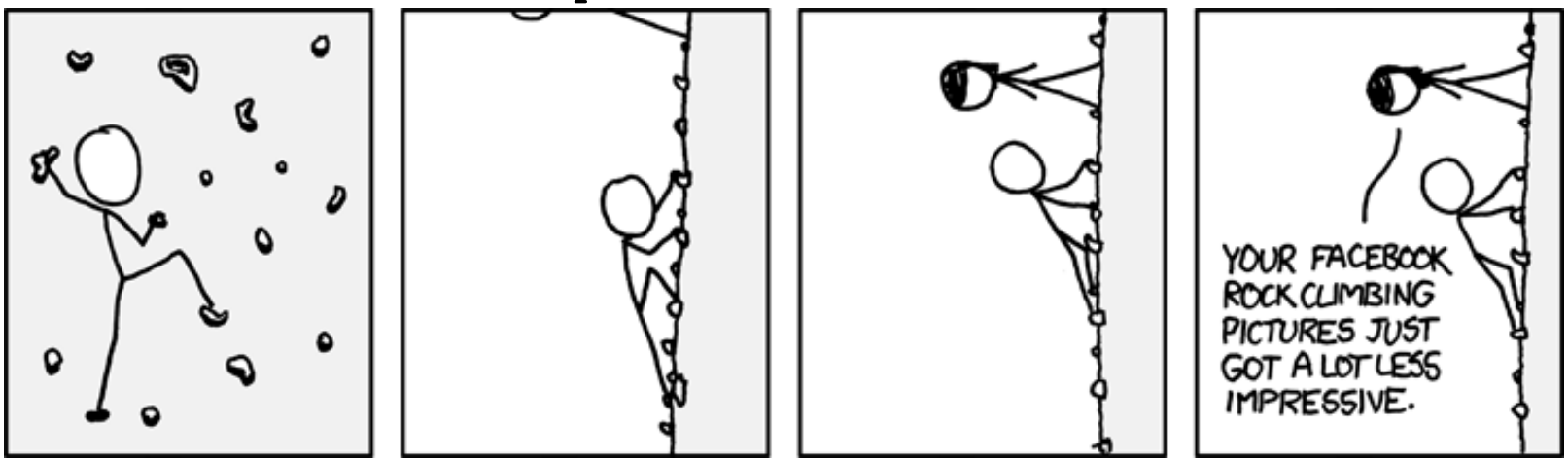




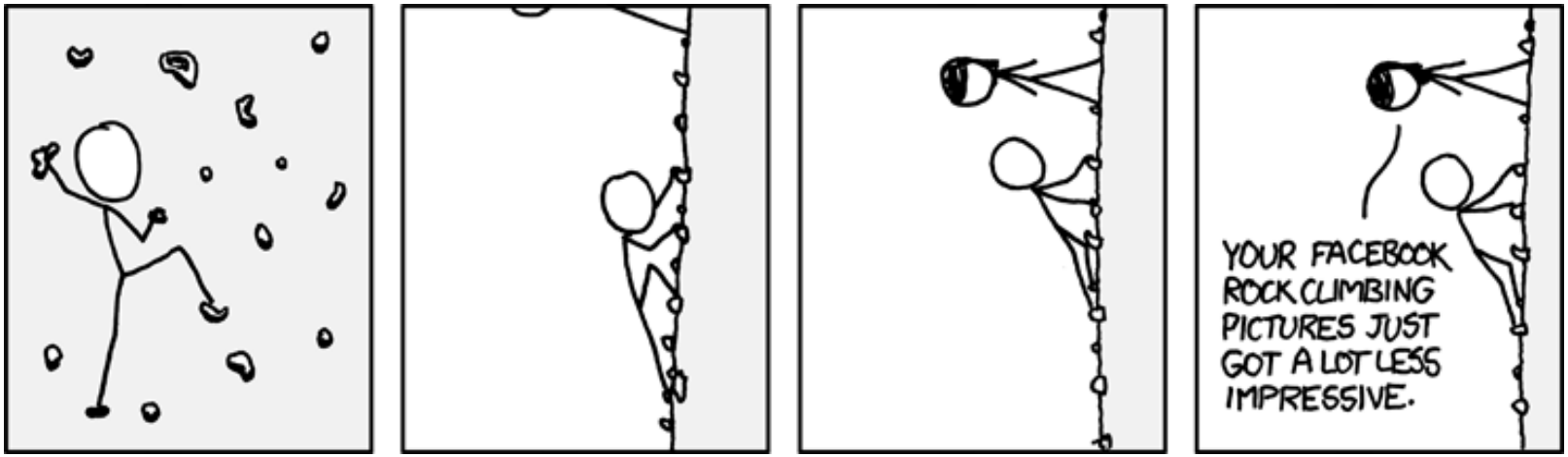
$$\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle_{\mathbb{R}^3} = x_1^2 z_1^2 + x_2^2 z_2^2 + 2(x_1x_2)(z_1z_2) = (x_1z_1 + x_2z_2)^2 = (\mathbf{x} \cdot \mathbf{z})^2$$

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

https://xkcd.com/655/



https://xkcd.com/655/



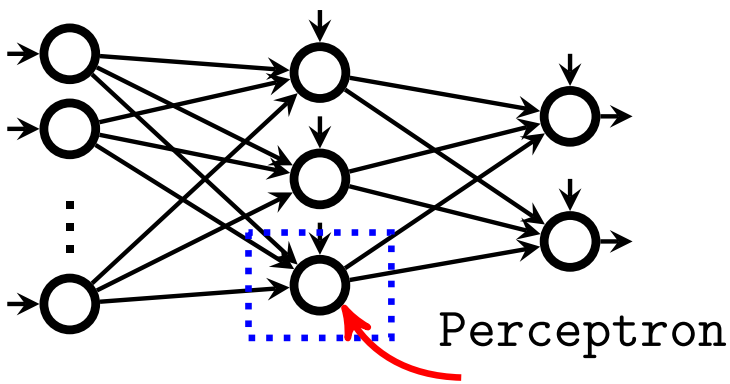
Recipe for ML Problems

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

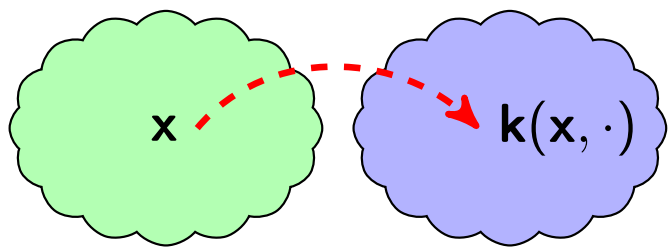
Perceptron[1]**:** $f(\mathbf{x}) = \boldsymbol{w}^\top \mathbf{x} + b$

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

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

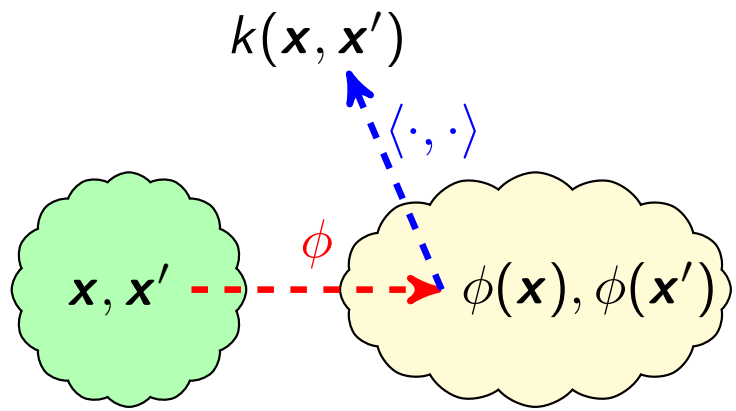
We call ϕ a **feature map** and $\mathcal{H}$ a **feature space** associated with k.

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

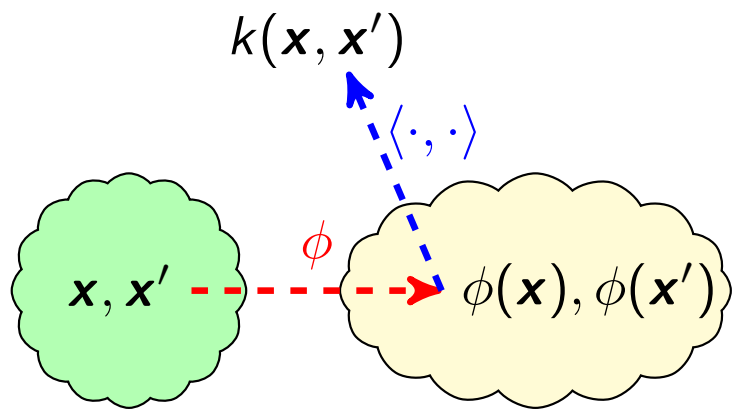
A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ we have

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

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 $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^2$
   - $\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2)$
   - $\mathcal{H} = \mathbb{R}^3$
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 $\mathbf{K}$ 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 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.

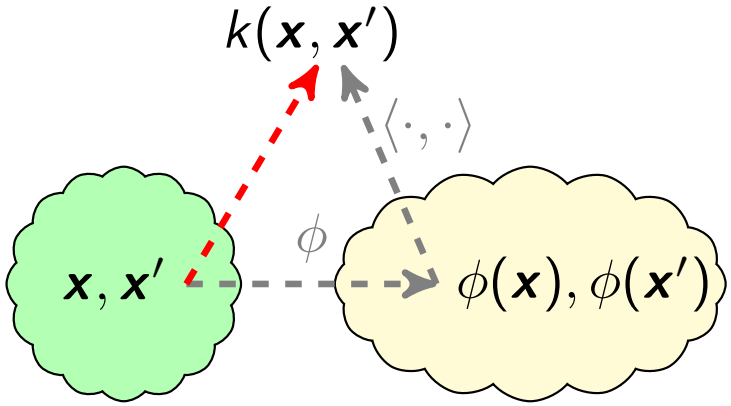
RKHS as Feature Space

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**.



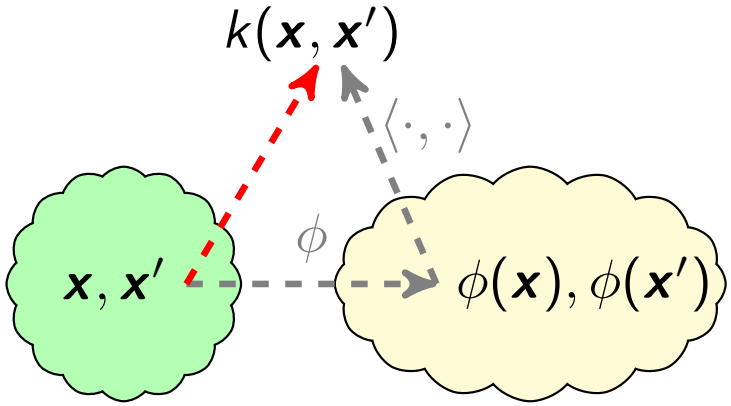
RKHS as Feature Space

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

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

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.

Recap on Kernel Methods

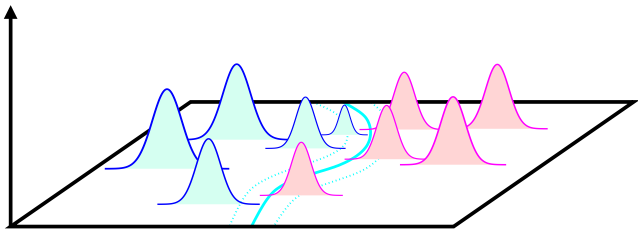
Kernel Mean Embedding of Distributions

Conditional Mean Embedding

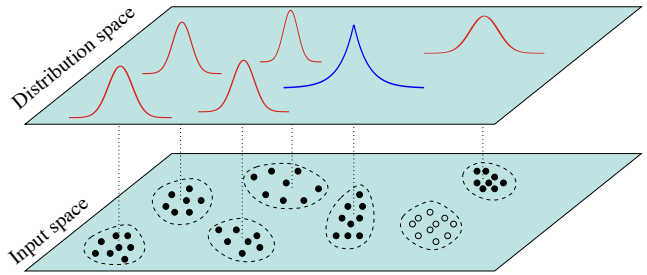
Deconfounded Causal Learning

Probability Measures

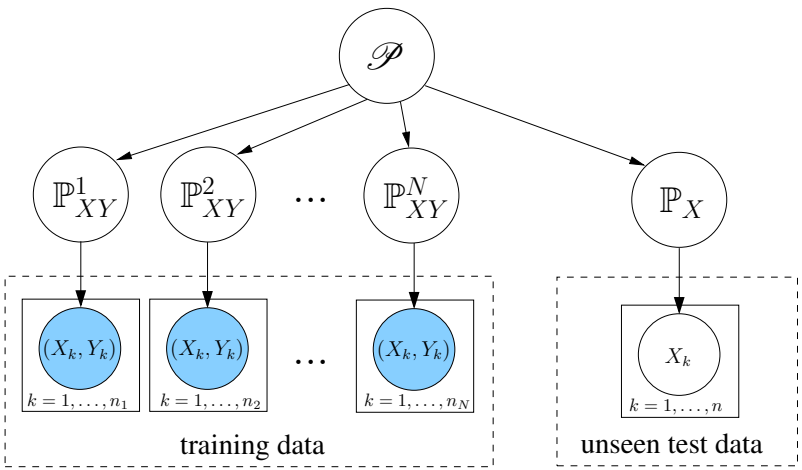
Learning on Distributions/Point Clouds



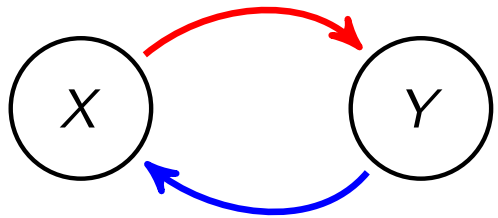
Group Anomaly/OOD Detection

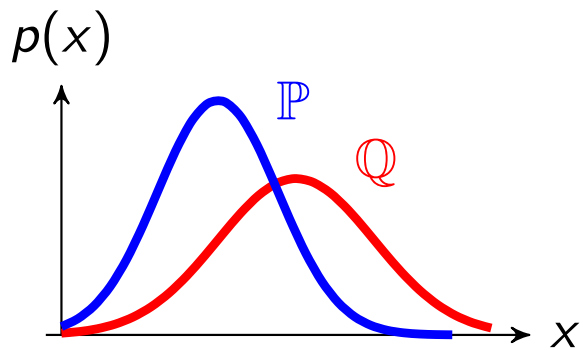


Generalization across Environments

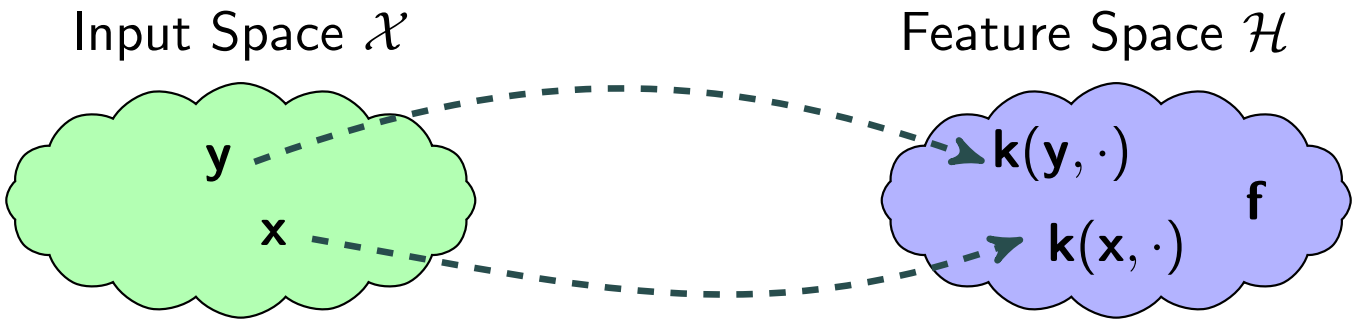


Statistical and Causal Inference

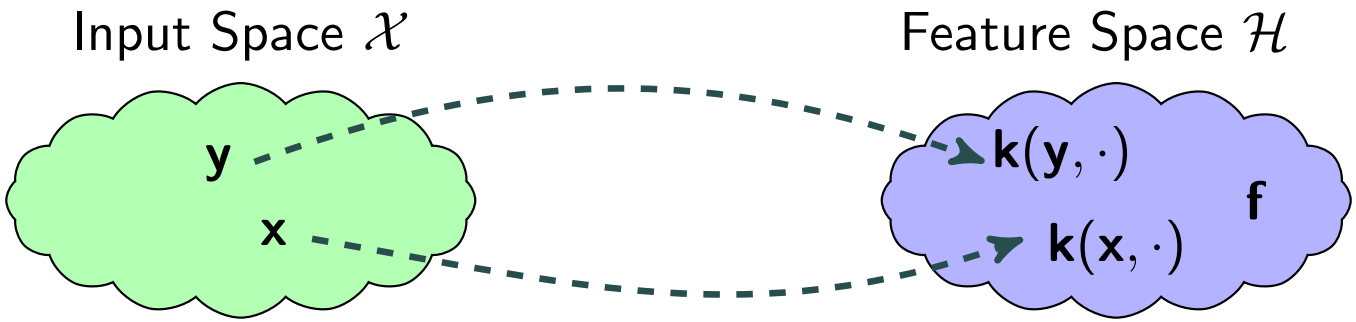


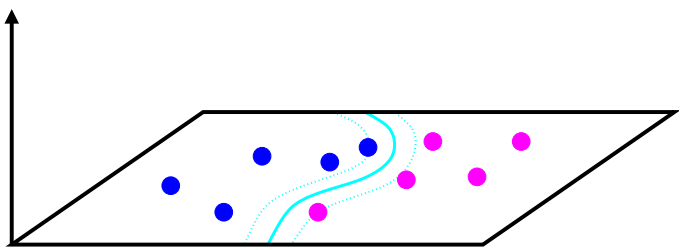


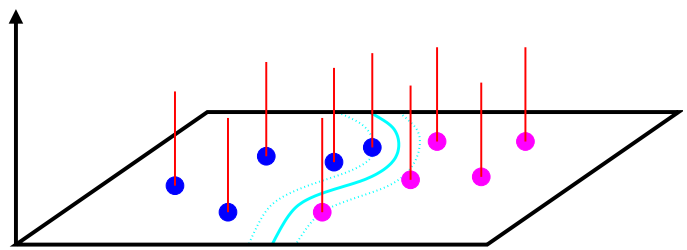
Embedding of Dirac Measures



Embedding of Dirac Measures



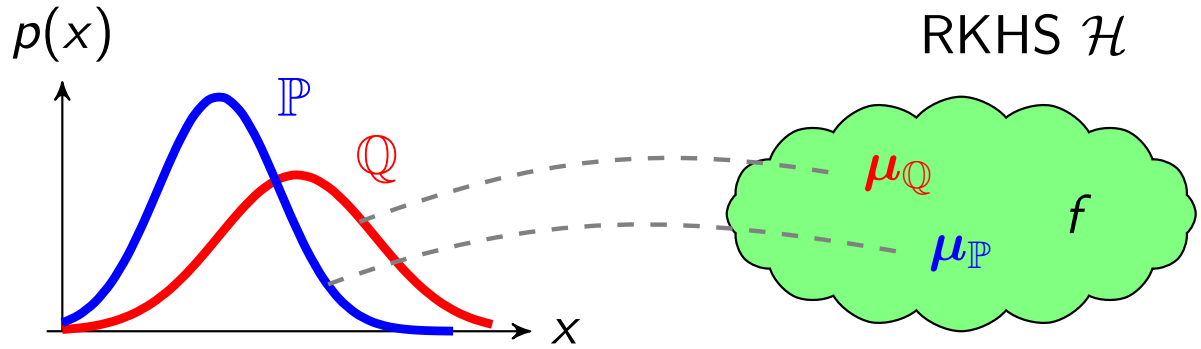




$x \mapsto k(\cdot, x)$

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

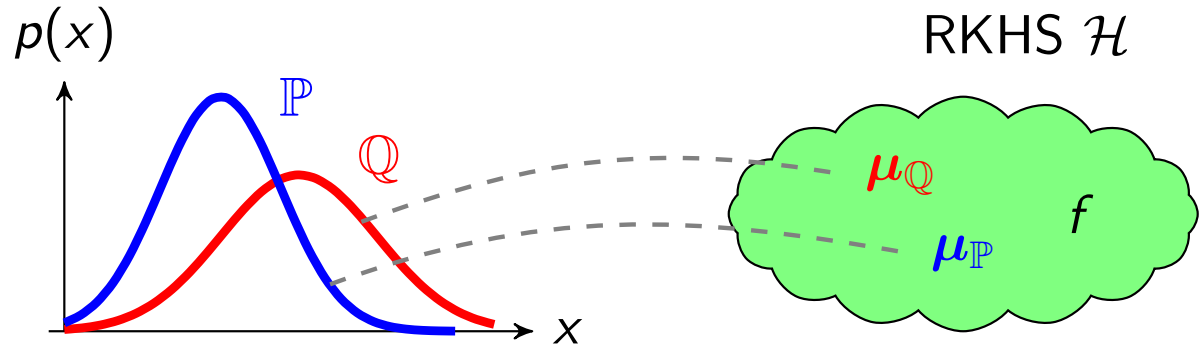
Embedding of Marginal Distributions



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

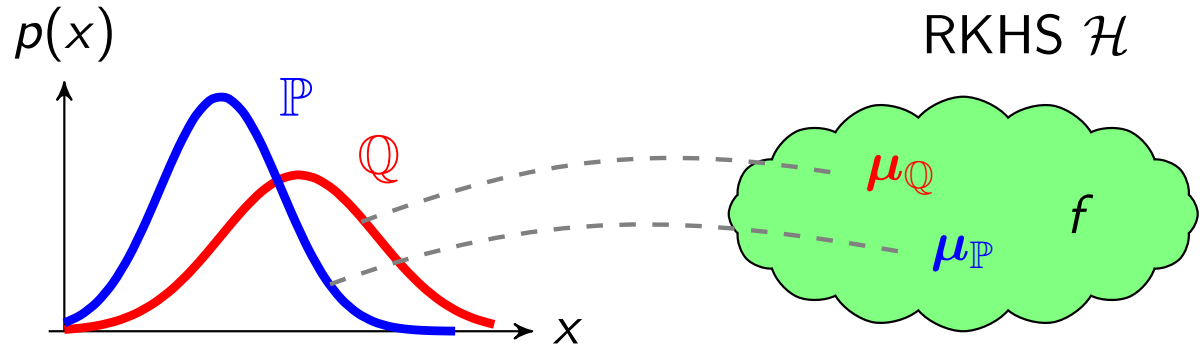
Let $\mathbb{P}$ be a probability measure defined on a measurable space $(\mathcal{X}, \Sigma)$ with a σ-algebra Σ.

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.

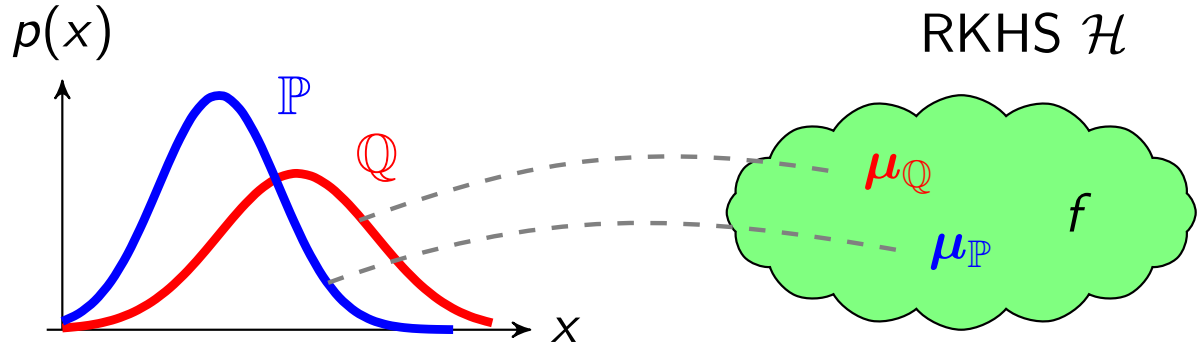
Kernel mean embedding

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**.

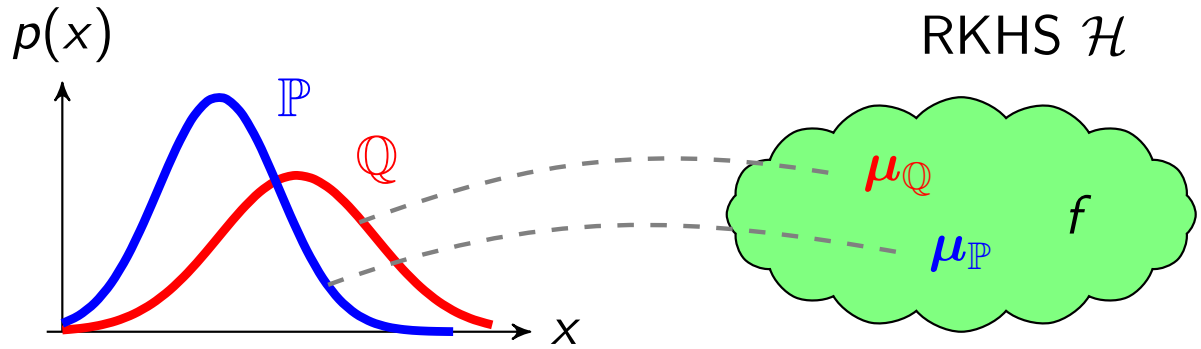
Embedding of Marginal Distributions



- 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 $x_1, x_2, \ldots, x_n$ 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 $1-\delta$,

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

- 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 $x_1, x_2, \ldots, x_n$ 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 $1-\delta$,

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

- 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]

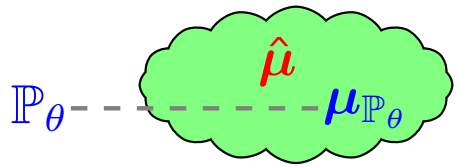
$$\hat{\boldsymbol{\mu}}_{\alpha} := \alpha f^* + (1-\alpha)\hat{\boldsymbol{\mu}}_{\mathbb{P}}, \quad f^* \in \mathcal{H}.$$

[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

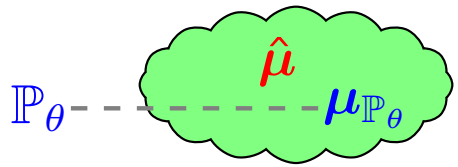
$$\theta^* = \arg\min_{\theta} \ \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\mathbb{P}_\theta}\|^2_{\mathcal{H}}.$$



Recovering Samples/Distributions

- An approximate pre-image problem

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



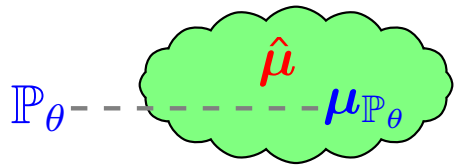
- The distribution $\mathbb{P}_\theta$ is assumed to be in a certain class

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

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

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

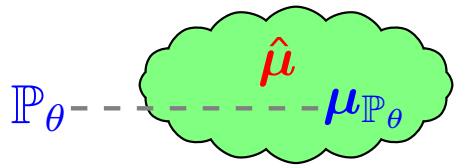
- **Kernel herding** generates deterministic *pseudo-samples* by greedily minimizing the squared error

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

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

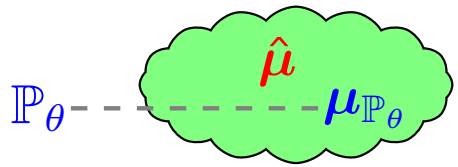
$$\mathcal{E}_T^2 = \left\| \boldsymbol{\mu}_{\mathbb{P}} - \frac{1}{T} \sum_{t=1}^{T} k(\cdot, \mathbf{x}_t) \right\|^2_{\mathcal{H}}.$$

- **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 $\mathbb{P}$

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

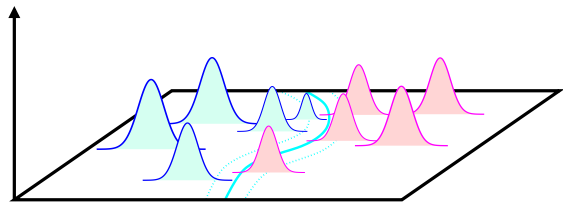
- If k is **characteristic**, $\boldsymbol{\mu}_{\mathbb{P}}$ captures all information about $\mathbb{P}$.
- All **universal** kernels are characteristic, but not vice versa.
- The empirical $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ requires **no parametric assumption** about $\mathbb{P}$.
- It can be estimated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

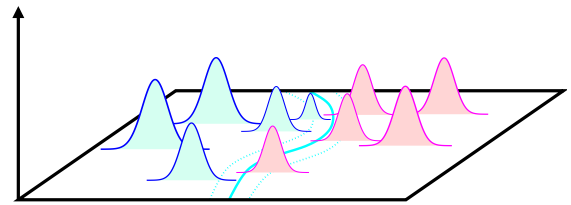
Learning from Distributions



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

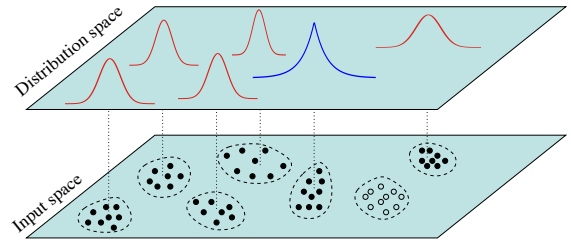
Application: High-Level Generalization

Learning from Distributions



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

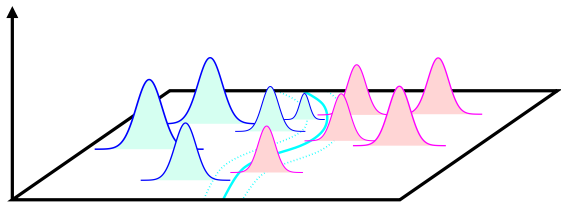
Group Anomaly Detection



KM. and Schölkopf, UAI 2013.

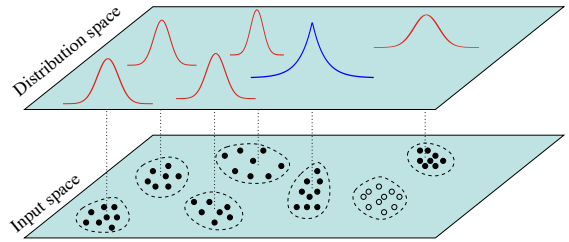
Application: High-Level Generalization

Learning from Distributions



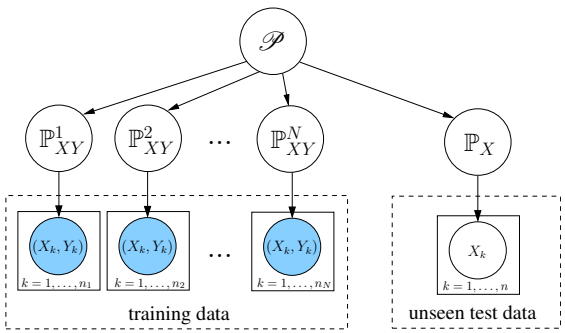
KM., Fukumizu, Dinuzzo, Schölkopf. NIPS 2012.

Group Anomaly Detection



KM. and Schölkopf, UAI 2013.

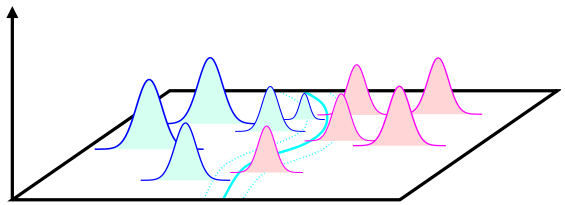
Domain Generalization



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

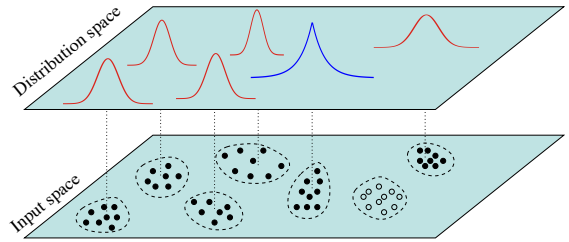
Application: High-Level Generalization

Learning from Distributions



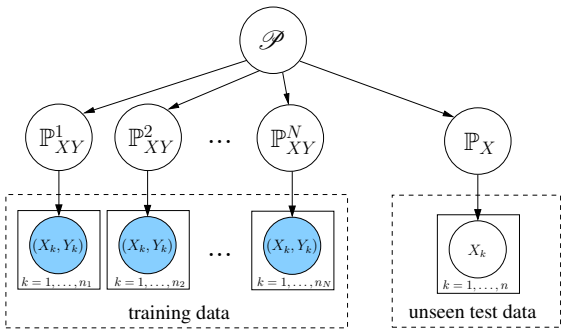
KM., Fukumizu, Dinuzzo, Schölkopf. NIPS 2012.

Group Anomaly Detection



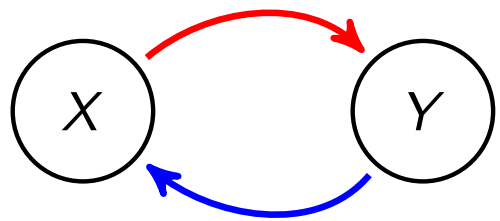
KM. and Schölkopf, UAI 2013.

Domain Generalization



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

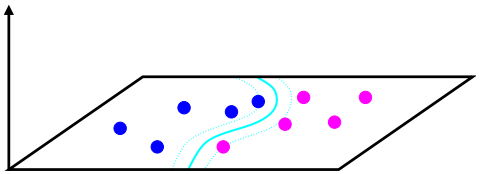
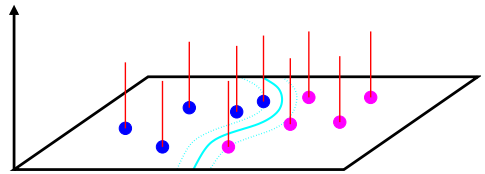
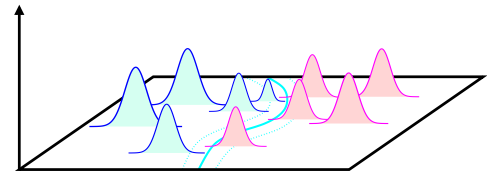
Cause-Effect Inference



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

Support Measure Machine (SMM)

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

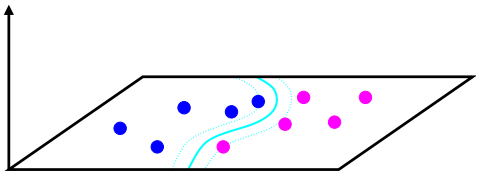
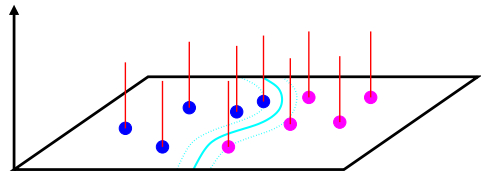
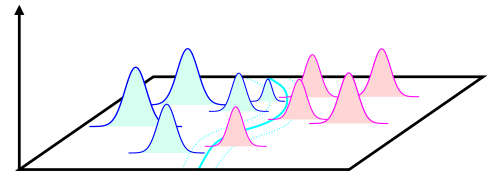




$x \mapsto k(\cdot, x)$ $\quad$ $\delta_x \mapsto \int k(\cdot, z) d\delta_x(z)$ $\quad$ $\mathbb{P} \mapsto \int k(\cdot, z) d\mathbb{P}(z)$

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

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

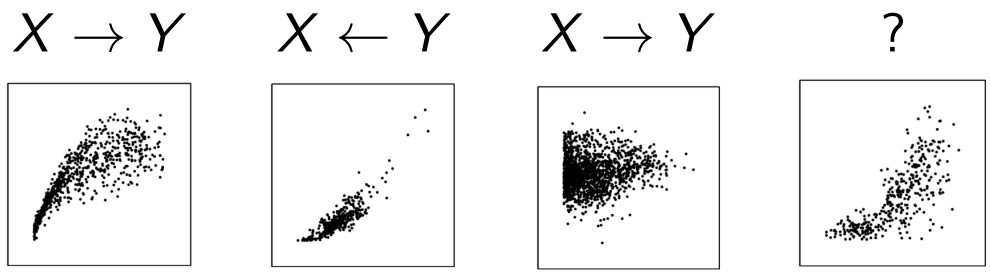
Supervised Learning on Point Clouds

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



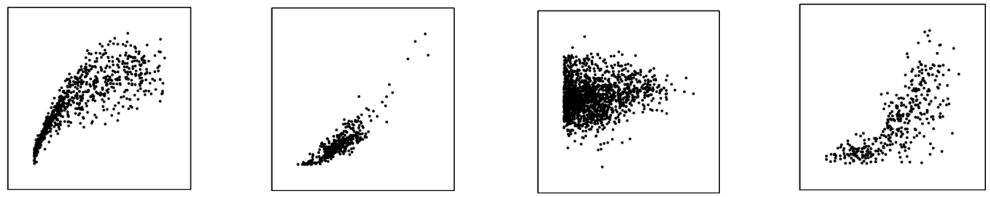
Lopez-Paz, **KM.**, B. Schölkopf, I. Tolstikhin. JMLR 2015, ICML 2015.

Supervised Learning on Point Clouds

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

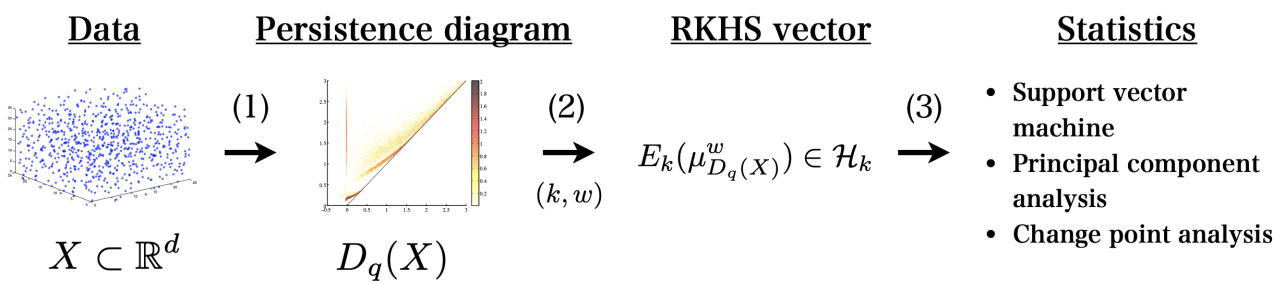
Causal Prediction

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



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

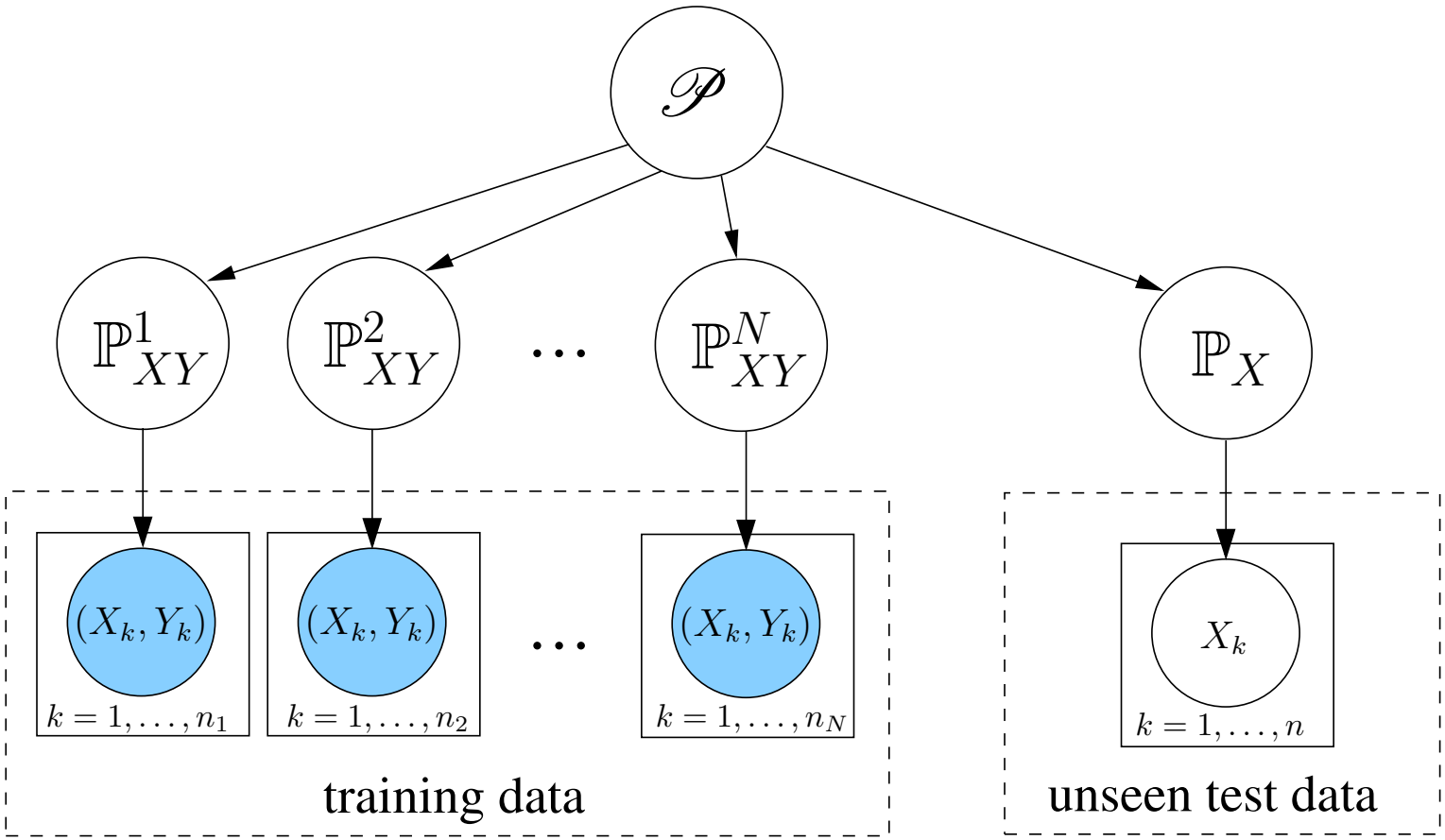
Topological Data Analysis



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:

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

- 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:

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

- 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:

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

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

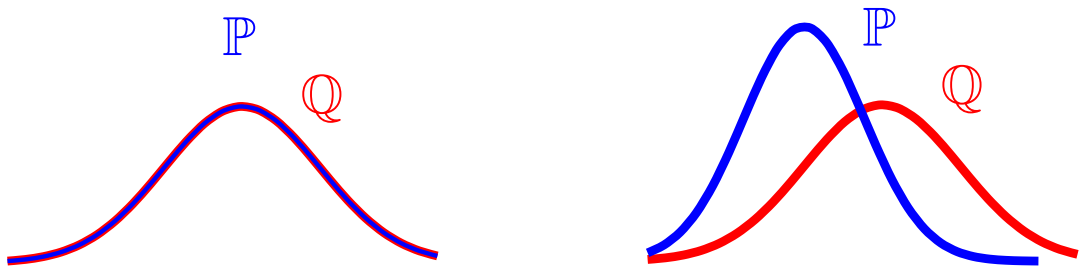
$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}, \mathcal{H}) := \sup_{h \in \mathcal{H}, \|h\| \leq 1} \left| \int h(x) \, \mathrm{d}\mathbb{P}(x) - \int h(x) \, \mathrm{d}\mathbb{Q}(x) \right|.$$

- 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

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

Kernel Two-Sample Testing

Gretton et al., JMLR2012

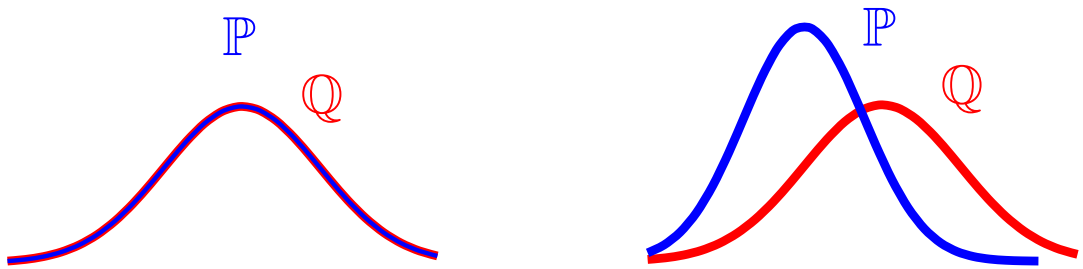


Question: Given $\{\boldsymbol{x}_i\}_{i=1}^n \sim \mathbb{P}$ and $\{\boldsymbol{y}_j\}_{j=1}^n \sim \mathbb{Q}$, check if $\mathbb{P} = \mathbb{Q}$.

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

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:

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

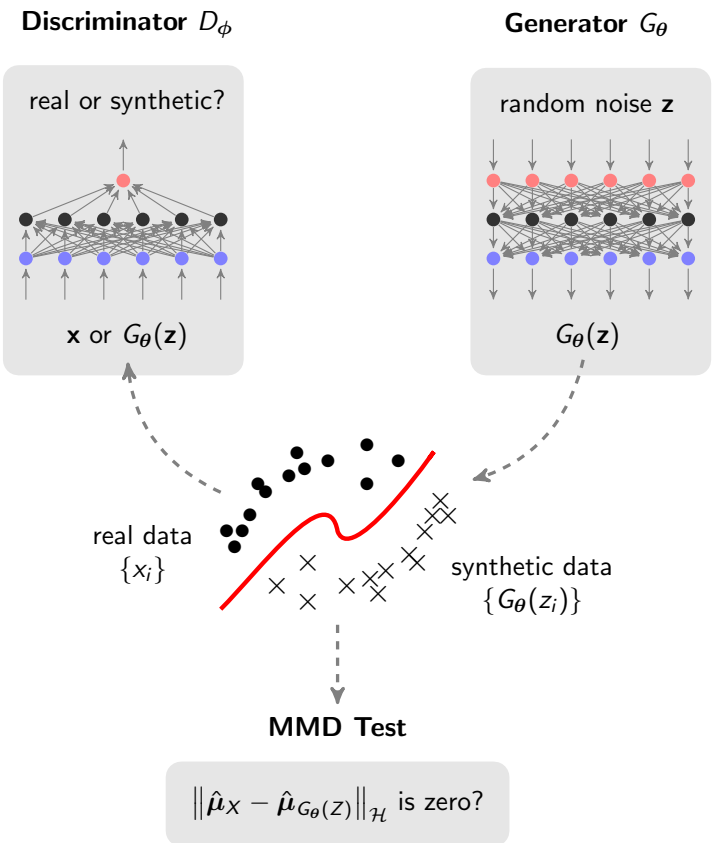
where $h((x_i, y_i), (x_j, y_j)) = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$.

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

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

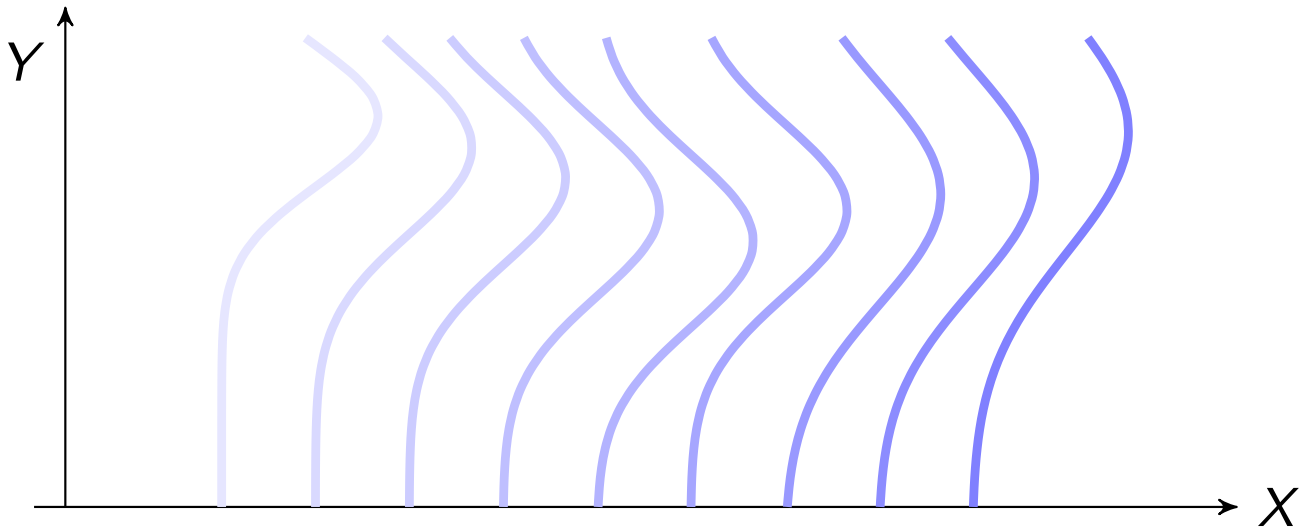
Generative Moment Matching Network

- 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

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

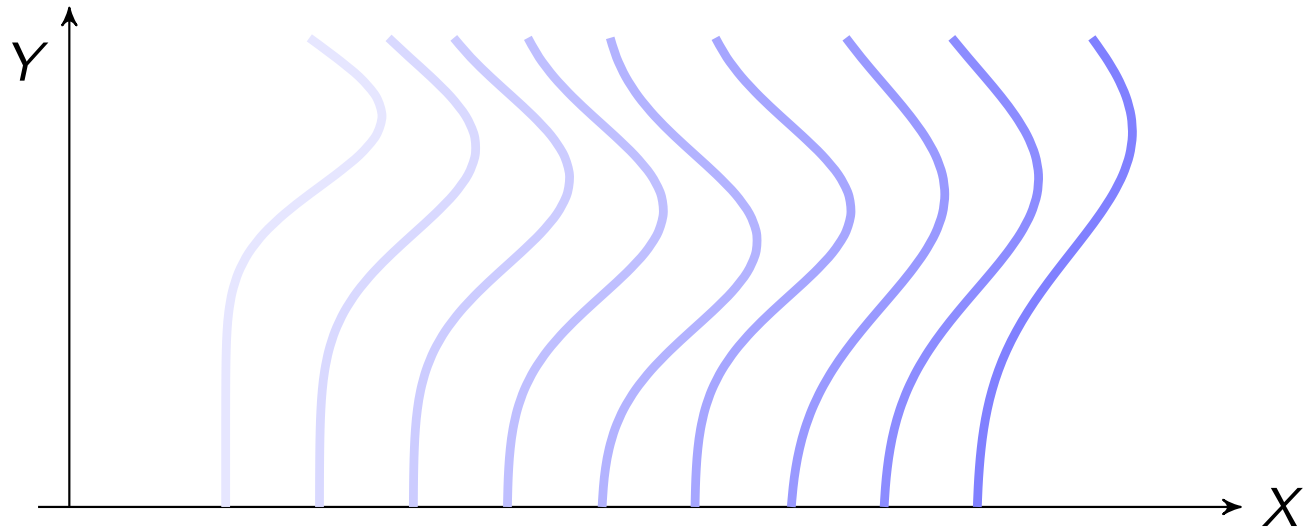
- Many tricks have been proposed to improve the GMMN:
  - Optimized kernels and feature extractors (Sutherland et al., 2017; Li et al., 2017a)
  - Gradient regularization (Binkowski et al., 2018; Arbel et al., 2018)
  - Repulsive loss (Wang et al., 2019)
  - Optimized witness points (Mehrjou et al., 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 $\mathbb{P}(Y|X)$



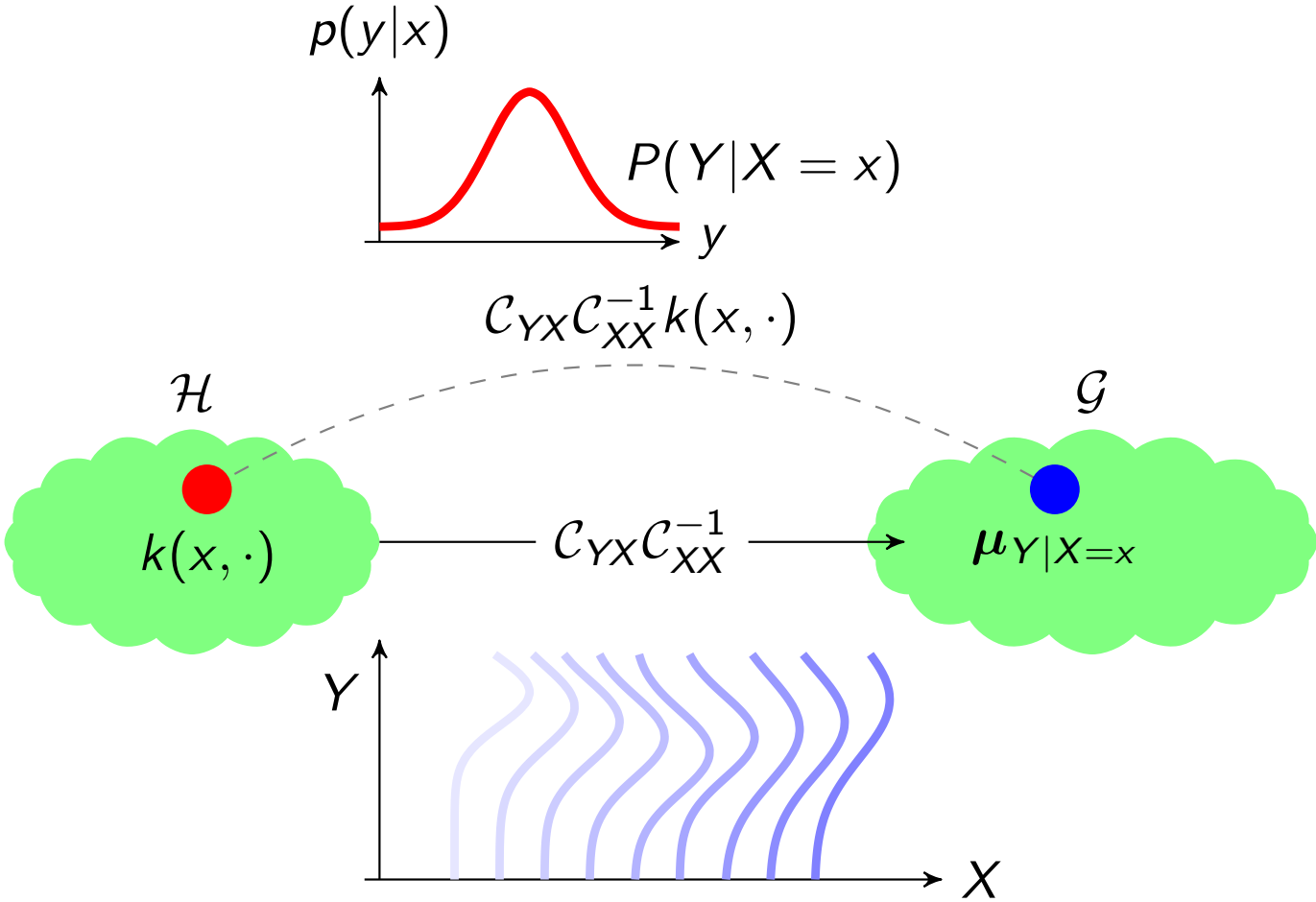
A collection of distributions $\mathscr{P}_Y := \{\mathbb{P}(Y|X = x) : x \in \mathcal{X}\}$.

- 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

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

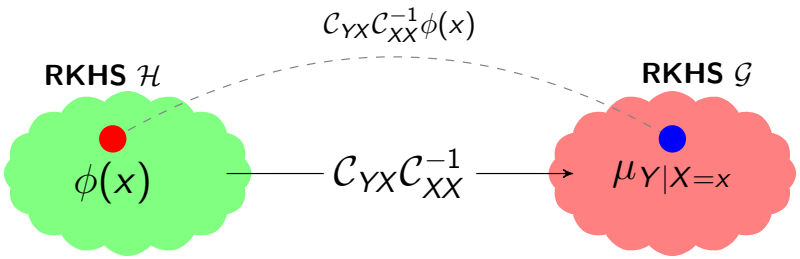
- In an infinite RKHS, $\mathcal{C}_{XX}^{-1}$ does not exists. Hence, we often use

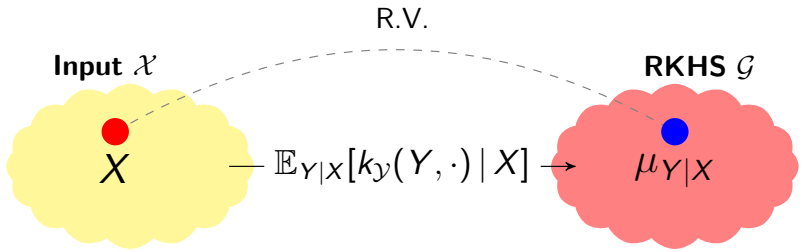
$$\mathcal{U}_{Y|X} := \mathcal{C}_{YX}(\mathcal{C}_{XX} + \varepsilon \mathcal{I})^{-1}.$$

- Conditional mean estimator

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

A Measure-Theoretic View





Operator-Based Approach

$$\mathcal{U}_{Y|X} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} : \mathcal{H} \to \mathcal{G}$$
$$\mathcal{U}_{Y|X} k_{\mathcal{X}}(x,\cdot) = \mathbb{E}[k_{\mathcal{Y}}(Y,\cdot)\,|\,X]$$

Assumptions

1. $\mathcal{C}_{XX}^{-1}$ exists.
2. $\mathbb{E}_{Y|X}[f(Y)\,|\,X=\cdot] \in \mathcal{H}$

Strong assumptions, not satisfied with commonly used kernels, e.g. the Gaussian kernel.

Measure-Theoretic Approach[1]

$$\mu_{Y|X} = \mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y,\cdot)\,|\,X]$$

- Bochner conditional expectation.
- No assumptions required.
- $\mu_{Y|X}$ is a X-measurable random variable with values in $\mathcal{H}_{\mathcal{Y}}$.
- Empirical estimation via natural and flexible regression set-up.

[1] Park and Muandet, "*A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings*". In NeurIPS 2020.

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 $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$, $\mathbf{L}_{ij} = l(y_i, y_j)$, 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

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

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

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

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.

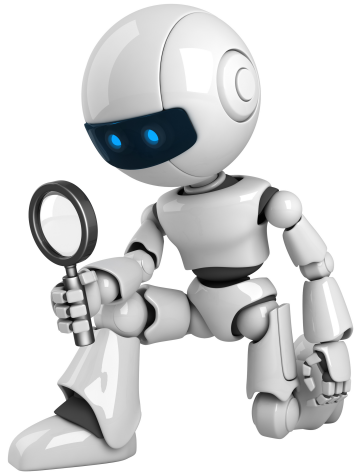
Recap on Kernel Methods

Kernel Mean Embedding of Distributions

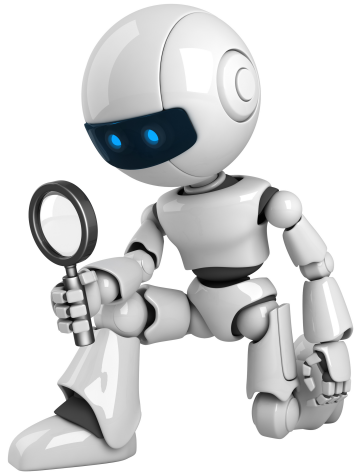
Conditional Mean Embedding

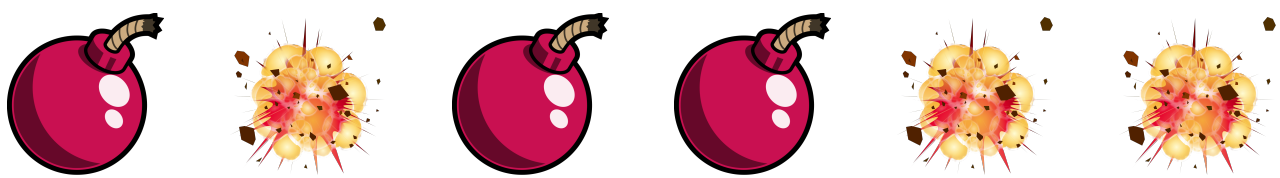
Deconfounded Causal Learning

Scenario I: Prediction

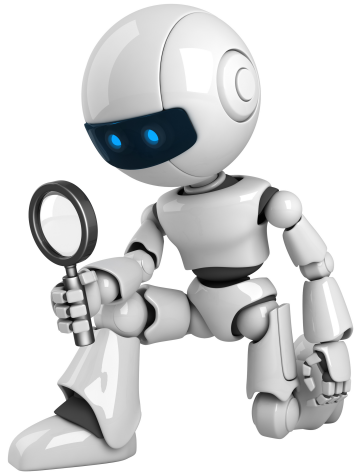


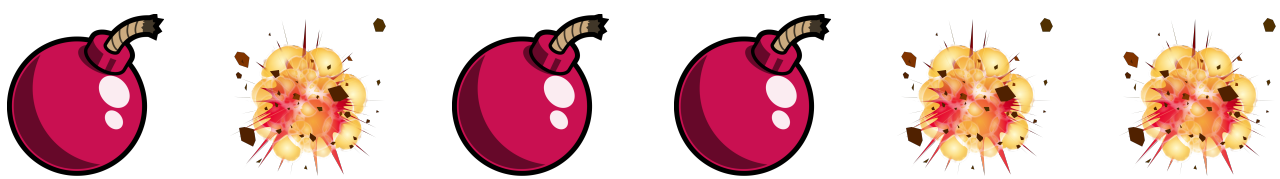
Scenario I: Prediction





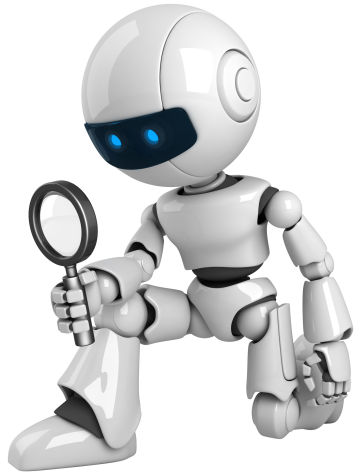
Scenario I: Prediction





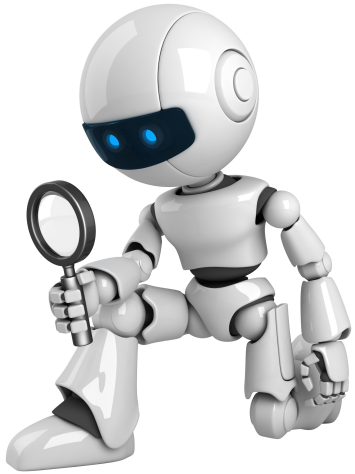
Scenario I: Prediction

That is easy!



Scenario I: Prediction

That is easy!



Scenario I: Supervised Learning

- Let $\mathcal{X}$ be an input space and $\mathcal{Y}$ be an output space.
- We observe an i.i.d. sample from $\mathbb{P}(X, Y)$:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}.$$

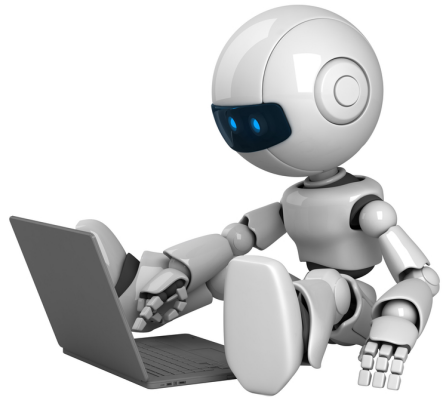
- Learn $f : \mathcal{X} \to \mathcal{Y}$ that generalizes to new data from $\mathbb{P}(X, Y)$.
  - Logistic regression, support vector machine, deep neural network, etc.
- Ultimately, we aim to estimate the **conditional mean**

$$f^*(x) = \mathbb{E}[Y \mid X = x].$$

- For classification problem, f^* is called the **Bayes classifier**.
- Nice theoretical guarantees via probably approximately correct (PAC) framework (Valiant 1984).

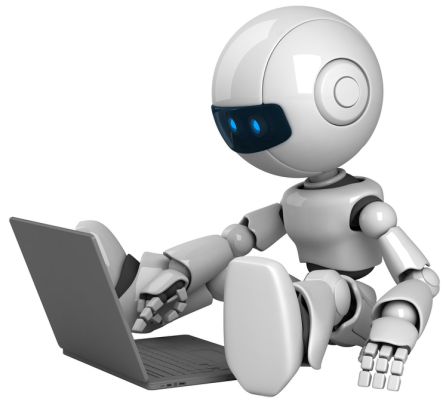
Scenario II: Intervention

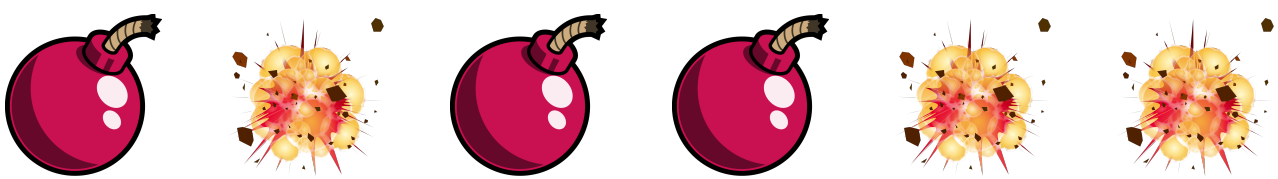
What if I press the **red** button?



Scenario II: Intervention

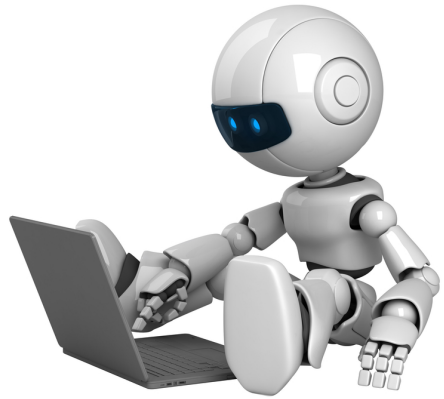
What if I press the **red** button?

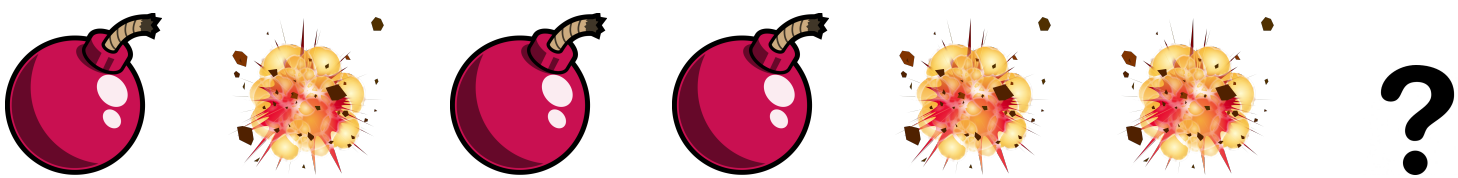




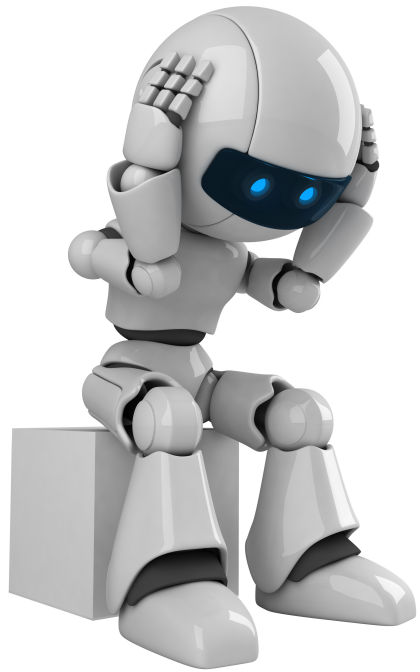
Scenario II: Intervention

What if I press the **red** button?



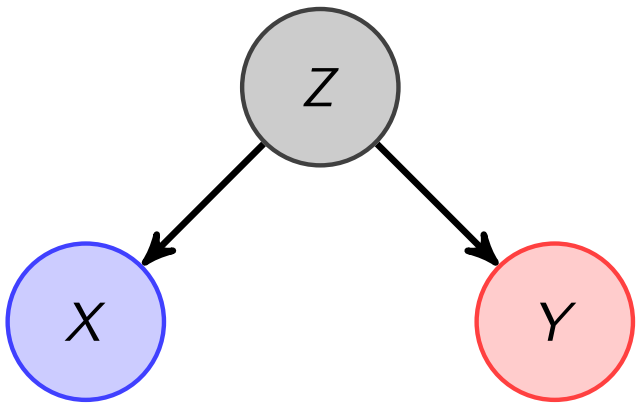


Scenario I $\neq$ Scenario II



Scenario I $\neq$ Scenario II

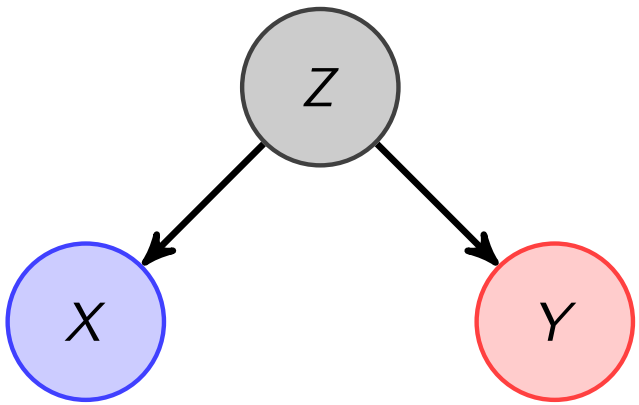
Pre-intervention



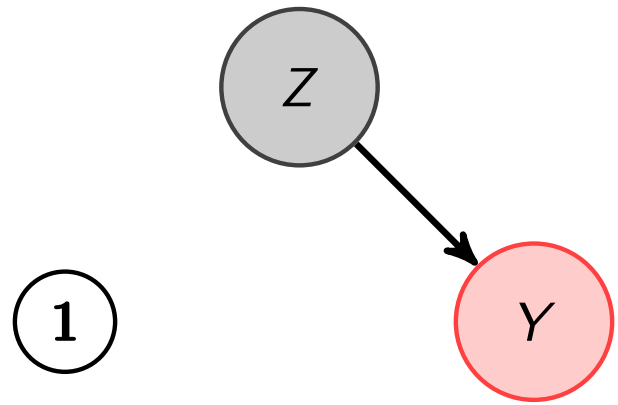
- Let X, Y, Z be binary random variables.
- If $Z = 0$, $X = 0$ (green button) and $Y = 0$ (no explosion).
- If $Z = 1$, $X = 1$ (red button) and $Y = 1$ (explosion).
- Let $Z = 1$ w.p. 0.5 (coin flip).

Scenario I $\neq$ Scenario II

Pre-intervention

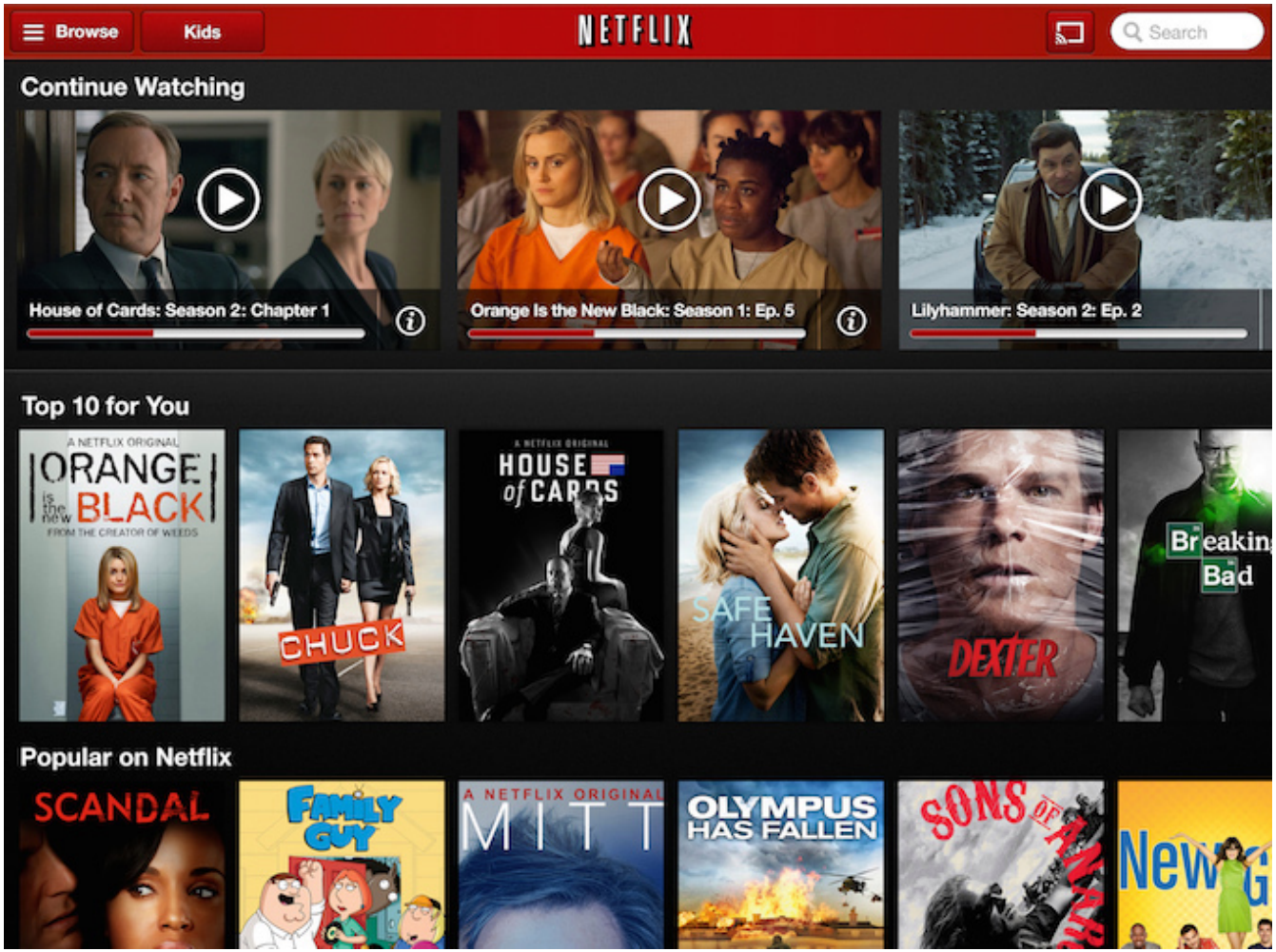


Post-intervention

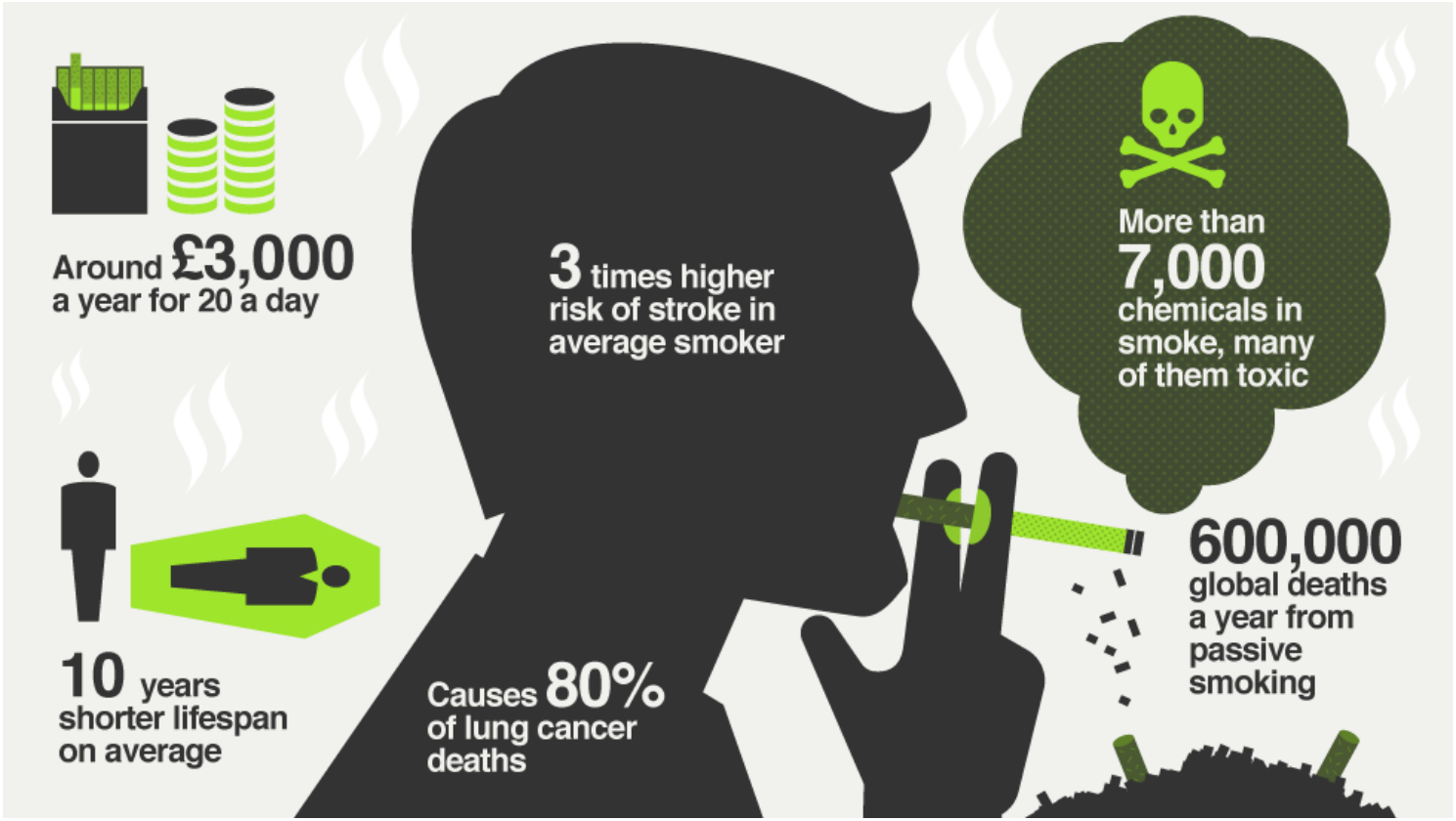


- Let X, Y, Z be binary random variables.
- If $Z = 0$, $X = 0$ (green button) and $Y = 0$ (no explosion).
- If $Z = 1$, $X = 1$ (red button) and $Y = 1$ (explosion).
- Let $Z = 1$ w.p. 0.5 (coin flip).

Example I : Recommendation Systems



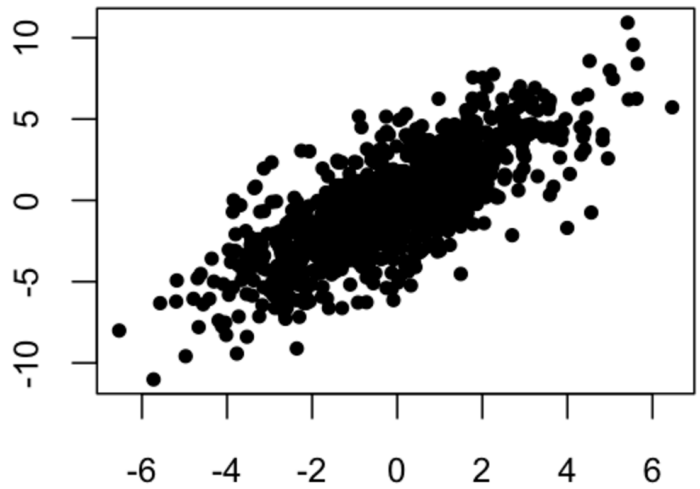
Example II : Healthcare



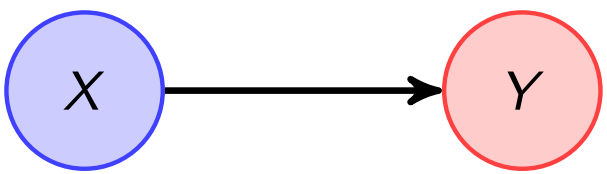
Example III : Public Policy

Causal Inference

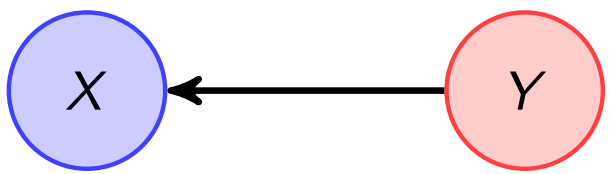
From observational data,



we are interested in distinguishing

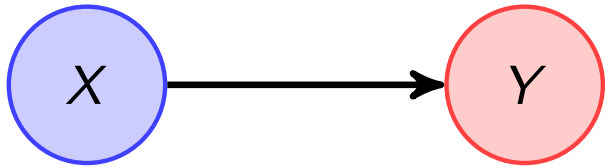


X **causes** Y



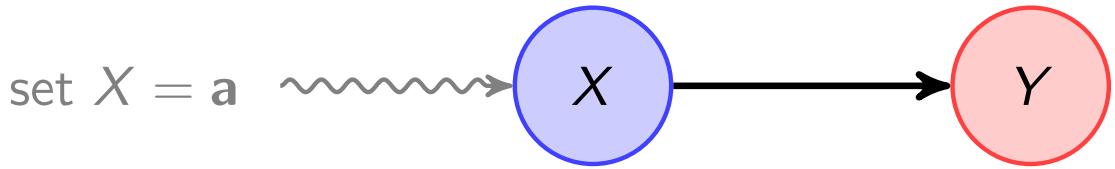
Y **causes** X

Intervention and Do Calculus



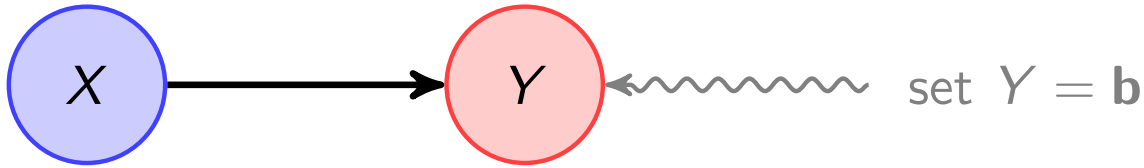
A variable X is said to **causes** another variable Y if the probability distribution of Y changes when we **intervene** on X.

Intervention and Do Calculus



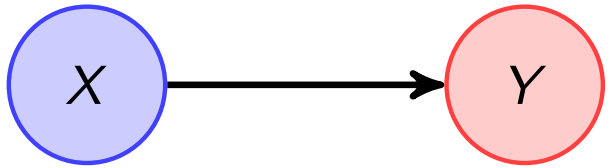
A variable X is said to **causes** another variable Y if the probability distribution of Y changes when we **intervene** on X.

Intervention and Do Calculus



A variable X is said to **causes** another variable Y if the probability distribution of Y changes when we **intervene** on X.

Intervention and Do Calculus



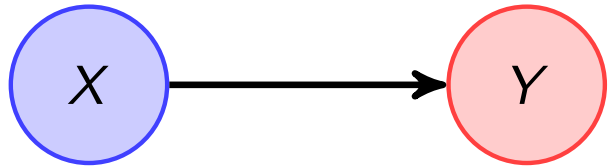
A variable X is said to **causes** another variable Y if the probability distribution of Y changes when we **intervene** on X.

- **Interventional distribution** : the distribution of Y if we **set** the value of X to x.

$$P(Y \mid \text{do}(X = x))$$

  - **Intervention**: pushing red button, knocking out genes, recommending movies, etc.

Intervention and Do Calculus



A variable X is said to **causes** another variable Y if the probability distribution of Y changes when we **intervene** on X.

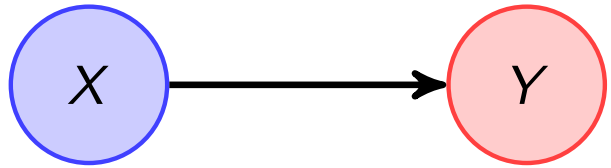
- **Interventional distribution** : the distribution of Y if we **set** the value of X to x.

$$P(Y \mid \text{do}(X = x))$$

  - **Intervention**: pushing red button, knocking out genes, recommending movies, etc.
- **Observational distribution** : the distribution of Y if we **observe** that the value of X is x.

$$P(Y \mid X = x)$$

Intervention and Do Calculus



A variable X is said to **causes** another variable Y if the probability distribution of Y changes when we **intervene** on X.

- **Interventional distribution** : the distribution of Y if we **set** the value of X to x.

$$P(Y \mid \text{do}(X = x))$$

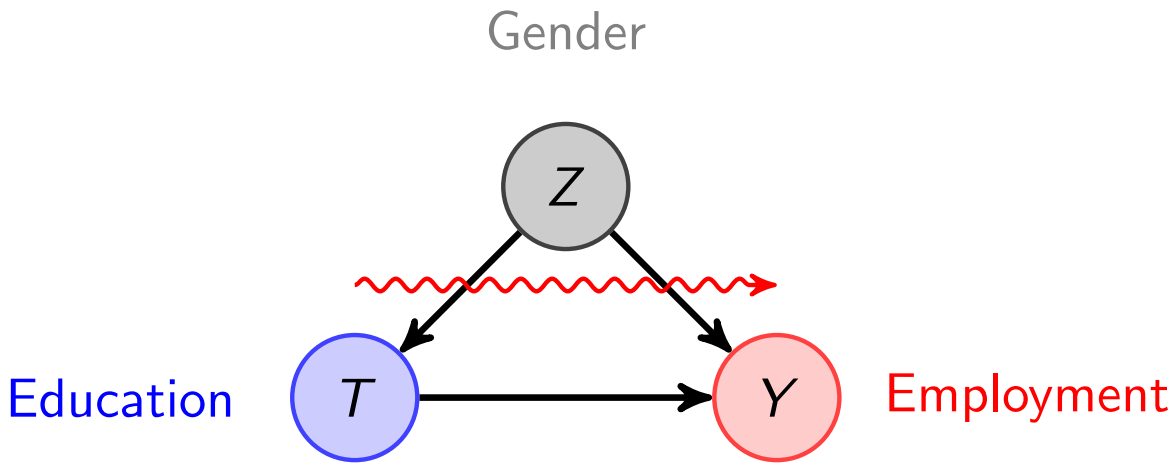
  - **Intervention**: pushing red button, knocking out genes, recommending movies, etc.
- **Observational distribution** : the distribution of Y if we **observe** that the value of X is x.

$$P(Y \mid X = x)$$

- $P(Y \mid \text{do}(X = x)) \neq P(Y \mid X = x)$.

Randomized Experiments

- A **randomized experiment**, pioneered by R. A. Fisher (1890–1962), is a gold standard for causal inference.
  - population is split into 2 groups by **random** lot.
  - assign **treatments** $T \in \{0, 1\}$ to each group.
  - observe **outcomes** Y_0 and Y_1 for both groups.
  - assess the causal effect $Y_0 - Y_1$.
- A randomization breaks the potential of unobserved confounders.



- Unfortunately, randomization is not always possible.
- We must rely on an **observational studies**.

Potential Outcome Framework

- The treatment is not assigned randomly.
- $\mathcal{T}$: the space of all possible treatments, $\mathcal{X}$: the space of covariates, $\mathcal{Y}$: the space of potential outcomes.
  1. Select the treatment group and control group using some covariates $X \in \mathcal{X}$.
  2. Assign the treatment T_k to the subject j with covariate $X_j \in \mathcal{X}$
  3. Observe the outcome $Y_k(j) \in \mathcal{Y}$ of the subject j
- **Potential outcomes**: $Y_0(j), Y_1(j)$
- Fundamental problem of causal inference (FPCI)

Treatment (take a pill)

Control (no pill)

Fundamental Problem of Causal Inference (FPCI)

- Consider $\mathcal{T} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$.

Subject	$Y_0(i)$	$Y_1(i)$	$Y_0(i) - Y_1(i)$
A	0	?	?
B	?	1	?
C	0	?	?
D	?	1	?

- ? is the **counterfactual** quantity.
- Individual treatment effect (ITE): $Y_0(i) - Y_1(i)$
- Average causal effect (ACE):

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$$

- Average treatment effects on the treated (ATET) :

$$\mathbb{E}[Y_0(i) \mid T_i = 1] - \mathbb{E}[Y_1(i) \mid T_i = 1]$$

Counterfactual Learning

- Recent studies by, e.g., Bottou et al. (2013), Johansson et al. (2016), Swaminathan and Joachims (2015).
- We receive an i.i.d. sample drawn according to $\mathbb{P}(X, T, Y)$

$$(x_1, t_1, y_1), (x_2, t_2, y_2), \ldots, (x_n, t_n, y_n) \in \mathcal{X} \times \mathcal{T} \times \mathcal{Y}.$$

- Learn $h : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ to estimate the causal effect (Hill 2011).
- We have no control over the **treatment assignment mechanism** in observational studies.
- In some applications, the **treatment assignment mechanism** is under our control.
  - Search engines, recommendation systems, ad placement, etc.
  - A counterfactual estimator of a system's performance under an alternative prediction.
- The learning involves the **counterfactual distribution**.

Assumption

Assumption I

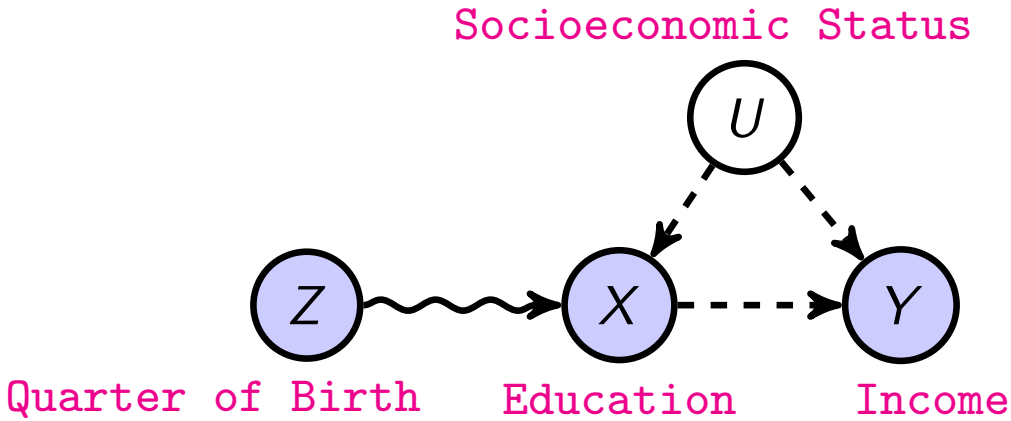
1. **Stable unit treatment value assumption (SUTVA)**: the outcome of subject i is independent of the outcomes of other subjects and their received treatments.
2. **Unconfoundedness/ignorability/exogeneity**:

$$Y_0, Y_1 \perp\!\!\!\perp T|X$$

3. **Common support assumption**:

$$\mathcal{X}_k \subseteq \mathcal{X}_j, \quad j, k = 1, 2, \ldots$$

Instrumental Variable Regression



Assumptions

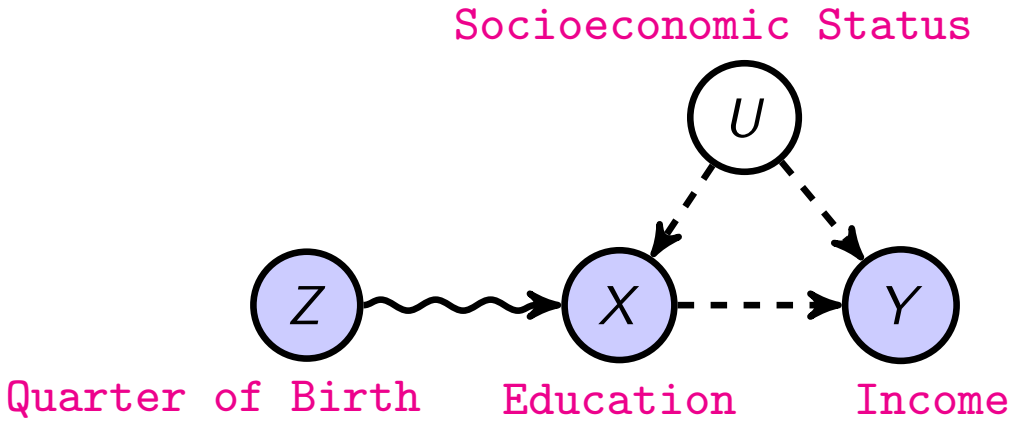
1. Relevance: $Z \to X$
2. Exclusion restriction:

$$Z \to X \to Y$$

3. Exchangeability: $Z \perp\!\!\!\perp U$

[1] Angrist, Joshua, D., and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." Journal of Economic Perspectives, 15 (4): 69-85.

Instrumental Variable Regression

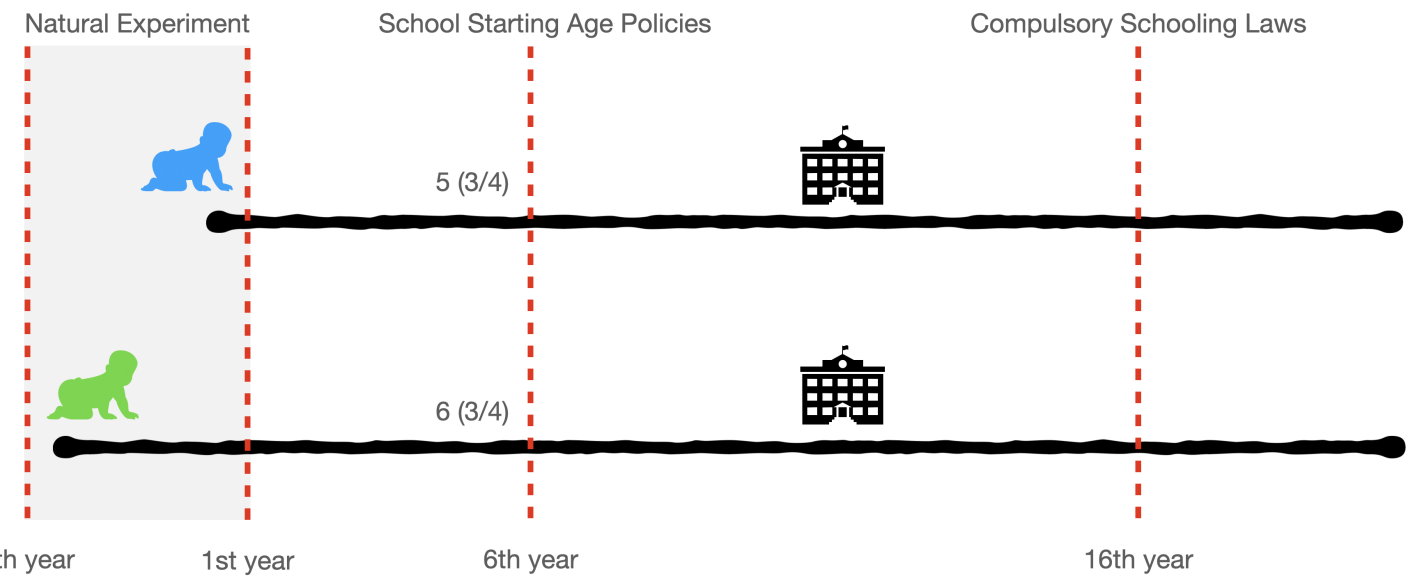


Assumptions

1. Relevance: $Z \to X$
2. Exclusion restriction:

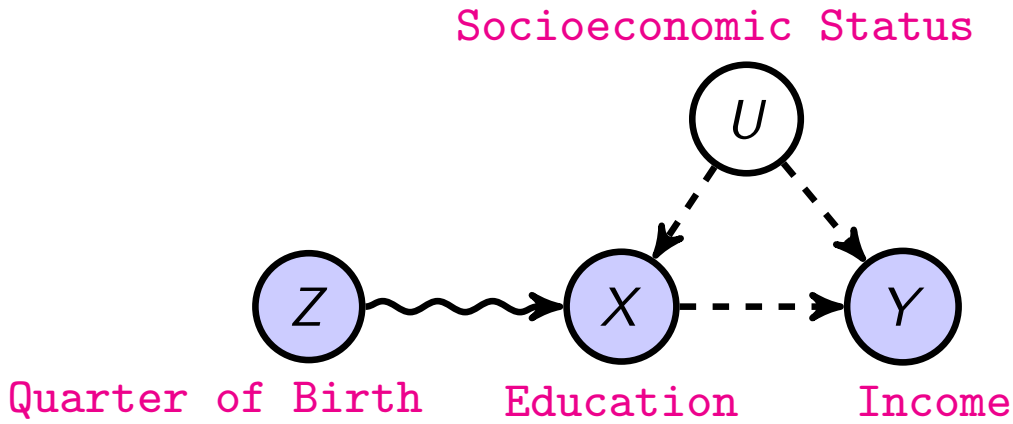
$$Z \to X \to Y$$

3. Exchangeability: $Z \perp\!\!\!\perp U$



[1] Angrist, Joshua, D., and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." Journal of Economic Perspectives, 15 (4): 69-85.

Instrumental Variable Regression



Assumptions

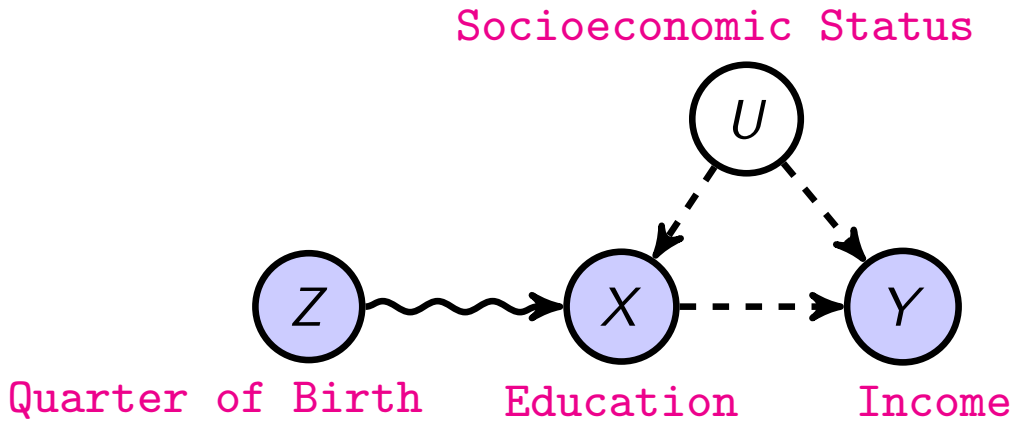
1. Relevance: $Z \to X$
2. Exclusion restriction:

$$Z \to X \to Y$$

3. Exchangeability: $Z \perp\!\!\!\perp U$

[1] Newey and Powell. Instrumental Variable Estimation of Nonparametric Models. Econometrica Vol. 71, No. 5 (Sep., 2003), pp. 1565-1578.

Instrumental Variable Regression



Assumptions

1. Relevance: $Z \to X$
2. Exclusion restriction:

$$Z \to X \to Y$$

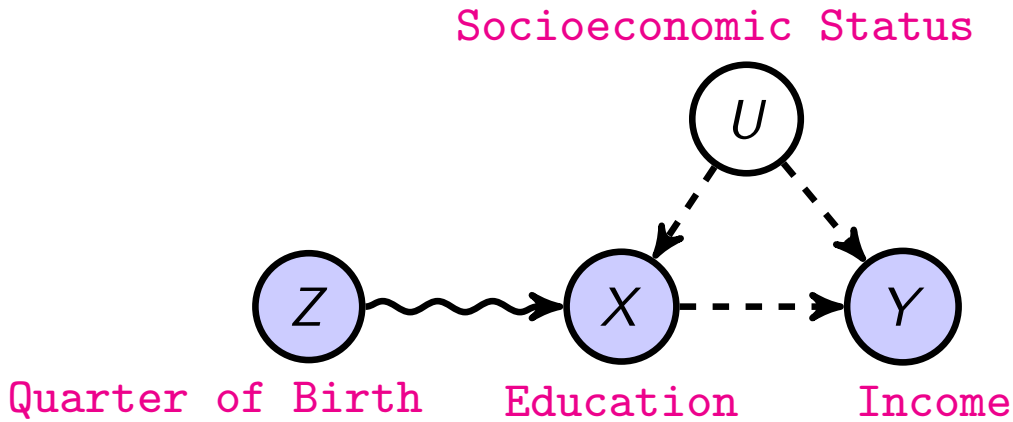
3. Exchangeability: $Z \perp\!\!\!\perp U$

- The structural model takes the form

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

[1] Newey and Powell. Instrumental Variable Estimation of Nonparametric Models. Econometrica Vol. 71, No. 5 (Sep., 2003), pp. 1565-1578.

Instrumental Variable Regression



Assumptions

1. Relevance: $Z \to X$
2. Exclusion restriction:

$$Z \to X \to Y$$

3. Exchangeability: $Z \perp\!\!\!\perp U$

- The structural model takes the form

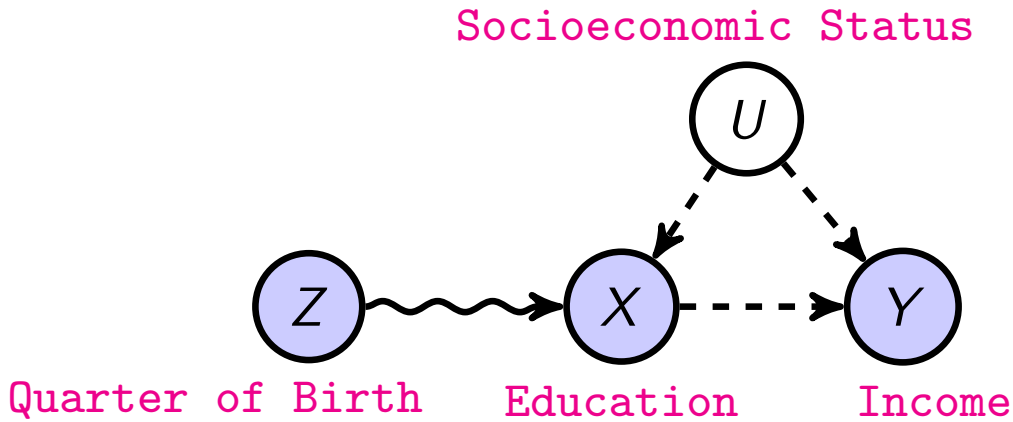
$$Y = f(X) + \varepsilon, \quad \mathbb{E}[\varepsilon] = 0, \quad \mathbb{E}[\varepsilon \mid X] \neq 0$$

- Using the instrumental variable (IV) Z, we have

$$\begin{aligned} \mathbb{E}[Y \mid Z] &= \mathbb{E}[f(X) \mid Z] + \mathbb{E}[\varepsilon \mid Z] \\ &= \mathbb{E}[f(X) \mid Z] \end{aligned}$$

[1] Newey and Powell. Instrumental Variable Estimation of Nonparametric Models. Econometrica Vol. 71, No. 5 (Sep., 2003), pp. 1565-1578.

Instrumental Variable Regression



Assumptions

1. Relevance: $Z \to X$
2. Exclusion restriction:

$$Z \to X \to Y$$

3. Exchangeability: $Z \perp\!\!\!\perp U$

- The structural model takes the form

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

- Using the instrumental variable (IV) Z, we have

$$\begin{aligned} \mathbb{E}[Y \mid Z] &= \mathbb{E}[f(X) \mid Z] + \mathbb{E}[\varepsilon \mid Z] \\ &= \mathbb{E}[f(X) \mid Z] \end{aligned}$$

- **Solutions**: (1) $\mathbb{E}[Y \mid Z] = \mathbb{E}[f(X) \mid Z]$ (2) $\mathbb{E}[Y - f(X) \mid Z] = 0$

[1] Newey and Powell. Instrumental Variable Estimation of Nonparametric Models. Econometrica Vol. 71, No. 5 (Sep., 2003), pp. 1565-1578.

Kernel Instrumental Variable Regression

- Assume that f belongs to the RKHS $\mathcal{F}$:

$$\begin{aligned}\mathbb{E}[Y \mid z] &= \mathbb{E}[f(X) \mid z] = \mathbb{E}[\langle f, k(X, \cdot)\rangle_{\mathcal{F}} \mid z] \\ &= \langle f, \mathbb{E}[k(X, \cdot) \mid z]\rangle_{\mathcal{F}} = \langle f, \mu_{X|z}\rangle_{\mathcal{F}}.\end{aligned}$$

[1] Singh, Sahani, and Gretton. Kernel Instrumental Variable Regression. In NeurIPS 2019.

Kernel Instrumental Variable Regression

- Assume that f belongs to the RKHS $\mathcal{F}$:

$$\begin{aligned}\mathbb{E}[Y \mid z] &= \mathbb{E}[f(X) \mid z] = \mathbb{E}[\langle f, k(X, \cdot)\rangle_{\mathcal{F}} \mid z] \\ &= \langle f, \mathbb{E}[k(X, \cdot) \mid z]\rangle_{\mathcal{F}} = \langle f, \mu_{X|z}\rangle_{\mathcal{F}}.\end{aligned}$$

- Given data $\{(x_i, y_i, z_i)\}_{i=1}^{n_1+n_2}$, the KIV proceeds in two steps:
  1. Estimate the conditional mean embedding $\mu_{X|Z}$:

$$\hat{\mu}_{X|z} = \sum_{i=1}^{n_1} \beta_i(z) k(x_i, \cdot), \quad \boldsymbol{\beta}(z) := (K_{ZZ} + n_1\lambda_1 I)^{-1} K_{Zz}$$

  2. Solve a vector-valued kernel ridge regression:

$$\hat{R}_\lambda(f) = \frac{1}{n_2} \sum_{i=1}^{n_2} \left(y_i - \langle f, \hat{\mu}_{X|z_i}\rangle_{\mathcal{F}}\right)^2 + \lambda_2 \|f\|_{\mathcal{F}}^2$$

[1] Singh, Sahani, and Gretton. Kernel Instrumental Variable Regression. In NeurIPS 2019.

Kernel Instrumental Variable Regression

- Assume that f belongs to the RKHS $\mathcal{F}$:

$$\begin{aligned}\mathbb{E}[Y \mid z] &= \mathbb{E}[f(X) \mid z] = \mathbb{E}[\langle f, k(X, \cdot)\rangle_{\mathcal{F}} \mid z] \\ &= \langle f, \mathbb{E}[k(X, \cdot) \mid z]\rangle_{\mathcal{F}} = \langle f, \mu_{X|z}\rangle_{\mathcal{F}}.\end{aligned}$$

- Given data $\{(x_i, y_i, z_i)\}_{i=1}^{n_1+n_2}$, the KIV proceeds in two steps:
  1. Estimate the conditional mean embedding $\mu_{X|Z}$:

$$\hat{\mu}_{X|z} = \sum_{i=1}^{n_1} \beta_i(z) k(x_i, \cdot), \quad \boldsymbol{\beta}(z) := (K_{ZZ} + n_1\lambda_1 I)^{-1} K_{Zz}$$

  2. Solve a vector-valued kernel ridge regression:

$$\hat{R}_\lambda(f) = \frac{1}{n_2} \sum_{i=1}^{n_2} \left(y_i - \langle f, \hat{\mu}_{X|z_i}\rangle_{\mathcal{F}}\right)^2 + \lambda_2 \|f\|_{\mathcal{F}}^2$$

- The solution $\hat{f}$ can be expressed as $\hat{f} = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$ where

$$\boldsymbol{\alpha} := (WW^\top + n_2\lambda_2 K_{XX})^{-1} W\mathbf{y}, \quad W = K_{XX}(K_{ZZ} + n_1\lambda_1 I)^{-1} K_{Z\tilde{z}}$$

[1] Singh, Sahani, and Gretton. Kernel Instrumental Variable Regression. In NeurIPS 2019.

Conditional Moment Restriction

Newey (1993), Ai and Chen (2003)

- Recall that $\mathbb{E}[\varepsilon \mid Z] = 0$, i.e.,

$$\mathbb{E}[Y - f(X) \mid Z] = 0, \quad P_Z - \text{ a.s.}$$

  which is known as a **conditional moment restriction** (CMR).
- The CMR implies unconditional moment restrictions:

$$\mathbb{E}[(Y - f(X))h(Z)] = 0, \quad \forall h,$$

  where h is a real-valued measurable function.
- Given the instruments $h_1, \ldots, h_m$, one can use the **generalized method of moment (GMM)** to learn the function f.

Maximum Moment Restriction (MMR)

Muandet, Jitkrittum, Kübler, **UAI** 2020

Let $\mathcal{H}$ be a space of instruments $h(z)$.

$$\mathbb{E}[Y - f(X) \mid Z] = 0 \quad \Leftrightarrow \quad \sup_{h \in \mathcal{H}, \|h\| \leq 1} |\mathbb{E}[(Y - f(X))h(Z)]| = 0$$

Maximum Moment Restriction (MMR)

Muandet, Jitkrittum, Kübler, **UAI** 2020

Let $\mathcal{H}$ be a space of instruments $h(z)$.

$$\mathbb{E}[Y - f(X) \,|\, Z] = 0 \quad \Leftrightarrow \quad \sup_{h \in \mathcal{H}, \|h\| \leq 1} |\mathbb{E}[(Y - f(X))h(Z)]| = 0$$

- The equivalence above holds if $\mathcal{H}$ is a **universal** RKHS.

Maximum Moment Restriction (MMR)

Muandet, Jitkrittum, Kübler, **UAI** 2020

Let $\mathcal{H}$ be a space of instruments $h(z)$.

$$\mathbb{E}[Y - f(X) \,|\, Z] = 0 \quad \Leftrightarrow \quad \sup_{h \in \mathcal{H}, \|h\| \leq 1} |\mathbb{E}[(Y - f(X))h(Z)]| = 0$$

- The equivalence above holds if $\mathcal{H}$ is a **universal** RKHS.
- Let $\boldsymbol{\mu}_f := \mathbb{E}[(Y - f(X))k(Z, \cdot)]$.

$$\begin{aligned}
\mathrm{MMR}(\mathcal{H}, f) &:= \sup_{h \in \mathcal{H}, \|h\| \leq 1} |\mathbb{E}\left[(Y - f(X))h(Z)\right]| \\
&= \sup_{h \in \mathcal{H}, \|h\| \leq 1} |\mathbb{E}\left[\langle h, (Y - f(X))k(Z, \cdot)\rangle_{\mathcal{H}}\right]| \\
&= \sup_{h \in \mathcal{H}, \|h\| \leq 1} |\langle h, \mathbb{E}\left[(Y - f(X))k(Z, \cdot)\right]\rangle_{\mathcal{H}}| \\
&= \|\mathbb{E}[(Y - f(X))k(Z, \cdot)]\|_{\mathcal{H}} = \|\boldsymbol{\mu}_f\|_{\mathcal{H}}.
\end{aligned}$$

Maximum Moment Restriction (MMR)

Muandet, Jitkrittum, Kübler, **UAI** 2020

Let $\mathcal{H}$ be a space of instruments $h(z)$.

$$\mathbb{E}[Y - f(X) \,|\, Z] = 0 \quad \Leftrightarrow \quad \sup_{h \in \mathcal{H}, \|h\| \leq 1} |\mathbb{E}[(Y - f(X))h(Z)]| = 0$$

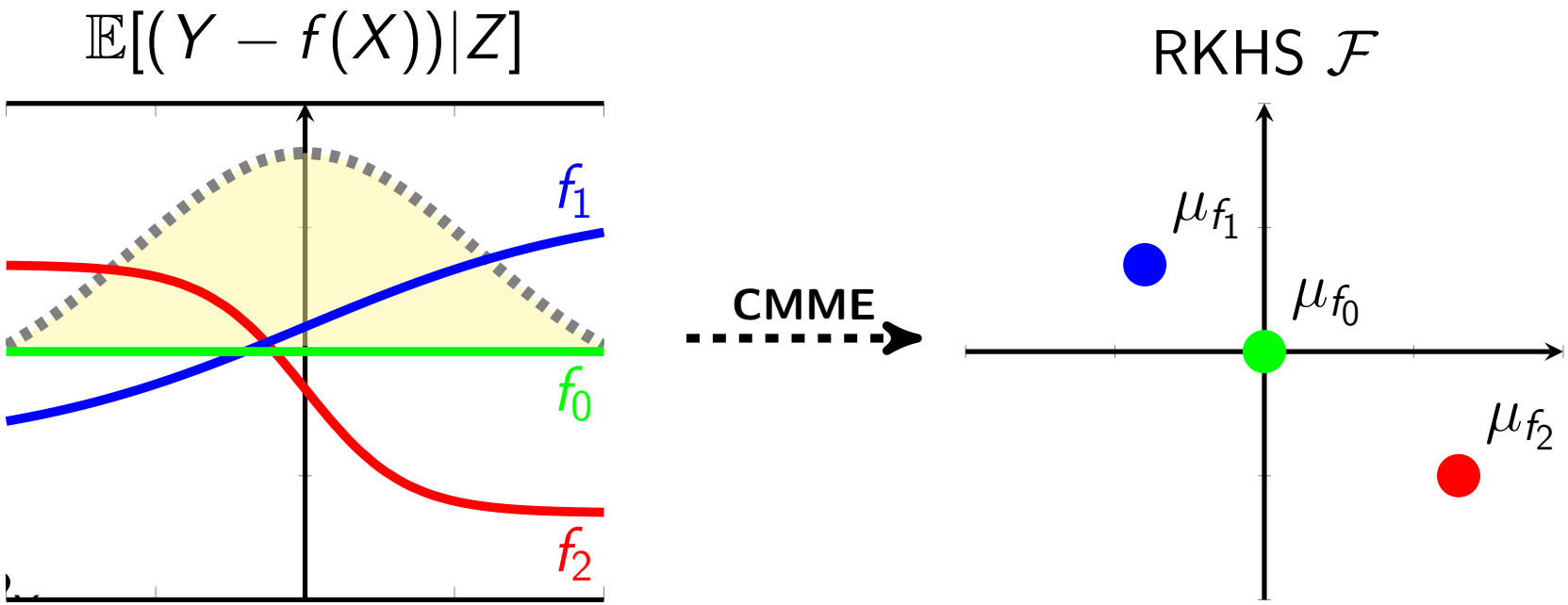
- The equivalence above holds if $\mathcal{H}$ is a **universal** RKHS.
- Let $\boldsymbol{\mu}_f := \mathbb{E}[(Y - f(X))k(Z, \cdot)]$.

$$\begin{aligned} \text{MMR}(\mathcal{H}, f) &:= \sup_{h \in \mathcal{H}, \|h\| \leq 1} |\mathbb{E}\left[(Y - f(X))h(Z)\right]| \\ &= \sup_{h \in \mathcal{H}, \|h\| \leq 1} |\mathbb{E}\left[\langle h, (Y - f(X))k(Z, \cdot)\rangle_{\mathcal{H}}\right]| \\ &= \sup_{h \in \mathcal{H}, \|h\| \leq 1} |\langle h, \mathbb{E}\left[(Y - f(X))k(Z, \cdot)\right]\rangle_{\mathcal{H}}| \\ &= \|\mathbb{E}[(Y - f(X))k(Z, \cdot)]\|_{\mathcal{H}} = \|\boldsymbol{\mu}_f\|_{\mathcal{H}}. \end{aligned}$$

- $\text{MMR}^2(\mathcal{H}, f) = \mathbb{E}[(Y - f(X))k(Z, Z')(Y' - f(X'))]$.

Conditional Moment Embeddings

$$\mu_f := \mathbb{E}[(Y - f(X))k(Z, \cdot)]$$



Instrumental Variable Regression

$$\boxed{\mathbb{E}[Y - f(X) \mid Z] = 0, \text{ a.s. } P_Z.}$$

Instrumental Variable Regression

$$\boxed{\mathbb{E}[Y - f(X) \,|\, Z] = 0, \text{ a.s. } P_Z.}$$

Then, we can write the MMR objective as

$$\text{MMR}^2(\mathcal{H}, f) = \mathbb{E}[(Y - f(X))(Y' - f(X'))k(Z, Z')].$$

Instrumental Variable Regression

$$\boxed{\mathbb{E}[Y - f(X) \,|\, Z] = 0, \text{ a.s. } P_Z.}$$

Then, we can write the MMR objective as

$$\mathrm{MMR}^2(\mathcal{H}, f) = \mathbb{E}[(Y - f(X))(Y' - f(X'))k(Z, Z')].$$

Regularized MMR

Given a sample $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$,

$$\begin{aligned} R_\lambda(f) &= \widehat{\mathrm{MMR}}^2(\mathcal{H}, f) + \lambda \|f\|^2 \\ &= \frac{1}{n^2} \sum_{i,j=1}^{n} (y_i - f(x_i))(y_j - f(x_j))k(z_i, z_j) + \lambda \|f\|^2. \end{aligned}$$

MMR-IV Algorithms

Neural Network

Let $\mathcal{F}_\Theta$ be a class of deep neural networks:

$$f_{\text{NN}} = \arg\min_{f_\theta \in \mathcal{F}_\Theta} \frac{1}{n^2} \sum_{i,j=1}^{n} (y_i - f_\theta(x_i))(y_j - f_\theta(x_j))k(z_i, z_j) + \lambda \|\theta\|^2.$$

This can be solved using standard SGD.

MMR-IV Algorithms

Neural Network

Let $\mathcal{F}_\Theta$ be a class of deep neural networks:

$$f_{\mathrm{NN}} = \arg\min_{f_\theta \in \mathcal{F}_\Theta} \frac{1}{n^2} \sum_{i,j=1}^{n} (y_i - f_\theta(x_i))(y_j - f_\theta(x_j)) k(z_i, z_j) + \lambda \|\theta\|^2.$$

This can be solved using standard SGD.

RKHS

Let $\mathcal{F}_l$ be the RKHS with the reproducing kernel $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$$\begin{aligned} f_{\mathrm{RKHS}} &= \arg\min_{f \in \mathcal{F}_l} \frac{1}{n^2} \sum_{i,j=1}^{n} (y_i - f\theta(x_i))(y_j - f(x_j)) k(z_i, z_j) + \lambda \|f\|^2_{\mathcal{F}_l} \\ &= \sum_{i=1}^{n} \alpha_i l(x_i, \cdot), \qquad \boldsymbol{\alpha} = (LWL + \lambda L)^{-1} LW \boldsymbol{y}. \end{aligned}$$

We use Nyström approximation to reduce the time complexity.

Asymptotic Normality

Parametric regime

Under some technical assumptions, we have that

$$\sqrt{n}(\hat{\theta}_V - \theta^*) \rightsquigarrow N(\mathbf{0}, \Sigma_V)$$

holds, where $\Sigma_V = 4H^{-1}\mathrm{diag}(\mathbb{E}_U[\mathbb{E}^2_{U'}[h_{\theta^*}(U, U')]])H^{-1}$ and $\rightsquigarrow$ denotes a convergence in law.

Asymptotic Normality

Parametric regime

Under some technical assumptions, we have that

$$\sqrt{n}(\hat{\theta}_V - \theta^*) \rightsquigarrow N(\mathbf{0}, \Sigma_V)$$

holds, where $\Sigma_V = 4H^{-1}\mathrm{diag}(\mathbb{E}_U[\mathbb{E}^2_{U'}[h_{\theta^*}(U, U')]])H^{-1}$ and $\rightsquigarrow$ denotes a convergence in law.

Nonparametric regime

If there exists $s \in (0, 2)$ and a constant $C_H > 0$ such that

$$\log \mathcal{N}(\varepsilon, \mathcal{H}_I, \|\cdot\|_{L^\infty}) \leq C_H \varepsilon^{-s}$$

for any $\varepsilon \in (0, 1)$. Then, there exists a Gaussian process $\mathbb{G}^*_P$ such that

$$\sqrt{n}(\hat{f}_V - f^*_{\lambda_0}) \rightsquigarrow \mathbb{G}^*_P \text{ in } \mathcal{H}_I$$

.

Experimental Results

Low-Dimensional Scenario

$$\textbf{DGP:} \quad Y = f^*(X) + e + \delta, \qquad X = Z_1 + e + \gamma,$$

$$Z := (Z_1, Z_2) \sim \mathrm{Uniform}([-3, 3]^2),\ e \sim \mathcal{N}(0, 1),\ \text{and}\ \gamma, \delta \sim \mathcal{N}(0, 0.1^2).$$

Experimental Results

Low-Dimensional Scenario

$$\textbf{DGP:} \quad Y = f^*(X) + e + \delta, \qquad X = Z_1 + e + \gamma,$$

$$Z := (Z_1, Z_2) \sim \text{Uniform}([-3, 3]^2),\ e \sim \mathcal{N}(0, 1),\ \text{and}\ \gamma, \delta \sim \mathcal{N}(0, 0.1^2).$$

Table: The MSE $\pm$ one standard deviation with $n = 2000$.

Algorithm	**True Function** f^*			
	abs	linear	sin	step
`DirectNN`	.116 ± .000	.035 ± .000	.189 ± .000	.199 ± .000
`2SLS`	.522 ± .000	**.000 ± .000**	.254 ± .000	.050 ± .000
`Poly2SLS`	.083 ± .000	**.000 ± .000**	.133 ± .000	.039 ± .000
`GMM+NN`	.318 ± .000	.044 ± .000	.694 ± .000	.500 ± .000
`AGMM`	.600 ± .001	.025 ± .000	.274 ± .000	.047 ± .000
`DeepIV`	.247 ± .004	.056 ± .003	.165 ± .003	.038 ± .001
`DeepGMM`	.027 ± .009	.005 ± .001	.160 ± .025	*.025 ± .006*
`KernelIV`	*.019 ± .000*	.009 ± .000	*.046 ± .000*	.026 ± .000
`AGMM-K`	181 ± .000	2.34 ± .000	19.4 ± .000	4.13 ± .000
`MMR-IV (NN)`	**.011 ± .002**	.005 ± .000	.153 ± .019	.040 ± .004
`MMR-IV (Nys)`	**.011 ± .001**	*.001 ± .000*	**.006 ± .002**	**.020 ± .002**

Experimental Results

High-Dimensional Scenario

We generate data using the same DGP:

$$\textbf{DGP:} \quad Y = f^*(X) + e + \delta, \qquad X = Z_1 + e + \gamma,$$

but map Z and X to high-dimensional MNIST images, i.e.,

1. **MNIST$_{\mathbf{Z}}$**: $Z \leftarrow \mathrm{RI}(\pi(Z_1^{\mathrm{low}}))$
2. **MNIST$_{\mathbf{X}}$**: $X \leftarrow \mathrm{RI}(\pi(X^{\mathrm{low}}))$
3. **MNIST$_{\mathbf{XZ}}$**: $X \leftarrow \mathrm{RI}(\pi(X^{\mathrm{low}}))$, $Z \leftarrow \mathrm{RI}(\pi(Z_1^{\mathrm{low}}))$

where

$$\pi(u) := \mathrm{round}(\min(\max(1.5u + 5, 0), 9))$$

is a function mapping inputs to an integer between 0 and 9.

Experimental Results

High-Dimensional Scenario

Table: The mean square error (MSE) $\pm$ one standard deviation on high-dimensional structured data. We run each method 10 times.

Algorithm	**Setting** MNIST_z	MNIST_x	MNIST_xz
`DirectNN`	.134 $\pm$ .000	*.229 $\pm$ .000*	*.196 $\pm$ .011*
`2SLS`	.563 $\pm$ .001	$>$1000	$>$1000
`Ridge2SLS`	.567 $\pm$ .000	.431 $\pm$ .000	.705 $\pm$ .000
`GMM+NN`	.121 $\pm$ .004	.235 $\pm$ .002	.240 $\pm$ .016
`AGMM`	*.017 $\pm$ .007*	.732 $\pm$ .107	.529 $\pm$.163
`DeepIV`	.114 $\pm$ .005	n/a	n/a
`DeepGMM`	.038 $\pm$ .004	.315 $\pm$ .130	.333 $\pm$ .168
`AGMM-K+NN`	.021 $\pm$ .007	1.05 $\pm$ .366	.327 $\pm$ .192
`MMR-IV (NN)`	.024 $\pm$ .006	**.124 $\pm$ .021**	**.130 $\pm$ .009**
`MMR-IV (Nys)`	**.015 $\pm$ .002**	.442 $\pm$ .000	.425 $\pm$ .002

Mendelian Randomization

Mendelian randomization relies on **genetic variants** as instruments.

$$Y = \beta X + c_1 e + \delta, \quad X = \sum_{i=1}^{m} \alpha_i Z_i + c_2 e + \gamma, \tag{1}$$

- $Z \in \mathbb{R}^{d'}$ with each entry $Z_i \sim B(2, p_i)$,
- $p_i \sim \mathrm{unif}(0.1, 0.9)$
- $e \sim \mathcal{N}(0, 1)$
- $\alpha_i \sim \mathrm{unif}([0.8/d', 1.2/d'])$
- $\gamma, \delta \sim \mathcal{N}(0, 0.1^2)$.
- We set $\alpha_i \sim \mathrm{unif}([0.8/d', 1.2/d'])$.

Mendelian Randomization

Mendelian randomization relies on **genetic variants** as instruments.

$$Y = \beta X + c_1 e + \delta, \quad X = \sum_{i=1}^{m} \alpha_i Z_i + c_2 e + \gamma, \tag{1}$$

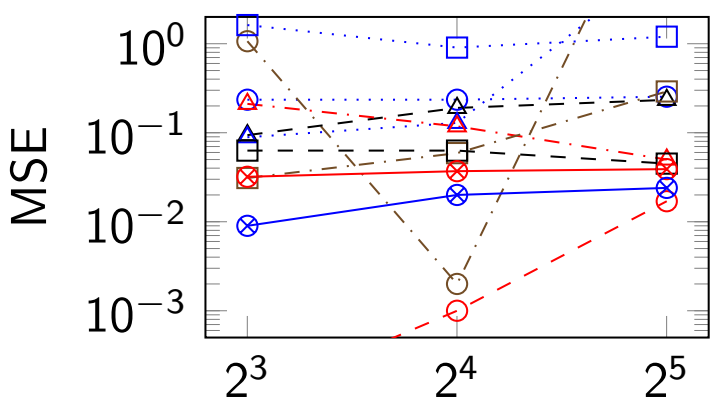
- $Z \in \mathbb{R}^{d'}$ with each entry $Z_i \sim B(2, p_i)$,
- $p_i \sim \text{unif}(0.1, 0.9)$
- $e \sim \mathcal{N}(0, 1)$
- $\alpha_i \sim \text{unif}([0.8/d', 1.2/d'])$
- $\gamma, \delta \sim \mathcal{N}(0, 0.1^2)$.
- We set $\alpha_i \sim \text{unif}([0.8/d', 1.2/d'])$.

We consider three experiments:

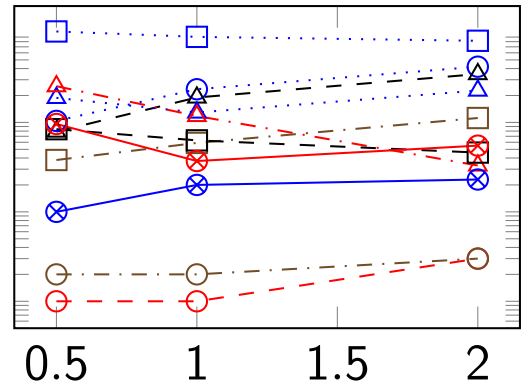
$$\text{(i) } d' = 8, 16, 32, \text{ (ii) } c_1 = 0.5, 1, 2, \text{ (iii) } c_2 = 0.5, 1, 2$$

with default parameter values: $\beta = 1$, $d' = 16$, $c_1 = 1$, $c_2 = 1$.

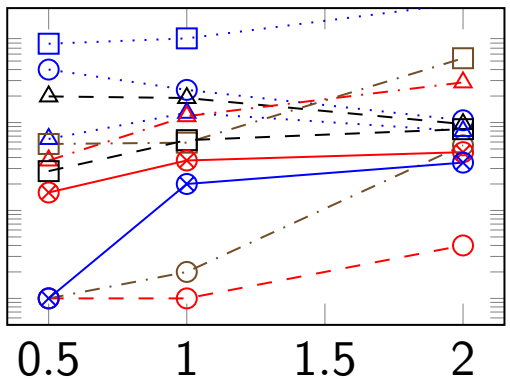
Mendelian Randomization



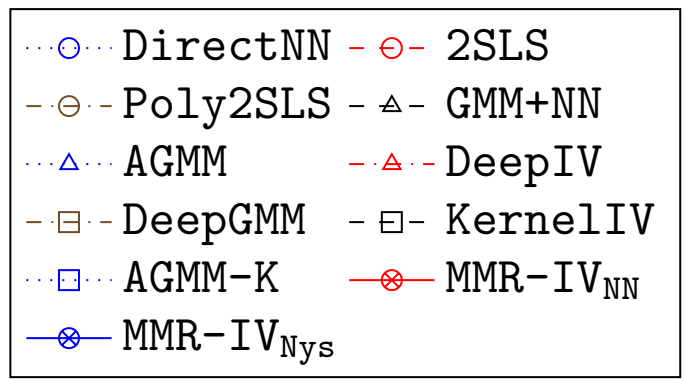
Number of IVs



Value of c_1



Value of c_2



Vitamin D Data

We apply our algorithm to the Vitamin D data (Sjolander and Martinussen 2019; Sec. 5.1).

- The data were collected from a 10-year study on 2571 individuals aged 40–71 and 4 variables are employed:
  1. *age* (at baseline)
  2. *filaggrin* (binary indicator of filaggrin mutations)
  3. *VitD* (Vitamin D level at baseline)
  4. *death* (binary indicator of death during study)

- The goal is to evaluate the potential effect of VitD on death.

Vitamin D Data

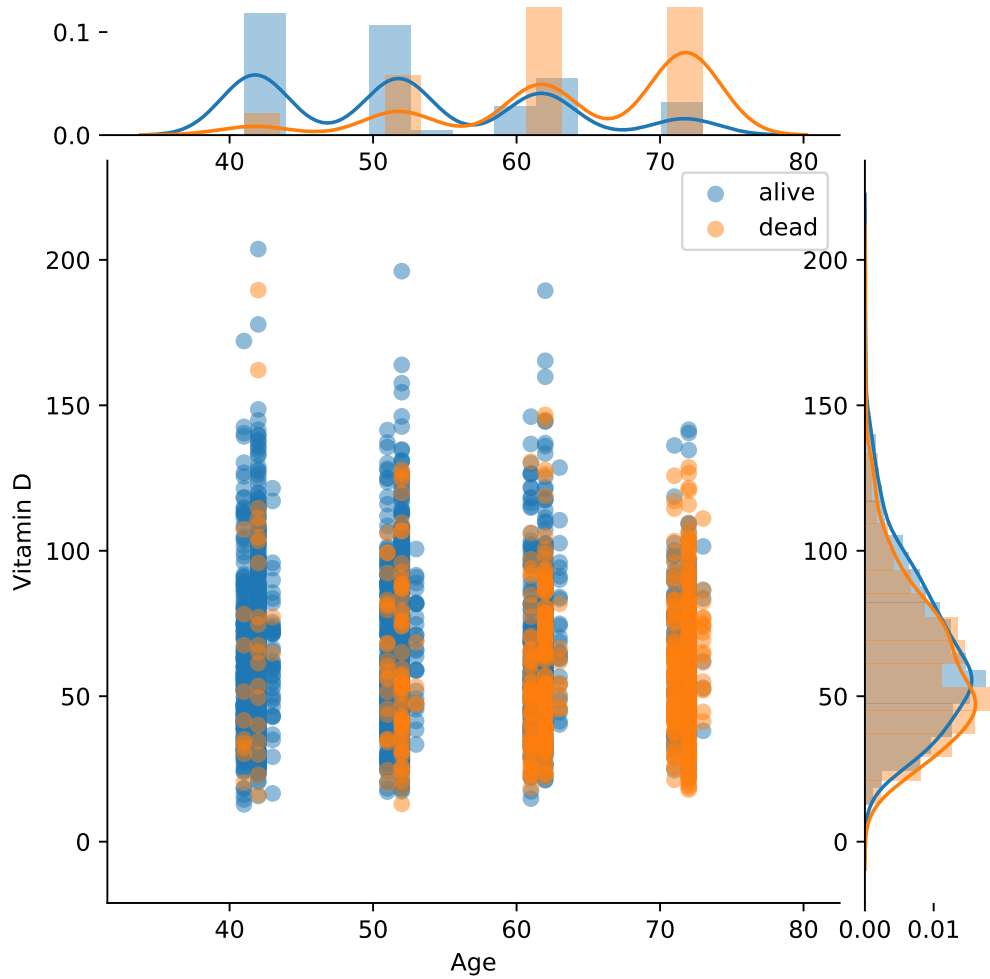
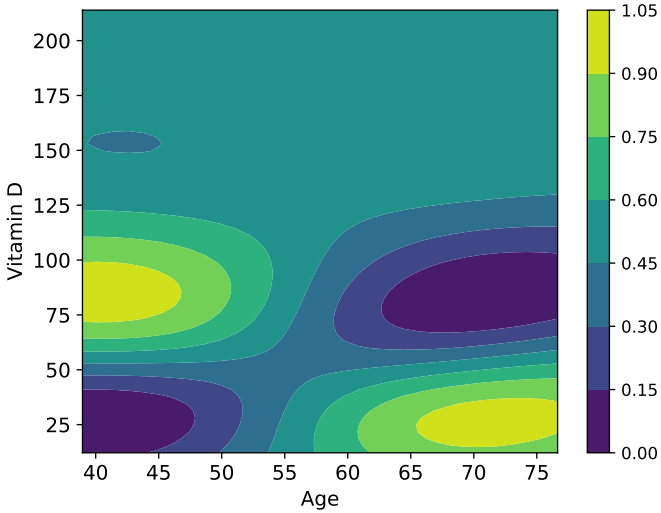
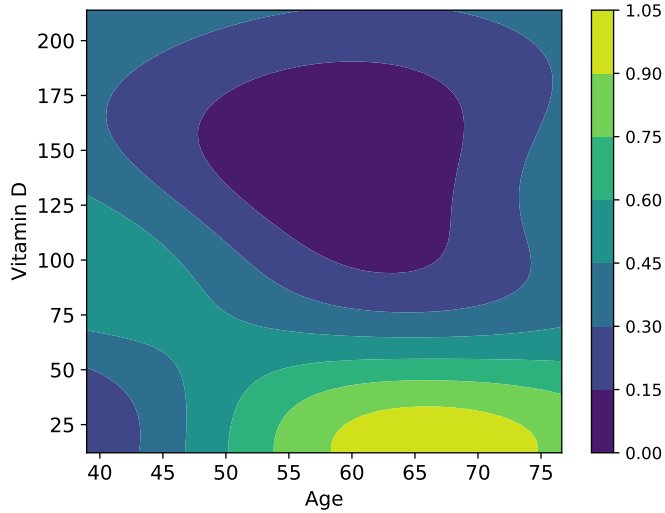


Figure: Distribution of Vitamin D data. Data points are plotted in the middle, the solid curve and histogram on the right describe the kernel density estimation and histogram of Vitamin D, and those on the top are for Age.

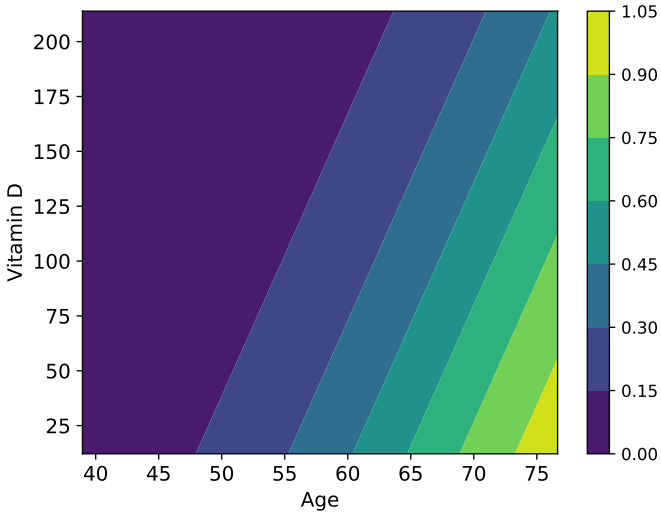
Vitamin D Data



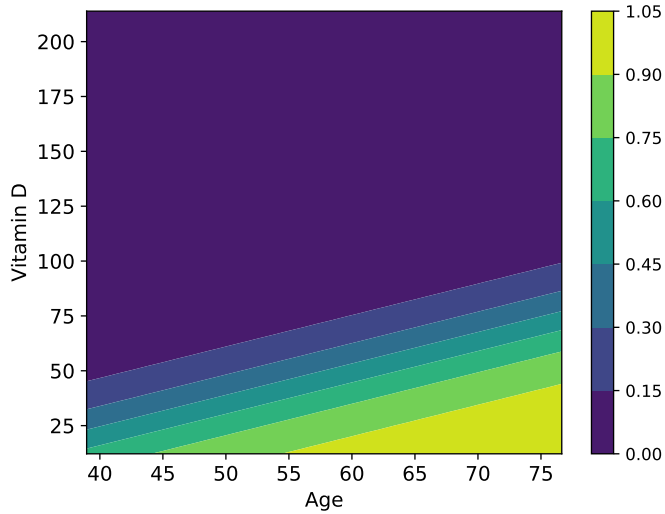
(a) Kernel Ridge Regression (IV: None)



(b) MMR-IV (IV: Filaggrin Mutation)



(c) GLM (IV: None)



(d) 2SLS + GLM (IV: Filaggrin Mutation)

References

- **Kernel Instrumental Variable Regression**
  Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019). pp. 4593–4605.
- **Kernel Conditional Moment Test via Maximum Moment Restriction**
  Uncertainty in Artificial Intelligence (UAI), PMLR 124:41-50, 2020.
  `http://proceedings.mlr.press/v124/muandet20a.html`

- **Maximum Moment Restriction for Instrumental Variable Regression**
  `https://arxiv.org/abs/2010.07684`

- **Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction**
  Proceedings of the 38th International Conference on Machine Learning, PMLR 139:7512-7523, 2021.
  `https://proceedings.mlr.press/v139/mastouri21a.html`

Q & A

References I

L. Bottou, J. Peters, J. Q. nonero Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

F. D. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *Proceeding of the 33rd International Conference on Machine Learning (ICML)*, pages xxx–xxx, 2016.

A. Sjolander and T. Martinussen. Instrumental variable estimation with the R package ivtools. *Epidemiologic Methods*, 8(1), 2019.

A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 814–823. JMLR.org, 2015.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, Nov. 1984.