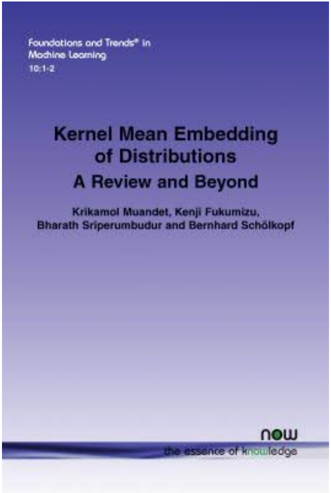
An Introduction to Hilbert Space Embedding of Probability Measures

Krikamol Muandet
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Jeju, South Korea, February 22, 2019

Reference



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

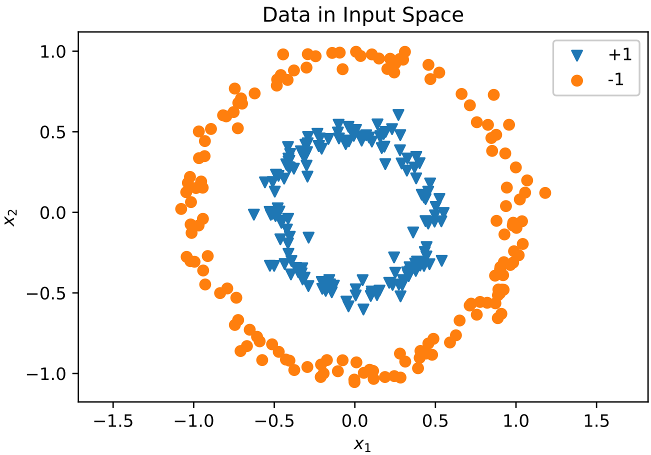
From Points to Measures

Embedding of Marginal Distributions

Embedding of Conditional Distributions

Future Directions

Classification Problem

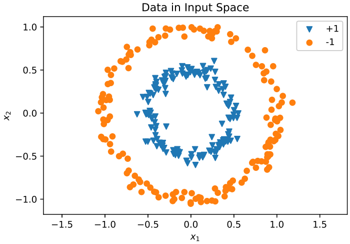


Feature Map

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

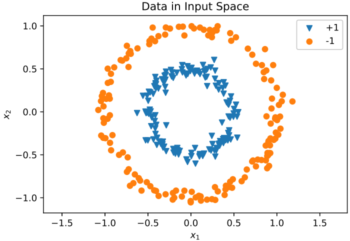
Feature Map

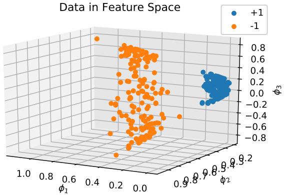
$$\phi : (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$



Feature Map

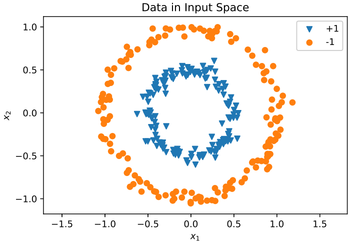
$$\phi \ : \ (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2} x_1 x_2)$$

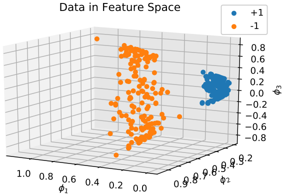




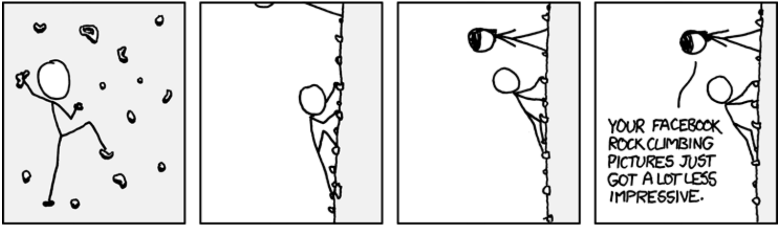
Feature Map

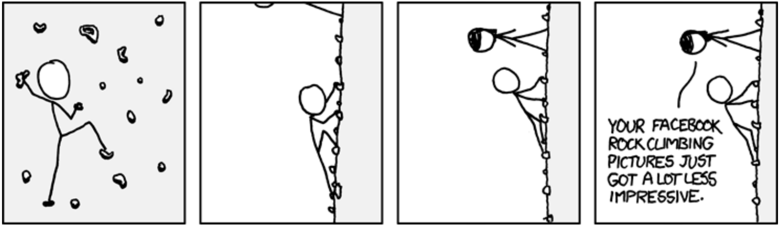
$$\phi \;:\; (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$





$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathbb{R}^3} = (\mathbf{x} \cdot \mathbf{x}')^2$$


YOUR FACEBOOK ROCK CLIMBING PICTURES JUST GOT A LOT LESS IMPRESSIVE.



Our recipe:

1. Construct a non-linear feature map $\phi : \mathcal{X} \to \mathcal{H}$.
2. Evaluate $D_\phi = \{\phi(x_1), \phi(x_2), \dots, \phi(x_n)\}$.
3. Solve the learning problem in $\mathcal{H}$ using D_ϕ.

Kernels

Definition

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** of k.

Kernels

Definition

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** of 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. $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + c)^m$ for $c > 0, \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$
   - $\dim(\mathcal{H}) = \binom{d+m}{m}$
3. $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|_2^2\right)$
   - $\mathcal{H} = \mathbb{R}^\infty$

Positive Definite Kernels

Definition (Positive definiteness)

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

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

Equivalently, we have that a **Gram** matrix **K** is positive definite.

Positive Definite Kernels

Definition (Positive definiteness)

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

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

Equivalently, we have that a **Gram** matrix $\mathbf{K}$ is positive definite.

Example (Any kernel is positive definite)

Let k be a kernel with feature map $\phi : \mathcal{X} \to \mathcal{H}$, then we have

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

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

Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space of functions mapping from $\mathcal{X}$ into $\mathbb{R}$.

Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space of functions mapping from $\mathcal{X}$ into $\mathbb{R}$.

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

$$f(x) = \langle f, k(\cdot, x) \rangle$$

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

Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space of functions mapping from $\mathcal{X}$ into $\mathbb{R}$.

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

$$f(x) = \langle f, k(\cdot, x) \rangle$$

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

2. 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.

Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space of functions mapping from $\mathcal{X}$ into $\mathbb{R}$.

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

$$f(x) = \langle f, k(\cdot, x) \rangle$$

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

2. 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.

Remark: If $\|f_n - f\|_{\mathcal{H}} \to 0$ for $n \to \infty$, then for all $x \in \mathcal{X}$, we have

$$\lim_{n \to \infty} f_n(x) = f(x)$$

.

Reproducing Kernels

Lemma (Reproducing kernels are kernels)

Let $\mathcal{H}$ be a Hilbert space over $\mathcal{X}$ 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

$$\phi(x) = k(\cdot, x), \qquad x \in \mathcal{X}.$$

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

Reproducing Kernels

Lemma (Reproducing kernels are kernels)

Let $\mathcal{H}$ be a Hilbert space over $\mathcal{X}$ 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

$$\phi(x) = k(\cdot, x), \qquad x \in \mathcal{X}.$$

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 yields

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

Kernels and RKHSs

Theorem (Every RKHS has a unique reproducing kernel)

Let $\mathcal{H}$ be an RKHS over $\mathcal{X}$. Then $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$k(\mathbf{x}, \mathbf{x}') = \langle \delta_{\mathbf{x}}, \delta_{\mathbf{x}'} \rangle_{\mathcal{H}}, \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}$$

is the only reproducing kernel of $\mathcal{H}$. Furthermore, if $(e_i)_{i \in I}$ is an orthonormal basis of $\mathcal{H}$, then for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ we have

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i \in I} e_i(\mathbf{x}) \overline{e_i(\mathbf{x}')}.$$

Kernels and RKHSs

Theorem (Every RKHS has a unique reproducing kernel)

Let $\mathcal{H}$ be an RKHS over $\mathcal{X}$. Then $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$k(\mathbf{x}, \mathbf{x}') = \langle \delta_{\mathbf{x}}, \delta_{\mathbf{x}'} \rangle_{\mathcal{H}}, \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}$$

is the only reproducing kernel of $\mathcal{H}$. Furthermore, if $(e_i)_{i \in I}$ is an orthonormal basis of $\mathcal{H}$, then for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ we have

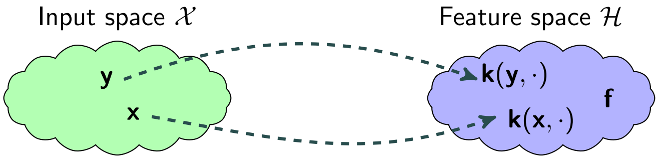
$$k(\mathbf{x}, \mathbf{x}') = \sum_{i \in I} e_i(\mathbf{x}) \overline{e_i(\mathbf{x}')}.$$

Universal kernels

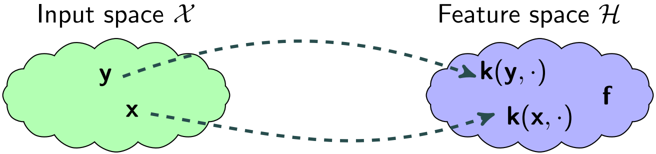
A continuous kernel k on a compact metric space $\mathcal{X}$ is called **universal** if the RKHS $\mathcal{H}$ of k is dense in $C(\mathcal{X})$, i.e., for every function $g \in C(\mathcal{X})$ and all $\varepsilon > 0$ there exist an $f \in \mathcal{H}$ such that

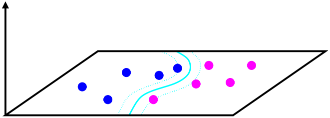
$$\|f - g\|_\infty \leq \varepsilon.$$

From Points to Measures

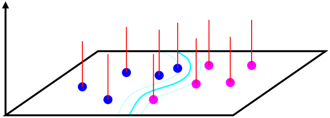


From Points to Measures



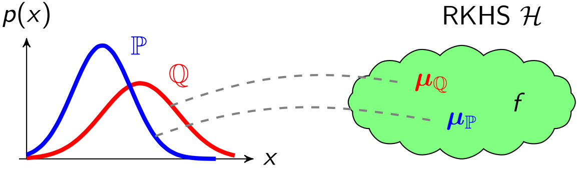


$x \mapsto k(\cdot, x)$

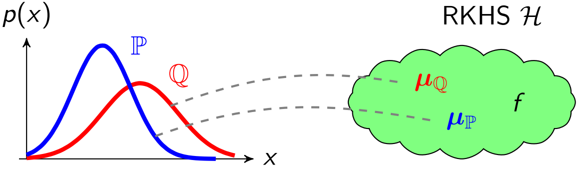


$\delta_x \mapsto \int k(\cdot, z) d\delta_x(z)$

Embedding of Marginal Distributions



Embedding of Marginal Distributions

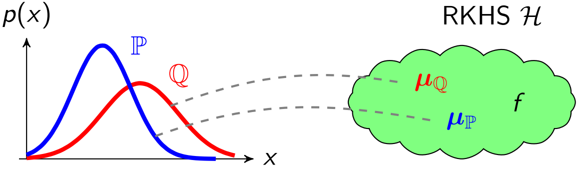


Definition

Let $\mathscr{P}$ be a space of all probability measures on a measurable space $(\mathcal{X}, \Sigma)$ and $\mathcal{H}$ an RKHS endowed with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. 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



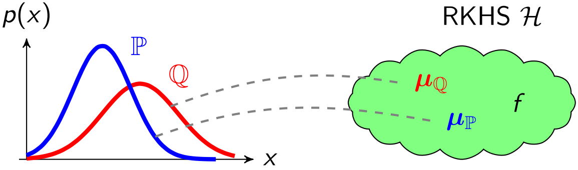
Definition

Let $\mathscr{P}$ be a space of all probability measures on a measurable space $(\mathcal{X}, \Sigma)$ and $\mathcal{H}$ an RKHS endowed with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. 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: For a Dirac measure $\delta_{\mathbf{x}}$, $\delta_{\mathbf{x}} \mapsto \mu[\delta_{\mathbf{x}}] \equiv \mathbf{x} \mapsto k(\cdot, \mathbf{x})$.

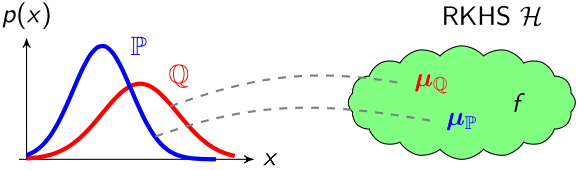
Embedding of Marginal Distributions



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

$$\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle, \quad f \in \mathcal{H}.$$

Embedding of Marginal Distributions



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

$$\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle, \quad f \in \mathcal{H}.$$

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

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

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

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

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

[1]Tolstikhin et al. *Minimax Estimation of Kernel Mean Embeddings*. JMLR, 2017.
[2]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

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

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

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

Kernel Mean Estimation

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

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

- For each $f \in \mathcal{H}$, we have $\mathbb{E}_{X\sim\widehat{\mathbb{P}}}[f(X)] = \langle f, \hat{\boldsymbol{\mu}}_{\mathbb{P}}\rangle$.
- Assume that $\|f\|_\infty \leq 1$ for all $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$. W.p.a.l $1-\delta$,

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

[1] Tolstikhin et al. *Minimax Estimation of Kernel Mean Embeddings*. JMLR, 2017.
[2] 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

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

- For each $f \in \mathcal{H}$, we have $\mathbb{E}_{X\sim\widehat{\mathbb{P}}}[f(X)] = \langle f, \hat{\boldsymbol{\mu}}_{\mathbb{P}} \rangle$.
- Assume that $\|f\|_\infty \leq 1$ for all $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$. W.p.a.l $1-\delta$,

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

- The convergence happens at a rate $O_p(n^{-1/2})$ which has been shown to be minimax optimal.[1]

[1] Tolstikhin et al. *Minimax Estimation of Kernel Mean Embeddings*. JMLR, 2017.
[2] 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

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

- For each $f \in \mathcal{H}$, we have $\mathbb{E}_{X\sim\widehat{\mathbb{P}}}[f(X)] = \langle f, \hat{\boldsymbol{\mu}}_{\mathbb{P}}\rangle$.
- Assume that $\|f\|_{\infty} \leq 1$ for all $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$. W.p.a.l $1-\delta$,

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

- The convergence happens at a rate $O_p(n^{-1/2})$ which has been shown to be minimax optimal.[1]
- In high dimensional setting, we can improve an estimation by shrinkage estimators:[2]

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

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

Explicit 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* $\mathbb{P}$

Explicit 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, $\boldsymbol{\mu}_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}[e^{\langle X, \cdot \rangle}]$.

Explicit 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

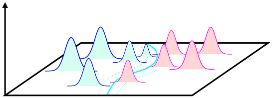
Consider $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$ where ψ is a positive definite function. Then,

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

for positive finite measure Λ.

Application: High-Level Generalization

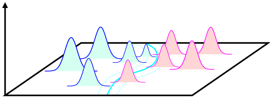
Learning from Distributions



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

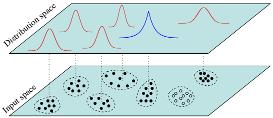
Application: High-Level Generalization

Learning from Distributions



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

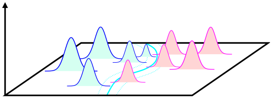
Group Anomaly Detection



KM. and Schölkopf, UAI 2013.

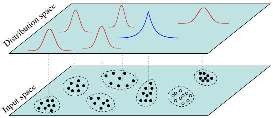
Application: High-Level Generalization

Learning from Distributions



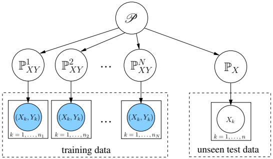
KM., Fukumizu, Dinuzzo, Schölkopf. NIPS 2012.

Group Anomaly Detection



KM. and Schölkopf, UAI 2013.

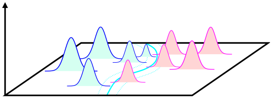
Domain Adaptation/Generalization



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

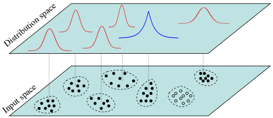
Application: High-Level Generalization

Learning from Distributions



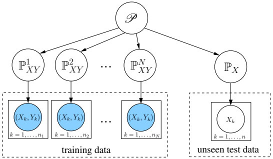
KM., Fukumizu, Dinuzzo, Schölkopf. NIPS 2012.

Group Anomaly Detection



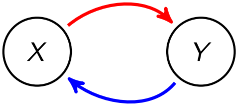
KM. and Schölkopf, UAI 2013.

Domain Adaptation/Generalization



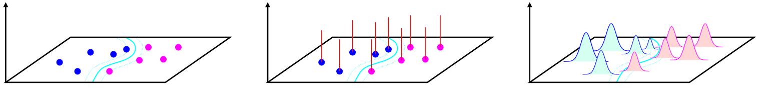
KM. et al. ICML 2013; Zhang, **KM.** et al. ICML 2013

Cause-Effect Inference



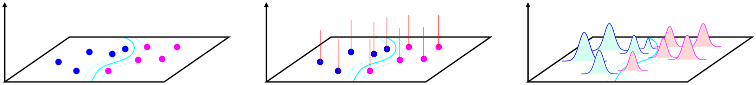
Lopez-Paz, **KM.** et al. JMLR 2015, ICML 2015.

Support Measure Machine (SMM)



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

Support Measure Machine (SMM)



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

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

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

Kernel Discrepancy Measure for Distributions

- Maximum mean discrepancy (MMD)

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

Kernel Discrepancy Measure for Distributions

- Maximum mean discrepancy (MMD)

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

- MMD is an **integral probability metric (IPM)** and corresponds to the RKHS distance between mean embeddings.

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

Kernel Discrepancy Measure for Distributions

- Maximum mean discrepancy (MMD)

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

- MMD is an **integral probability metric (IPM)** and corresponds to the RKHS distance between mean embeddings.

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

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

Kernel Discrepancy Measure for Distributions

- Maximum mean discrepancy (MMD)

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

- MMD is an **integral probability metric (IPM)** and corresponds to the RKHS distance between mean embeddings.

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

- If k is **universal**, 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

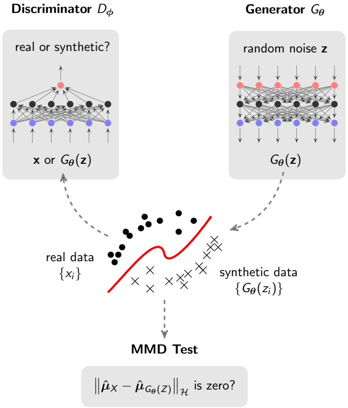
$$\widehat{\text{MMD}}^2_u(\mathbb{P},\mathbb{Q},\mathcal{H}) = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n k(\mathbf{x}_i,\mathbf{x}_j) + \frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j\neq i}^m k(\mathbf{y}_i,\mathbf{y}_j) - \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m k(\mathbf{x}_i,\mathbf{y}_j).$$

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

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

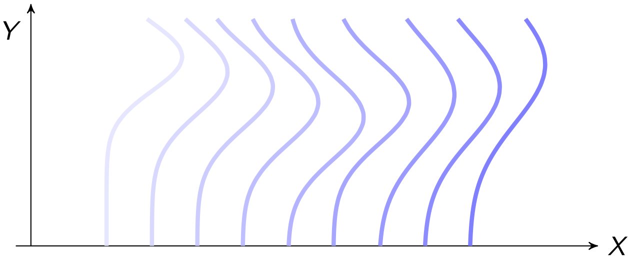
Generative Moment Matching Network

- 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

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

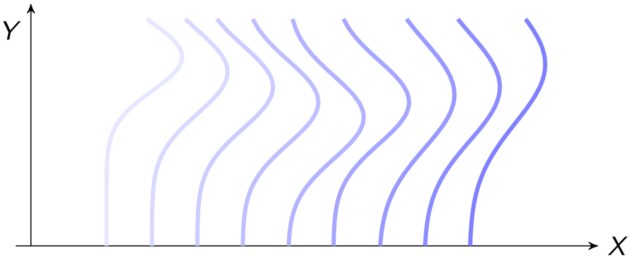
- Many tricks have been proposed to improve the GMMN:
  - Optimized kernels and feature extractors (Sutherland et al., 2017; Li et al., 2017a),
  - Gradient regularization (Binkowski et al., 2018; Arbel et al., 2018)
  - Repulsive loss (Wang et al., 2019)
  - Optimized witness points (Mehrjou et al.,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

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.

Covariance Operators

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

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

Covariance Operators

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

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

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

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

Covariance Operators

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

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

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

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

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

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

Covariance Operators

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

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

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

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

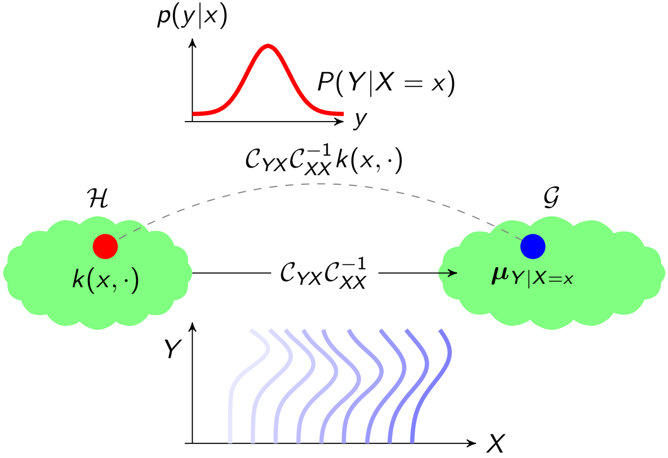
- Alternatively, $\mathcal{C}_{YX}$ is a unique bounded operator satisfying

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

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

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

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

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

  for all $g \in \mathcal{G}$.

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

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

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

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

Conditional Mean Estimation

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

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

Conditional Mean Estimation

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

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

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

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

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

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

Conditional Mean Estimation

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

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

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

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

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

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

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

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

- By the law of total expectation,

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

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

- By the law of total expectation,

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

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

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

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

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

[3] 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 $\mu_Y^{\otimes} = \mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)]$.

[3] 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

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

[3] 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

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

[3] 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**.[3]

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

From Points to Measures

Embedding of Marginal Distributions

Embedding of Conditional Distributions

Future Directions

Future Directions

- Representation learning and embedding of distributions
- Kernel methods in deep learning
  - MMD-GAN
  - Wasserstein autoencoder (WAE)
  - Invariant learning in deep neural networks
- Kernel mean estimation in high dimensional setting
- Recovering (conditional) distributions from mean embeddings

Q & A