Kernel Mean Embedding (Part II)

Maximum Mean Discrepancy (MMD)

Recap: RKHS

Kernel

- A kernel is a "similarity" measure:
  Given $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and a feature map $\phi : \mathcal{X} \to \mathcal{F}$ for some Hilbert space $\mathcal{F}$,

$$k : (\boldsymbol{x}, \boldsymbol{y}) \mapsto \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle_{\mathcal{F}}.$$

- We don't care about ϕ:
  If k is positive definite (pd), then such ϕ must exists (aka canonical feature map).

Reproducing Kernel Hilbert Space (RKHS)

- RKHS $\mathcal{H}$ of space $\mathcal{X}$ is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ such that:
  every $\boldsymbol{x} \in \mathcal{X}$ uniquely corresponds to $k_{\boldsymbol{x}} \in \mathcal{H}$ and

$$f(\boldsymbol{x}) = \langle k_{\boldsymbol{x}}, f \rangle_{\mathcal{H}}, \ \forall f \in \mathcal{H}. \qquad \text{(Reproducing property)}$$

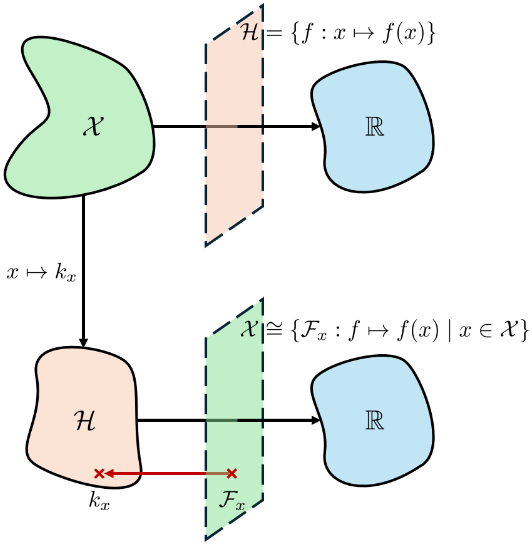
- $\mathcal{H}$ uniquely corresponds to a pd kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) = \langle k_{\boldsymbol{x}}, k_{\boldsymbol{y}} \rangle_{\mathcal{H}}. \qquad \text{(Reproducing kernel)}$$

- Every pd kernel uniquely corresponds to an RKHS.

Recap: RKHS

To illustrate how an RKHS is constructed:



Recap: Kernel Mean Embedding

- Let $\mathcal{X}$ be a probability space and $\mathcal{H}$ be an RKHS of $\mathcal{X}$ with reproducing kernel k. The kernel mean embedding of $\mathbb{P}$, $\mu_{\mathbb{P}} \in \mathcal{H}$, is defined as

$$\mu_{\mathbb{P}} : \boldsymbol{y} \mapsto \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[k(X, \boldsymbol{y})],$$
$$\text{recall } k_{\boldsymbol{x}} : \boldsymbol{y} \mapsto k(\boldsymbol{x}, \boldsymbol{y}).$$

- Reproducing property:
  Every distribution $\mathbb{P}$ that satisfies $\mathbb{E}_{X \sim \mathbb{P}}[\sqrt{k(X, X)}] < \infty$ has a unique kernel mean embedding $\mu_{\mathbb{P}}$ such that

$$\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle \mu_{\mathbb{P}}, f \rangle_{\mathcal{H}}, \ \forall f \in \mathcal{H}$$
$$\text{recall } f(X) = \langle k_{\boldsymbol{x}}, f \rangle_{\mathcal{H}}.$$

- Generalized kernel trick:

$$\mathbb{E}_{X, Y \sim \mathbb{P}, \mathbb{Q}}[k(X, Y)] = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}},$$
$$\text{recall } k(X, Y) = \langle k_{\boldsymbol{x}}, k_{\boldsymbol{y}} \rangle_{\mathcal{H}}.$$

Maximum Mean Discrepancy

From now on, let $\mathcal{X}$ be a probability space and $\mathcal{H}$ be an RKHS of $\mathcal{X}$ with reproducing kernel k. Also assume all distributions have mean embedding.

- The maximum mean discrepancy (MMD) between two distributions $\mathbb{P}, \mathbb{Q}$ on $\mathcal{X}$ is defined as

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

- Q: How to estimate MMD in practice?

$$\begin{aligned}\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}}\|_{\mathcal{H}}^2 + \|\mu_{\mathbb{Q}}\|_{\mathcal{H}}^2 - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}} \\ &= \mathbb{E}_{X,X' \sim \mathbb{P},\mathbb{P}}[k(X, X')] + \mathbb{E}_{Y,Y' \sim \mathbb{Q},\mathbb{Q}}[k(Y, Y')] \\ &\quad - 2\mathbb{E}_{X,Y \sim \mathbb{P},\mathbb{Q}}[k(X, Y)].\end{aligned}$$

- Suppose we sample $X_1, \dots, X_n \sim \mathbb{P}, Y_1, \dots, Y_m \sim \mathbb{Q}$ i.i.d., then we have an unbiased estimator

$$\widehat{\mathrm{MMD}^2}(\mathbb{P}, \mathbb{Q}) := \frac{1}{n(m-1)} \sum_{i \neq j} k(X_i, X_j) + k(Y_i, Y_j) - 2k(X_i, Y_j).$$

Maximum Mean Discrepancy

MMD as integral probability metric

- Given a collection $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$, the integral probability metric is defined as

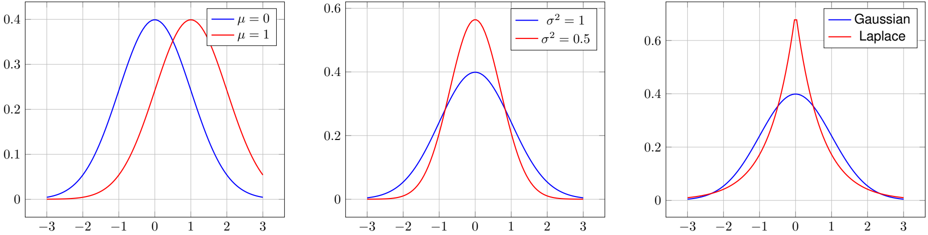
$$D_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)].$$

- When $\mathcal{F}$ is the unit ball in $\mathcal{H}$, namely $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$, $D_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})^2 = \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})$.
  The key is applying the reproducing property:

$$\begin{aligned} &\sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} \underbrace{\mathbb{E}_{X \sim \mathbb{P}}[f(X)]}_{\langle \mu_{\mathbb{P}}, f \rangle_{\mathcal{H}}} - \underbrace{\mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]}_{\langle \mu_{\mathbb{Q}}, \rangle_{\mathcal{H}}} \\ = &\sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, f \rangle_{\mathcal{H}} = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}. \end{aligned}$$

Application: Statistical Testing

- Consider a binary hypothesis test with null hypothesis $h_0 : \mathbb{P} = \mathbb{Q}$ and alternative hypothesis $h_1 : \mathbb{P} \neq \mathbb{Q}$.
- Suppose we sample $X_1, \ldots, X_n \sim \mathbb{P}, Y_1, \ldots, Y_n \sim \mathbb{Q}$ i.i.d. Our goal is to design a decision criterion whether we should reject h_0 or not.



- Instead of comparing empirical mean, we can use the empirical estimator of MMD.
- From left to right: two distributions are harder to distinguish, so we should use more complicated kernels.

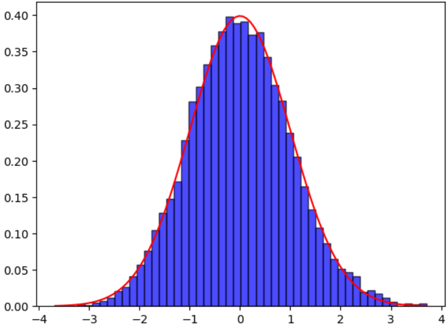
Application: Statistical Testing

Key facts [Gretton et. al. (2006), Theorem 8]

- When $\mathbb{P} \neq \mathbb{Q}$,

$$\sqrt{n} \cdot \frac{\widehat{\mathrm{MMD}^2}(\mathbb{P}, \mathbb{Q}) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})}{\sigma(\mathbb{P}, \mathbb{Q})} \to \mathcal{N}(0, 1).$$

- Example: let $\mathbb{P} = \mathcal{N}(0, 1), \mathbb{Q} = \mathcal{N}(1, 1)$, then $\sqrt{n} \cdot \frac{\widehat{\mathrm{MMD}^2}(\mathbb{P},\mathbb{Q}) - 1}{\sqrt{12}} \to \mathcal{N}(0, 1)$.



Application: Statistical Testing

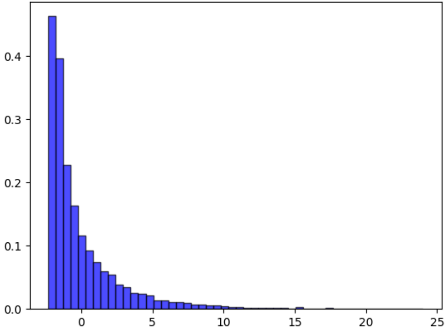
Key facts [Gretton et. al. (2006), Theorem 8]

- When $\mathbb{P} = \mathbb{Q}$,

$$n \cdot \widehat{\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})} \to \sum_{i=1}^{\infty} 2\lambda_i (Z_i^2 - 1),$$

  where $Z_i \sim \mathcal{N}(0,1)$ i.i.d. and λ_i's are eigenvalues of

$$f \mapsto \mathbb{E}_{X,X' \sim \mathbb{P}}[\tilde{k}(X, X') f(X)], \quad \tilde{k}(\boldsymbol{x}, \boldsymbol{x}') = \langle k_{\boldsymbol{x}} - \mu_{\mathbb{P}}, k_{\boldsymbol{x}'} - \mu_{\mathbb{P}} \rangle_{\mathcal{H}}.$$

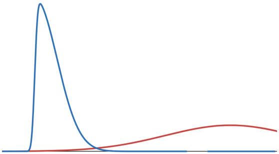


Application: Statistical Testing

Decision rule

- Let $T_0 := n \cdot \widehat{\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})}$, then

$$T_0 \approx \begin{cases} n \cdot \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) + \sqrt{n} \cdot \mathcal{N}(0, \sigma^2), & \mathbb{P} \neq \mathbb{Q} \\ \sum_{i=1}^{\infty} 2\lambda_i (Z_i^2 - 1), & \mathbb{P} = \mathbb{Q}. \end{cases}$$



- If T_0 is large, then we should believe h_0 is unlikely.

Decision rule: reject h_0 if $T_0 \geq c_\alpha$.

- The goal is to determine c_α so that the error $\mathbb{P}\{T_0 \geq c_\alpha | h_0\} \leq \alpha$, where α is some fixed confidence level.
  In other words, c_α is the $1 - \alpha$ quantile of $T_0 | h_0 \Rightarrow$ this is impractical to compute.

Application: Statistical Testing

A more feasible decision rule

- Suppose we can sample T from the same distribution as T_0. Given $T_0 = t_0$,

$$\mathbb{P}_T\{T \geq t_0\} \leq \alpha \implies t_0 \geq c_\alpha.$$

  In other words, t_0 is above the $1-\alpha$ quantile.

- Instead of directly computing c_α, estimate $\mathbb{P}_T\{T \geq T_0\}$:
  Given $T_1, \ldots, T_m$ sampled i.i.d. under h_0,

$$\mathbb{P}_T\{T \geq T_0\} \approx \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{T_i \geq T_0\}.$$

Application: Statistical Testing

Permutation test

- Given $X_1, \ldots, X_n \sim \mathbb{P}, Y_1, \ldots, Y_n \sim \mathbb{Q}$, permute all samples and partition into $\tilde{X}_1, \ldots, \tilde{X}_n, \tilde{Y}_1, \ldots, \tilde{Y}_n$. Then compute

$$T_l = \widehat{\mathrm{MMD}^2}(\tilde{\mathbb{P}}, \tilde{\mathbb{Q}}) := \frac{1}{n(n-1)} \sum_{i \neq j} k(\tilde{X}_i, \tilde{X}_j) + k(\tilde{Y}_i, \tilde{Y}_j) - 2k(\tilde{X}_i, \tilde{Y}_j).$$

  Under null hypothesis $h_0 : \mathbb{P} = \mathbb{Q}$, T_l and T_0 have the same distribution.

- New decision rule:
  Permute samples and compute T_l for $l = 1, \ldots, m$. Reject h_0 if

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{T_i \geq T_0\} \leq \alpha.$$

- Example