

Chapter 4, Part 5: Representing Probability Distributions in RKHSs

Advanced Topics in Statistical Machine Learning

Tom Rainforth

Hilary 2022

rainforth@stats.ox.ac.uk

Representing Probability Distributions with Features

Intuitively, two distributions are similar if they produce similar expectations across a wide range of functions

We can thus form a feature representation $\mu_\varphi(P)$ of a **distribution** P by introducing some set of "feature functions" $\varphi_\ell : \mathcal{X} \mapsto \mathbb{R},\ \ell = 1, \ldots, r$ and calculating each of their expectations with respect to P:

$$\begin{aligned}\mu_\varphi(P) &:= \left[\mathbb{E}_{X\sim P}\varphi_1(X), \mathbb{E}_{X\sim P}\varphi_2(X), \ldots, \mathbb{E}_{X\sim P}\varphi_r(X)\right], \\ &= \mathbb{E}_{X\sim P}\left[\varphi_1(X), \varphi_2(X), \ldots, \varphi_r(X)\right] \\ &=: \mathbb{E}_{X\sim P}\varphi(X) \quad \text{where} \quad \varphi : \mathcal{X} \to \mathbb{R}^r\end{aligned}$$

Examples:

- Histograms: $\varphi_\ell(X) = \mathbb{I}(X \in \mathcal{X}_\ell)/\|\mathcal{X}_\ell\|$ where each $\mathcal{X}_\ell$ represent a bin and $\|\mathcal{X}_\ell\|$ the bin size
- Monte Carlo representations: informally, we can take the limit of small bin sizes of the histogram estimate above with $r \to \infty$
- Various density estimators, e.g. moment matching methods

Kernel Mean Embeddings

- $\mu_\varphi : P \to \mathbb{R}^r$ is a feature map for probability distributions: it is analogous to $\varphi : \mathcal{X} \mapsto \mathbb{R}^r$ for individual datapoints
- We can extend this to RKHS feature spaces by using the canonical feature map $\varphi(x) = k(\cdot, x)$ and defining the **kernel mean embedding** $\mu_k : P \mapsto \mathcal{H}_k$ as follows

$$\mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \tag{1}$$

- $\mu_k(P)$ represents P as a function in $\mathcal{H}_k$
- For **characteristic kernels** (includes RBF, Matèrn, rational quadratic), this representation is **unique**, in the sense that no two distinct distributions will have the same kernel mean embedding (i.e. μ_k is injective): if $\mathcal{H}_k$ is sufficiently rich, $\mu_k(P)$ can store all the information of P

Kernel Density Estimators

Kernel density estimation is a simple classical non–parametric density estimation approach that forms the density estimate

$$\hat{p}(X = x) = \frac{1}{n} \sum_{i=1}^{n} k(x, x_i),$$

where x_i are the datapoints, and k is a non–negative, normalized, kernel (i.e. $k(x, x') \geq 0 \; \forall x, x'$, $\int k(x, x') dx' = 1 \; \forall x$),[1] most commonly a Gaussian (i.e. the RBF kernel with an appropriate normalization constant).

It is straightforward to see that the function $\hat{p} : \mathcal{X} \mapsto \mathbb{R}^+$ equates to a Monte Carlo estimate of a kernel mean embedding with some additional assumptions on the kernel.

[1] Note that this can sometimes lead to some inconsistencies to how the term kernel is defined in the literature: the idea of a kernel as a "window" function is not always synonymous with how we have been using it

Link to Monte Carlo Estimators

- The standard Monte Carlo empirical measure (i.e. $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x)$) can further be viewed as limit of such a kernel density estimator.
- For example, if $\mathcal{X} = \mathbb{R}^p$ then we can use a normalized RBF kernel and take the limit $\gamma \to 0$
- Using kernels allows us to make additional smoothness assumptions about P: the kernel density estimate is effectively a smoothing of our samples
- Kernel mean embeddings also have uses beyond simple density estimation: we can carry out operations in the RKHS do things like form divergence metrics or perform independence tests

Operator View

Using the reproducing property, we see that a kernel mean embedding converts a function $f \in \mathcal{H}_k$ to its mean with respect to P when we take an inner product:

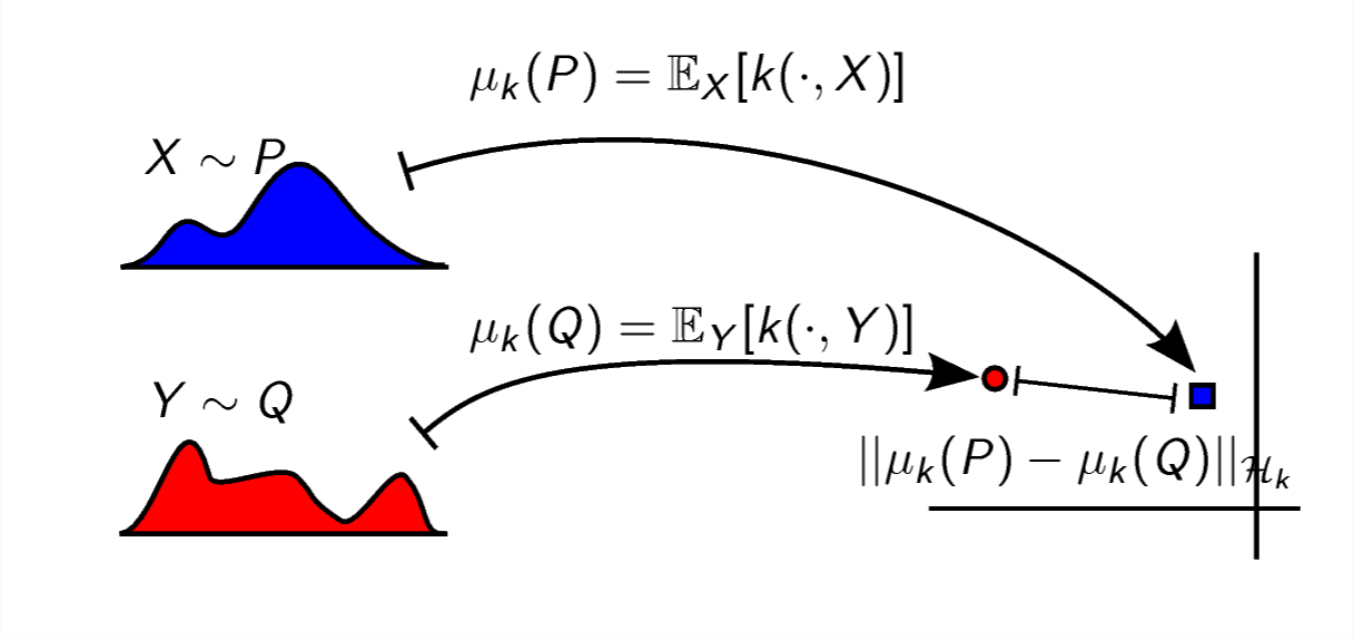
$$\begin{aligned}\langle \mu_k(P), f \rangle_{\mathcal{H}_k} &= \langle \mathbb{E}_{X \sim P} k(\cdot, X), f \rangle_{\mathcal{H}_k} \\ &= \mathbb{E}_{X \sim P} \langle k(\cdot, X), f \rangle_{\mathcal{H}_k} \\ &= \mathbb{E}_{X \sim P} f(X), \quad \forall f \in \mathcal{H}_k.\end{aligned}$$

Of particular note is the case where f is itself a kernel mean embedding: inner products between kernel mean embeddings can be computed as

$$\begin{aligned}\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} &= \langle \mathbb{E}_{X \sim P} k(\cdot, X), \mathbb{E}_{Y \sim Q} k(\cdot, Y) \rangle_{\mathcal{H}_k} \\ &= \mathbb{E}_{X \sim P, Y \sim Q} k(X, Y).\end{aligned}$$

Comparing Distributions

One of the most powerful uses of kernel mean embeddings is to measure discrepancies between distributions by considering their distance in RKHS norm.



Maximum Mean Discrepancy

Such distances are called maximum mean discrepancies (MMDs)

$$\begin{aligned}
&\mathrm{MMD}_k\left(P, Q\right) = \left\|\mu_k\left(P\right) - \mu_k\left(Q\right)\right\|_{\mathcal{H}_k} \\
&= \sqrt{\left\langle \mu_k\left(P\right), \mu_k\left(P\right)\right\rangle_{\mathcal{H}_k} + \left\langle \mu_k\left(Q\right), \mu_k\left(Q\right)\right\rangle_{\mathcal{H}_k} - 2\left\langle \mu_k\left(P\right), \mu_k\left(Q\right)\right\rangle_{\mathcal{H}_k}} \\
&= \sqrt{\mathbb{E}_{X\sim P, X'\sim P} k(X, X') + \mathbb{E}_{Y\sim Q, Y'\sim Q} k(Y, Y') - 2\mathbb{E}_{X\sim P, Y\sim Q} k(X, Y)},
\end{aligned}$$

where X and X' are independent random samples from P, and similarly Y and Y' are independent samples from Q.

We can interpret the MMD squared as (twice) the average similarity of pairs of points from the same distribution minus the average similarity of pair of points from different distributions

For characteristic kernels, the MMD is a proper metric on probability distributions: $\mathrm{MMD}_k\left(P, Q\right) = 0$ if and only if $P = Q$

Estimating the MMD

Given sets of independent[2] samples $\{x_i\}_{i=1}^{n_x} \overset{i.i.d.}{\sim} P$, $\{y_i\}_{i=1}^{n_y} \overset{i.i.d.}{\sim} Q$, a simple unbiased estimator of the squared MMD in is given by

$$\begin{aligned}\widehat{\mathrm{MMD}}_k^2(P,Q) = &\frac{1}{n_x(n_x-1)}\sum_{i\neq j} k(x_i,x_j) \\ &+ \frac{1}{n_y(n_y-1)}\sum_{i\neq j} k(y_i,y_j) - \frac{2}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} k(x_i,y_j)\,.\end{aligned} \tag{2}$$

Though its cost scales as $O((n_x+n_y)^2)$, it has the highly beneficial property of only requiring samples from each distribution: most divergences (e.g. KL) require at least one of the density functions.

This leads to applications in non–parametric hypothesis testing and training implicit models (e.g. GANs)

[2]We can still construct an unbiased estimator if certain y_i depend on certain x_i (or vice versa) by omitting the dependent pairs from the third empirical average.

Maximum Mean Discrepancy

The name MMD comes from an alternative insightful formulation that it is the largest possible discrepancy between the two expectations of a function in the RKHS with bounded norm.

More formally, as proved in the examples sheet,

$$\mathrm{MMD}_k(P, Q) = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{Y \sim Q} f(Y)|$$

The worst case f (i.e. the f that forms the supremum) turns out to be proportional to the **witness function** $\mu_k(P) - \mu_k(Q)$

Visualization

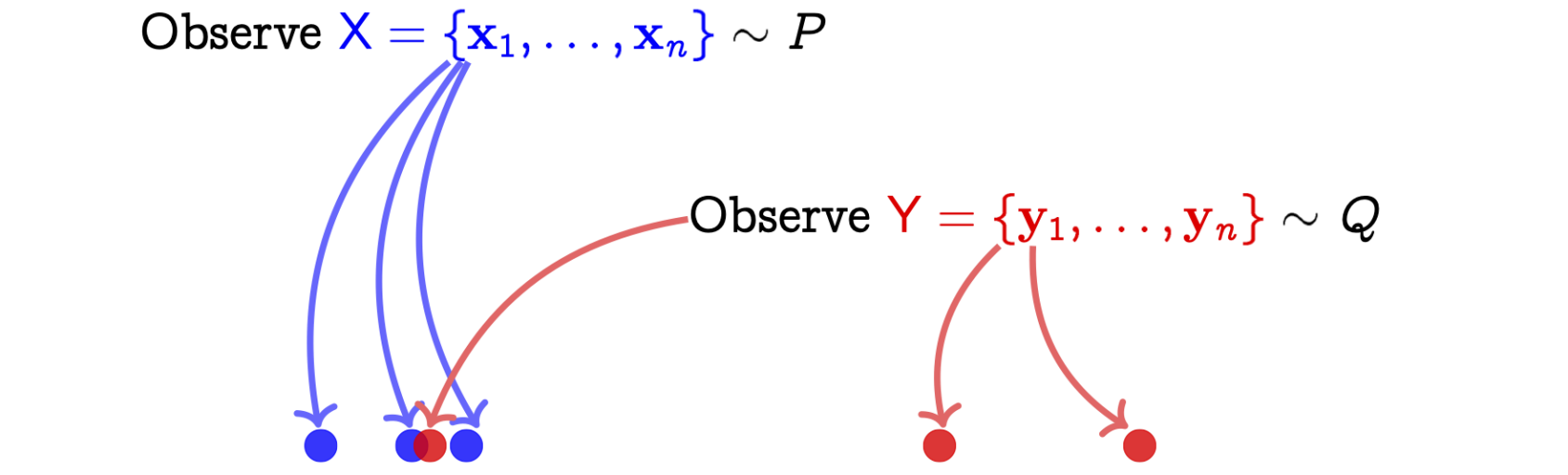


Figure credit: Arthur Gretton, `http://www.gatsby.ucl.ac.uk/~gretton/papers/testing_workshop.pdf`

Visualization

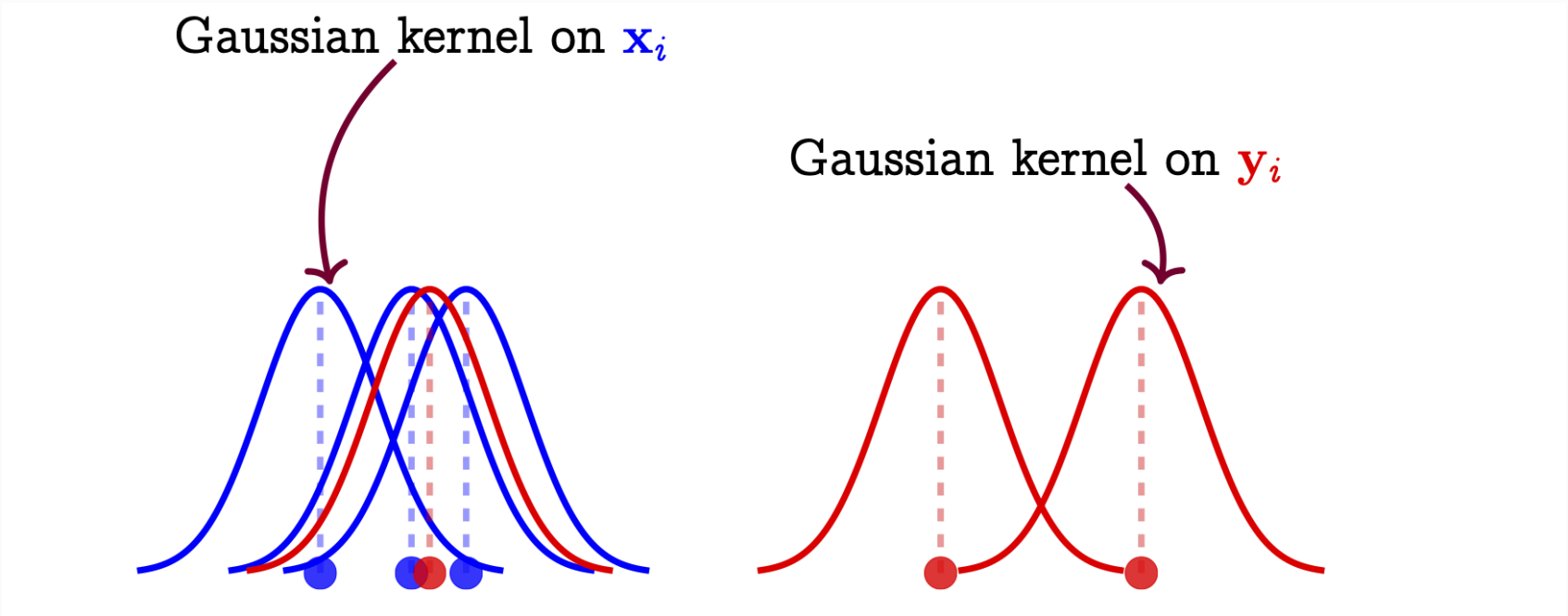


Figure credit: Arthur Gretton, http://www.gatsby.ucl.ac.uk/~gretton/papers/testing_workshop.pdf

Visualization

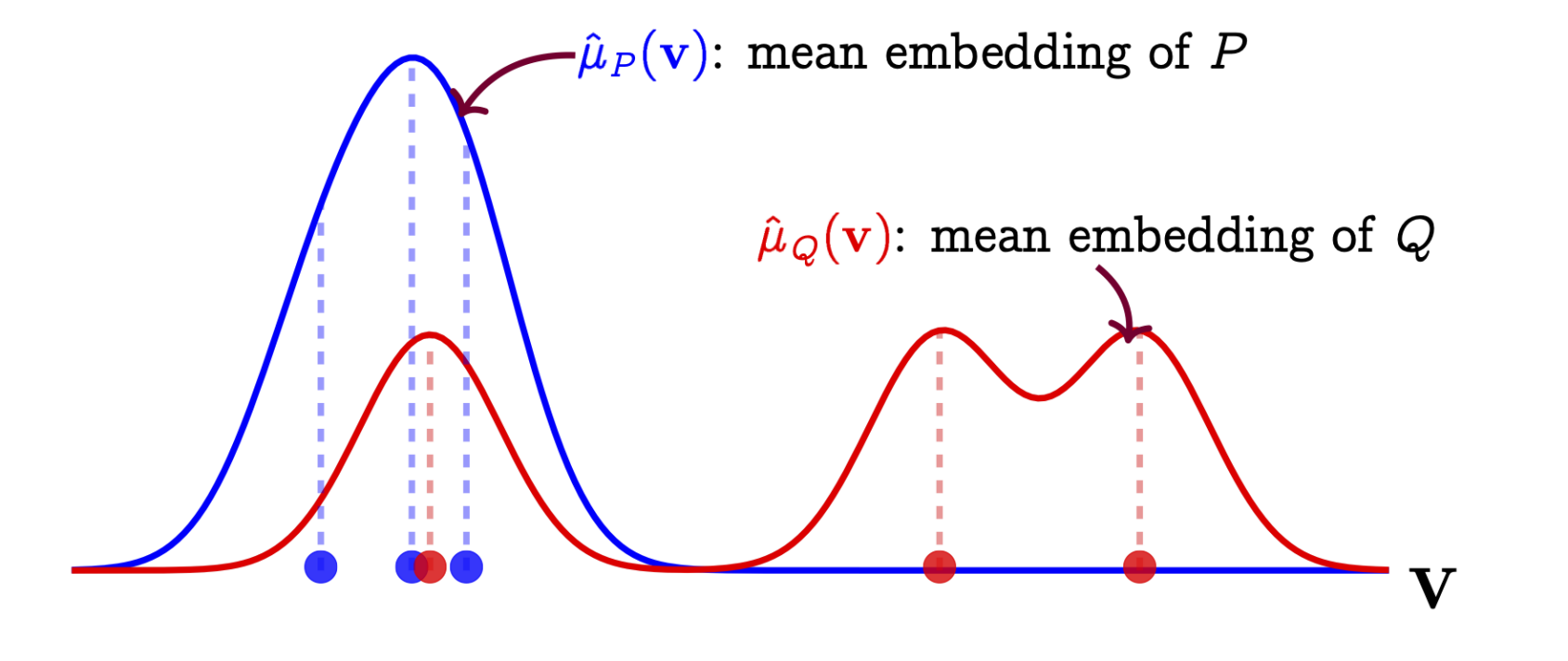


Figure credit: Arthur Gretton, http://www.gatsby.ucl.ac.uk/~gretton/papers/testing_workshop.pdf

Visualization

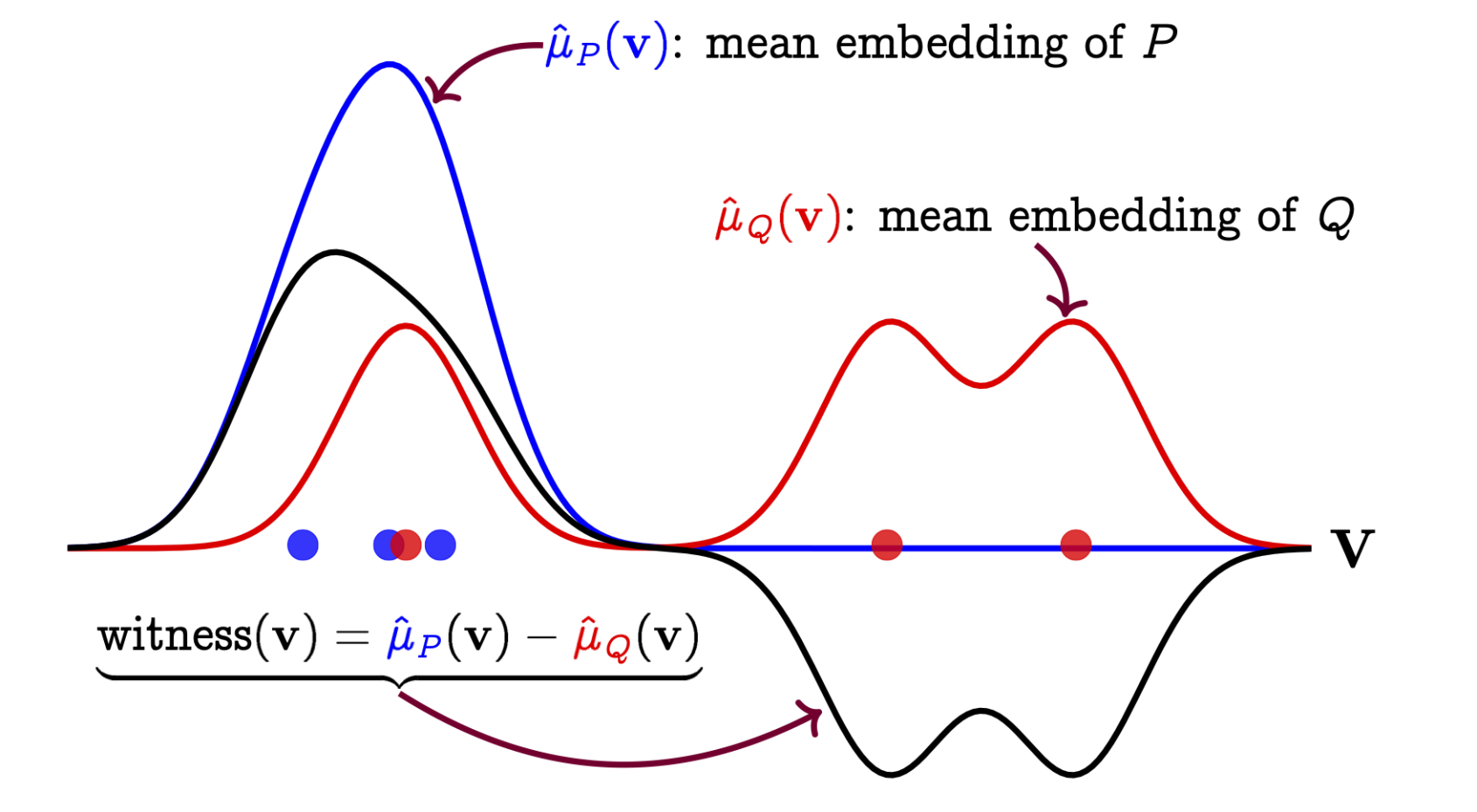
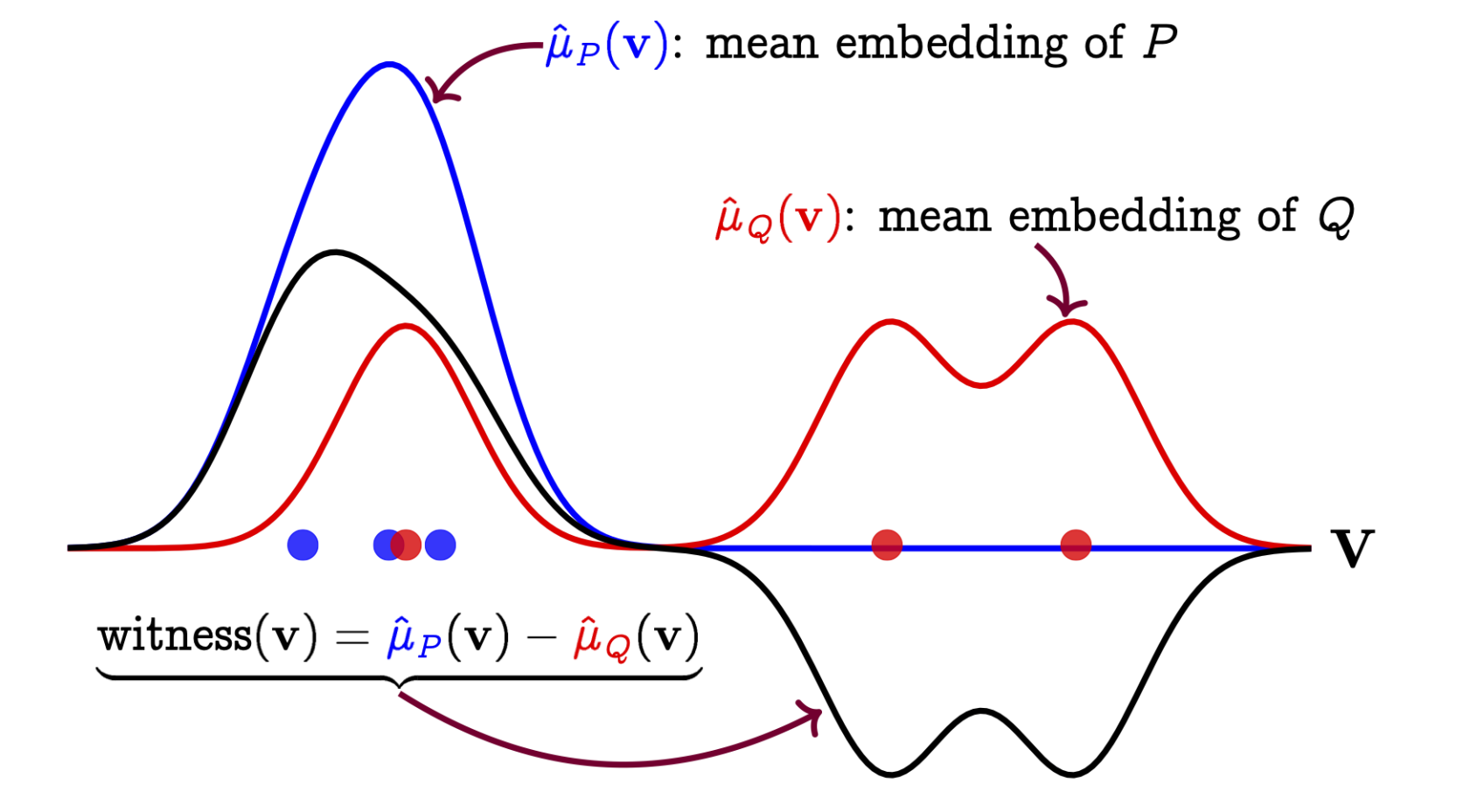


Figure credit: Arthur Gretton, http://www.gatsby.ucl.ac.uk/~gretton/papers/testing_workshop.pdf

Visualization



$$\text{MMD}_k\left(P, Q\right) = \|\text{witness}\|_{\mathcal{H}_k} = \|\mu_k\left(P\right) - \mu_k\left(Q\right)\|_{\mathcal{H}_k}$$

Figure credit: Arthur Gretton, `http://www.gatsby.ucl.ac.uk/~gretton/papers/testing_workshop.pdf`

Testing for Independence

As well as being a distance metric on probabilities more generally, the MMD also provides a mechanism for testing if two distributions are the same: $P = Q$ iff $\mathrm{MMD}_k(P, Q) = 0$.

We can exploit this to construct a measure of the dependency between two random variables X and Y by calculating the MMD between their joint distribution P_{XY} and the product of their marginals $P_X P_Y$:

$$\begin{aligned} \Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}(X, Y) &:= \mathrm{MMD}_k^2(P_{XY}, P_X P_Y) \\ &= \|\mu_k(P_{XY}) - \mu_k(P_X P_Y)\|_{\mathcal{H}_k}^2 . \end{aligned} \tag{3}$$

This is known as the **Hilbert–Schmidt independence criterion (HSIC)**, with X and Y being independent if $\Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}(X, Y) = 0$, with the reverse also holding (i.e. $\Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}(X, Y) = 0$ implying independence) for many common kernels.

Hilbert–Schmidt Independence Criterion

- To fully define $\Xi_{k_\mathcal{X},k_\mathcal{Y}}(X,Y)$, we need to define k which needs to be a valid kernel on the product space $\mathcal{X} \times \mathcal{Y}$
- For this, we can exploit our product rule from the last lecture: given kernels $k_\mathcal{X}$ on $\mathcal{X}$ and $k_\mathcal{Y}$ on $\mathcal{Y}$, we can define $k = k_\mathcal{X} \otimes k_\mathcal{Y}$, such that

$$k\left((x,y),\left(x',y'\right)\right) = k_\mathcal{X}(x,x')k_\mathcal{Y}(y,y'),$$

  which is a valid kernel on the product domain $\mathcal{X} \times \mathcal{Y}$
- The canonical feature map is now $\varphi(x,y) = k_\mathcal{X}(\cdot,x) \otimes k_\mathcal{Y}(\cdot,y)$ and the associated RKHS comprises of functions on $\mathcal{X} \times \mathcal{Y}$

Estimating the HSIC

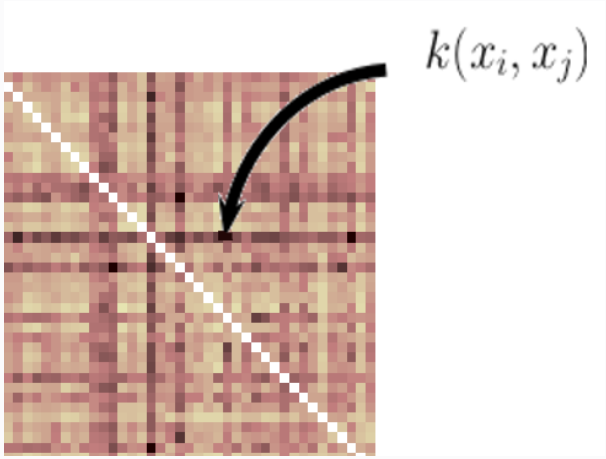
The HSIC can be estimated via the following process (see notes for derivation):

1. Draw m independent samples $\{(x_i, y_i)\}_{i=1}^m$ from the joint distribution P_{XY};
2. Construct the kernel matrices $K_{ij} = k_{\mathcal{X}}(x_i, x_j)$ and $L_{ij} = k_{\mathcal{Y}}(y_i, y_j)$;
3. Center the kernel matrices by computing $\widetilde{K} = HKH$ and $\widetilde{L} = HLH$ where $H = I - \frac{1}{m}1$ (i.e. $H_{ij} = \mathbb{I}(i = j) - \frac{1}{m}$);
4. Construct the estimate (noting $\widetilde{K}$ and $\widetilde{L}$ are symmetric)

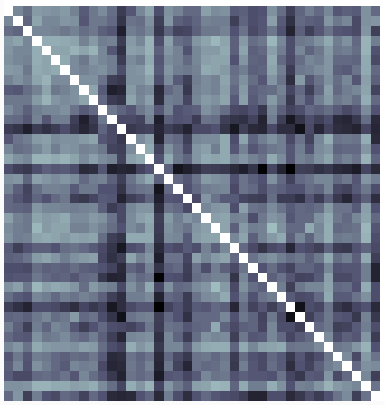
$$\widehat{\Xi_{k_{\mathcal{X}},k_{\mathcal{Y}}}}(X, Y) = \frac{1}{m^2}\mathrm{tr}\left(\widetilde{K}\widetilde{L}\right) = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\widetilde{K}_{ij}\widetilde{L}_{ij}. \quad (4)$$

HSIC Intuition

$k($  ,  $) \rightarrow \mathbf{K} =$



$\ell($ The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps... , Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament... $) \rightarrow \mathbf{L} =$



HSIC: similarity between the kernel matrices

$$\frac{1}{m^2}\left\langle \widetilde{\mathbf{K}}, \widetilde{\mathbf{L}} \right\rangle_{\mathcal{F}} = \frac{1}{m^2}\mathrm{tr}\left(\widetilde{\mathbf{K}}^\top \widetilde{\mathbf{L}}\right) = \frac{1}{m^2}\mathrm{tr}\left(\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}\right) = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\widetilde{\mathbf{K}}_{ij}\widetilde{\mathbf{L}}_{ij}$$

Recap

- The **kernel mean embedding** $\mu_k(P) = \mathbb{E}_{X\sim P} k(\cdot, X)$ provides a representation of distribution P in the RKHS of k
- $\mu_k(P)$ can be thought of as an operator that converts $f \in \mathcal{H}_k$ to their mean:

$$\langle \mu_k(P), f \rangle_{\mathcal{H}_k} = \mathbb{E}_{X\sim P} f(X), \ \forall f \in \mathcal{H}_k$$

- The **maximum mean discrepancy** $\mathrm{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}$ provides a measure of the difference between two distributions P and Q
- The **Hilbert-Schmidt Independence Criterion**

$$\begin{aligned}\Xi_{k_\mathcal{X}, k_\mathcal{Y}}(X, Y) &:= \mathrm{MMD}^2_{k_\mathcal{X}\otimes k_\mathcal{Y}}(P_{XY}, P_X P_Y) \\ &= \left\|\mu_{k_\mathcal{X}\otimes k_\mathcal{Y}}(P_{XY}) - \mu_{k_\mathcal{X}\otimes k_\mathcal{Y}}(P_X P_Y)\right\|^2_{\mathcal{H}_{k_\mathcal{X}\otimes k_\mathcal{Y}}}\end{aligned}$$

provides a measure of the dependency between X and Y

Further Reading

- Arthur Gretton's MLSS course on kernels: `http://www.gatsby.ucl.ac.uk/~gretton/teaching.html` (recommend Madrid 2018 version)