Chapter 4, Part 5: Representing Probability Distributions in RKHSs

Advanced Topics in Statistical Machine Learning

Eugenio Clerico

Hilary 2026

eugenio.clerico@stats.ox.ac.uk

Comparing measures via a RKHS

- A RKHS $\mathcal{H}$ is a space of functions $\mathcal{X} \to \mathbb{R}$.
- The reproducing kernel K can often be seen as a measure of *similarity* between points (it measures how different $h(x)$ and $h(y)$ are, for various $h \in \mathcal{H}$)
- We have a canonical feature mapping $\mathcal{X} \to \mathcal{H}$, $x \mapsto k_x$. Since on $\mathcal{H}$ we have some notion of *distance* (the norm), we can induce a notion of *pseudo-distance* on $\mathcal{X}$:

$$\Delta_K(x, y) = \|k_x - k_y\|_{\mathcal{H}} = \sqrt{K(x, x) + K(y, y) - 2K(x, y)}\,.$$

- If P and Q probabilities on $\mathcal{X}$, we can extend this idea and find a way to compare P and Q based on how different $\mathbb{E}_P[h]$ and $\mathbb{E}_Q[h]$ are for various $h \in \mathcal{H}$.

Measurability

Last time we said that regularity of K implies regularity of each $h \in \mathcal{H}$. This is true also for measurability. Henceforth we let $\mathcal{X}$ be a measurable space (e.g., $\mathbb{R}^d$ with the standard (Borel) σ-algebra).

Lemma 1

If K is measurable, then each $h \in \mathcal{H}$ is measurable.

Recall that $\mathcal{H} = \overline{\mathrm{Span}(k_x : x \in \mathcal{X})}$. So, h can be written as

$$h = \sum_{t=1}^{\infty} \alpha_t k_{x_t}$$

and the sequence converges point-wise. Each k_{x_t} is measurable, so h itself is measurable as a limit of measurable functions.

For the rest of this lecture we always assume that K is a measurable kernel! (This is essentially true for any reasonable kernel you might encounter...)

Averaging functionals

Say P is a probability measure on $\mathcal{X}$, and assume that $\mathbb{E}_P[\sqrt{K(X,X)}] = C_P < \infty$.[1]
Define $\mathrm{av}_P : \mathcal{H} \to \mathbb{R}$ as $\mathrm{av}_P(h) = \mathbb{E}_P[h]$.

Lemma 2

If $\mathbb{E}_P[\sqrt{K(X,X)}] = C_P < \infty$, av_P is linear and continuous.

It is trivial that av_P is linear (exercise). To show that av_P is continuous it is enough to show that if $\|h_n - h\|_{\mathcal{H}} \to 0$, then $\mathrm{av}_P(h_n) \to \mathrm{av}_P(h)$. Recall that, for all x, $|h_n(x) - h(x)| \leq \sqrt{K(x,x)}\|h_n - h\|_{\mathcal{H}}$ by CS. So, $|\mathrm{av}_P(h_n) - \mathrm{av}_P(h)| \leq \mathbb{E}_P[|h - h_n|] \leq \mathbb{E}_P[\sqrt{K(X,X)}]\|h_n - h\|_{\mathcal{H}} \leq C_P\|h_n - h\|_{\mathcal{H}} \to 0$.

[1] Note that this expectation always makes sense as long as K is measurable, since $K(x,x) \geq 0$ for all x...

Kernel mean embeddings

- Let P, K be such that $\mathbb{E}_P[\sqrt{K(X,X)}] < \infty$. We know that $\mathrm{av}_P : \mathcal{H} \to \mathbb{R}$ is continuous.
- We can use Riesz representation theorem to say that there is a unique $k_P \in \mathcal{H}$ such that

$$\mathrm{av}_P(h) = \langle k_P, h \rangle_{\mathcal{H}}\,, \quad \forall h \in \mathcal{H}\,.$$

- k_P is called the **kernel mean embedding** (KME) of P in $\mathcal{H}$.
- For $P = \delta_x$ (the Dirac delta on x, i.e. $\mathbb{E}_{\delta_x}[h] = h(x)$, $\forall h$), then $k_{\delta_x} = k_x$!

Properties of KMEs

Fix any P and Q probability measures such that $\mathbb{E}_P[\sqrt{K(X,X)}] < \infty$ and $\mathbb{E}_Q[\sqrt{K(X,X)}] < \infty$. Note that for bounded kernels ($K(x,x) \leq C < \infty$ for all x) this is automatic!

- $k_P \in \mathcal{H}$, so k_P is a function from $\mathcal{X} \to \mathbb{R}$. We have $k_P(x) = \langle k_x, k_P \rangle_{\mathcal{H}} = \mathbb{E}_P[k_x(X)] = \int_{\mathcal{X}} K(x,y) \mathrm{d}P(y)$.
- k_P can be defined as $\mathbb{E}_P[k_X]$, as an *average* of functions[2]: $\mathbb{E}_P[k_X](x) = \mathbb{E}_P[k_X(x)] = \int_{\mathcal{X}} K(y,x) \mathrm{d}P(y) = k_P(x)$. This allows us to write that $\mathbb{E}_P[\langle k_X, h \rangle_{\mathcal{H}}] = \langle \mathbb{E}_P[k_X], h \rangle_{\mathcal{H}}$.
- $\langle k_P, k_Q \rangle_{\mathcal{H}} = \mathbb{E}_{X \sim P, Y \sim Q}[K(X,Y)]$.
- $\|k_P\|_{\mathcal{H}} = \sqrt{\mathbb{E}_{X \sim P, X' \sim P}[K(X,X')]}$.

[2] This can be rigorously defined as a Bochner integral.

Characteristic kernels

- After embedding P and Q in $\mathcal{H}$, we might use k_P and k_Q to **compare** these two distribution. However, not every kernel can distinguish between different distributions!
- Say $K(x, y) = x \cdot y$. Then, $k_P(x) = \mathbb{E}_P[k_x] = x \cdot \mathbb{E}_P[X]$. In particular, if $\mathbb{E}_Q[X] = \mathbb{E}_P[X]$, then $k_P = k_Q$.
- More generally, for a kernel $K(x, y) = \phi(x) \cdot \phi(y)$ (where $\phi : \mathcal{X} \to \mathbb{R}^p$), $k_Q = k_P$ whenever $\mathbb{E}_P[\phi(X)] = \mathbb{E}_Q[\phi(X)]$.
- To always be able to distinguish P from Q, the mapping $P \mapsto k_P$ must be **injective**. A kernel for which this is true is called a **characteristic kernel**.
- The RBF kernel and Matérn kernels are characteristic kernels.

Comparing Distributions

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

Such distances are called **maximum mean discrepancies** (MMDs). Fixed a kernel K (and so a RKHS $\mathcal{H}$) we define

$$\Delta_K(P, Q) = \|k_P - k_Q\|_{\mathcal{H}}.$$

For characteristic kernels, the MMD is a proper metric on probability distributions: $\Delta_K(P, Q) = 0$ if and only if $P = Q$

Variational formulation of MMD

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.

Lemma 3

$$\Delta_K(P, Q) = \sup_{h\,:\,\|h\|_{\mathcal{H}}=1} \left(\mathbb{E}_P[h] - \mathbb{E}_Q[h]\right).$$

If $k_P = k_Q$ it's trivial, so assume $k_P \neq k_Q$.
Let $h^\star = \frac{k_P - k_Q}{\|k_P - k_Q\|_{\mathcal{H}}}$, then $\mathbb{E}_P[h^\star] - \mathbb{E}_Q[h^\star] = \|k_P - k_Q\|_{\mathcal{H}}$. So, $\Delta_K(P, Q) \leq \sup_{\|h\|_{\mathcal{H}}=1}(\mathbb{E}_P[h] - \mathbb{E}_Q[h])$.
For all $h \in \mathcal{H}$, such that $\|h\|_{\mathcal{H}} = 1$, we have

$$\mathbb{E}_P[h] - \mathbb{E}_Q[h] = \langle k_P - k_Q, h\rangle_{\mathcal{H}} \leq \|k_P - k_Q\|_{\mathcal{H}}$$

by Cauchy-Schwarz. So, $\Delta_K(P, Q) \geq \sup_{\|h\|_{\mathcal{H}}=1}(\mathbb{E}_P[h] - \mathbb{E}_Q[h])$.

Testing for Independence

When K is characteristic, the MMD provides a mechanism for testing if two distributions are the same: $P = Q$ iff $\Delta_K(P, Q) = 0$.

We can exploit this to construct a measure of the dependency between two random variables. Let X be valued in $\mathcal{X}$ and Y in $\mathcal{Y}$. Fix a kernel K on $\mathcal{X} \times \mathcal{Y}$. Let $P_{X,Y}$ be the joint distribution of X and Y, and $P_X \otimes P_Y$ the product of the marginals. Let

$$\Xi_K(X, Y) = \Delta_K^2 \left(P_{X,Y}, P_X \otimes P_Y\right) .$$

The **Hilbert–Schmidt independence criterion (HSIC)** states then $\Xi_K(X, Y) = 0$ is a necessary condition for X and Y to be independent. If K is **characteristic**, then it is also a sufficient condition.

Hilbert–Schmidt Independence Criterion

- To define $\Xi_K(X, Y)$, we need a kernel K on the product space $\mathcal{X} \times \mathcal{Y}$
- For this, we can exploit the tensor product rule: given kernels $K^{\mathcal{X}}$ on $\mathcal{X}$ and $K^{\mathcal{Y}}$ on $\mathcal{Y}$, we can define $K = K^{\mathcal{X}} \otimes K^{\mathcal{Y}}$, so that $K((x, y), (x', y')) = K^{\mathcal{X}}(x, x')K^{\mathcal{Y}}(y, y')$.
- Note that even if $K^{\mathcal{X}}$ and $K^{\mathcal{Y}}$ are characteristic, it might be that $K = K^{\mathcal{X}} \otimes K^{\mathcal{Y}}$ is not. However, if $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^{d'}$, $K^{\mathcal{X}}(x, x') = \kappa(\|x - x'\|)$ and $K^{\mathcal{Y}}(y, y') = \tilde{\kappa}(\|y - y'\|)$ for some continuous functions κ and $\tilde{\kappa}$, and both $K^{\mathcal{X}}$ and $K^{\mathcal{Y}}$ are characteristic, then K is also characteristic.

Empirical KME

Often P is unknown: we have only sample $X_1 \dots X_n$ iid from P.
We can define the **empirical distribution**: $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.
The empirical KME is $\hat{k}_n = k_{\hat{P}_n} = \frac{1}{n}\sum_{i=1}^n K(X_i, \cdot)$.
NB: $\hat{k}_n$ and $\hat{P}_n$ are random, as the sample is random!

Lemma 4

If $M_P^2 = \mathbb{E}_P[K(X,X)] < \infty$, *then* $\mathbb{E}[\Delta_K(\hat{P}_n, P)^2] \leq M_P^2/n$.

$\mathbb{E}[\|\hat{k}_n\|_{\mathcal{H}}^2] = \frac{1}{n}\mathbb{E}_P[K(X,X)] + \frac{n-1}{n}\mathbb{E}_{P\otimes P}[K(X,X')] = \frac{M_P^2}{n} + \frac{n-1}{n}\|k_P\|_{\mathcal{H}}^2$;
$\mathbb{E}[\langle \hat{k}_n, k_P\rangle_{\mathcal{H}}] = \frac{1}{n}\sum_{i=1}^n \mathrm{av}_P(k_P) = \|k_P\|_{\mathcal{H}}^2$;
$\mathbb{E}[\Delta_K(\hat{P}_n, P)^2] = \mathbb{E}[\|\hat{k}_n\|_{\mathcal{H}}^2] - 2\,\mathbb{E}[\langle \hat{k}_n, k_P\rangle_{\mathcal{H}}] + \|k_P\|_{\mathcal{H}}^2 = \frac{M_P^2 - \|k_P\|_{\mathcal{H}}^2}{n}$.

Estimating the MMD

Given sets of independent samples $\{X_i\}_{i=1}^n \overset{i.i.d.}{\sim} P$, $\{Y_i\}_{i=1}^m \overset{i.i.d.}{\sim} Q$.
A simple estimator of the squared MMD in is given by

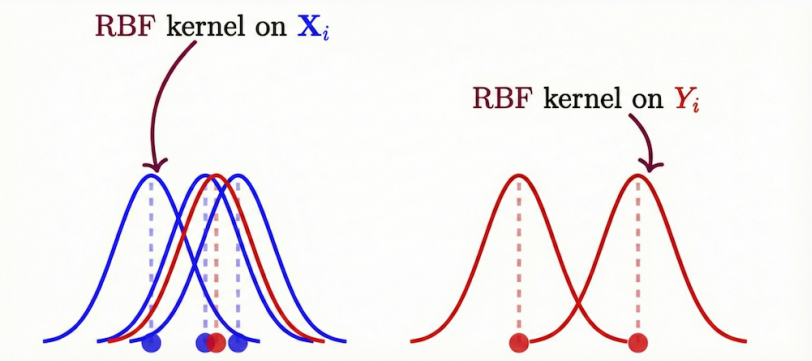
$$\Delta_K(\hat{P}_n, \hat{Q}_m)^2 = \sum_{i,j} \frac{K(X_i,X_j)}{n^2} + \sum_{i,j} \frac{K(Y_i,Y_j)}{m^2} - 2\sum_{i,j} \frac{K(X_i,Y_j)}{mn}.$$

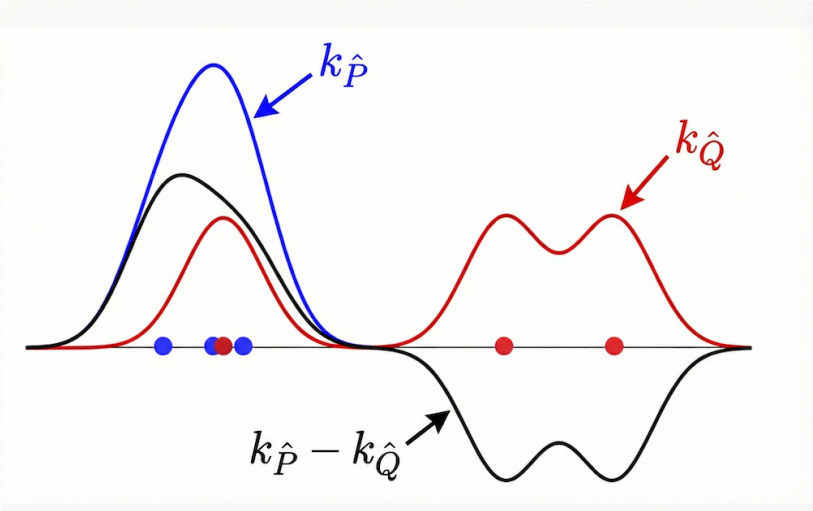
Though its cost scales as $O((n+m)^2)$, it only requires samples from each distribution: most divergences (e.g. KL) require at least one of the density functions. This leads to applications in non–parametric hypothesis testing and training implicit models (e.g. GANs).

Note that $\Delta_K(\hat{P}_n, \hat{Q}_m)^2$ is a biased estimator. However, from the previous lemma: $\mathbb{E}[(\Delta_K(\hat{P}_n, \hat{Q}_m) - \Delta_K(P,Q))^2] \leq 2\frac{M_P^2}{n} + 2\frac{M_Q^2}{m}$.

Exercise: Compute $\mathbb{E}[\Delta_K(\hat{P}_n, \hat{Q}_m)^2]$. Removing the diagonal terms $K(X_i, X_i)$ and $K(Y_i, Y_i)$, find an unbiased estimator for $\Delta_K(P,Q)^2$.

Visualization





$$\Delta_K\left(\hat{P}, \hat{Q}\right) = \|k_{\hat{P}} - k_{\hat{Q}}\|_{\mathcal{H}}$$

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

Kernel Density Estimators

Kernel density estimation (KDE) is a simple classical non–parametric density estimation approach that forms the (Lebesgue) density estimate

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

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

The function $\hat{p} : \mathcal{X} \mapsto \mathbb{R}^+$ is simply the empirical KME of K.[3]

[3] Note however that sometimes the term KDE is used also for similar procedure that use a non-negative and normalised function K which is not a kernel (not PD).

Viewing histograms as KDE

A simple example of KDE is given by histograms. For simplicity say that $\mathcal{X} = [0, 1)$. Fix a number of bins B and let $h = 1/B$. Each $x \in [0, 1)$ belongs to a single bin $b_j = [(j-1)h, jh)$, where $j = 1 \dots B$. In particular, let $J(x)$ denote the index of the bin of x. Define

$$K(x, x') = \begin{cases} 1/h & \text{if } J(x) = J(x'); \\ 0 & \text{otherwise.} \end{cases}$$

Exercise: Check that K is a kernel function, that it is non-negative and normalised.

The KDE with K returns exactly a normalised histogram: to the bin j is associated the value $\#\{i : J(x_i) = j\}/(hn)$.

The resulting $\hat{p}$ is a discontinuous density. Using a smoother kernel leads to smoother estimated densities.

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 $\mathcal{X} = \mathbb{R}^d$ then we can use a normalised 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

Further Reading

- Arthur Gretton's MLSS course on kernels: `http://www.gatsby.ucl.ac.uk/~gretton/teaching.html` (recommend Madrid 2018 version)
- For a more technical overview: Kernel Mean Embedding of Distributions: A Review and Beyond, Muandet et al., 2017. (Ofc, this goes far beyond the scope of this course!)