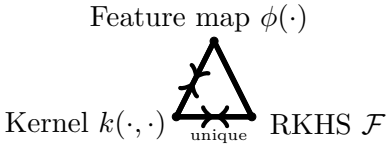
12 Kernel Methods: Properties & Applications

We have focused on showing that the three key concepts in kernel methods, 1- the feature map ϕ, 2- the kernel k, and 3- Reproducing Kernel Hilbert Space (RKHS) commute according to the following diagram.



Along the way, we derived a few key properties associated with the kernel and its RKHS that will be useful in the sequel. We recall them below.

- Reproducing property: Function evaluations can be written as inner products

$$f(x) = \langle R_x, f\rangle = \langle k(x,\cdot), f\rangle, \quad f \in \mathcal{F}, x \in \mathcal{X}.$$

- Moore-Aronszajn theorem: Given a kernel k, its RKHS is set of functions $f, g \in \mathcal{F}$ given as

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) \text{ for some } \alpha_i \text{ and } g(x) = \sum_{i=1}^{n} \beta_i k(x, x_i) \text{ for some } \beta_i$$

$$\begin{aligned}\text{Inner product in } \mathcal{F}: \ \langle f, g\rangle &= \left\langle \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \sum_{i=1}^{n} \beta_i k(\cdot, x_i) \right\rangle \\ &= \sum_{ij} \alpha_i \beta_j \underbrace{\langle k(\cdot, x_i), k(\cdot, x_j)\rangle}_{=k(x_i,x_j) \text{ by reproducing prop}} \\ &= \sum_{ij} \alpha_i \beta_j k(x_i, x_j).\end{aligned}$$

Perhaps, the above properties of kernels are the most useful ones as far as machine learning is concerned. We first start with a few more basics related to kernels and continue with some applications in machine learning.

12.1 Basic properties and examples

We first look at a few simple examples of kernels and identify their associated RKHS.
Example. [Linear kernel] Consider the kernel $k(x, x') = \langle x, x'\rangle$ where $x, x' \in \mathcal{X} = \mathbb{R}^d$. The RKHS for this kernel can be written as

$$\begin{aligned}\text{RKHS}(k) \ = \mathcal{F} &= \left\{ f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i), \forall n \in \mathbb{N}, \forall x_i \in \mathbb{R}^d, \forall \alpha_i \in \mathbb{R} \right\} \\ &= \left\{ f(x) = \sum_{i=1}^{n} \alpha_i \langle x, x_i\rangle, \forall n \in \mathbb{N}, \forall x_i \in \mathbb{R}^d, \forall \alpha_i \in \mathbb{R} \right\} \\ &= \{ f(x) = \left\langle x, \sum_{i=1}^{n} \alpha_i x_i \right\rangle, \forall n \in \mathbb{N}, \forall x_i \in \mathbb{R}^d, \forall \alpha_i \in \mathbb{R} \} \\ &= \{ f(x) = \langle x, \beta\rangle, \beta \in \mathbb{R}^d \} \quad \text{since } \mathbb{R}^d \text{ is a vector space.}\end{aligned}$$

Accordingly, for $f(x) = \langle \beta, x \rangle = \sum_{i=1}^n \alpha_i \langle x_i, x \rangle$ and $f'(x) = \langle \beta', x \rangle = \sum_{i=1}^n \alpha'_i \langle x'_i, x \rangle$, the inner product is given as

$$\langle f, f' \rangle = \langle \beta, \beta' \rangle.$$

This can be seen by writing

$$\begin{aligned}
\langle f, f' \rangle &= \sum_{ij} \alpha_i \alpha'_j k(x_i, x'_j) \\
&= \sum_{ij} \alpha_i \alpha'_j \langle x_i, x'_j \rangle \\
&= \left\langle \sum_{i=1}^n \alpha_i x_i, \sum_{j=1}^n \alpha'_j x'_j \right\rangle \\
&= \langle \beta, \beta' \rangle.
\end{aligned}$$

Example. [Common kernels]

- **Identity kernel** is given as $k(x, x') = 1$. This is a kernel since for any $x_1, ..., x_n$, the kernel matrix defined as $K_{ij} = k(x_i, x_j) = 1$ is a matrix of 1's, and it is positive semidefinite.
- **Indicator function** $k(x, x') = \mathbb{1}_{\{\|x - x'\| \leq 0\}}$ is a kernel since the kernel matrix K is an identity matrix (when $x_i \neq x_j$).
- **Indicator function** $k(x, x') = \mathbb{1}_{\{\|x - x'\| \leq \epsilon\}}$ for $\epsilon > 0$ is **not** a kernel. This can be seen by choosing $x_1 = 0$, $x_2 = \epsilon e_1$ and $x_3 = 2\epsilon e_1$ where $e_1 = [1, 0, 0, ...]^\top$ is the first standard basis vector in $\mathbb{R}^d$. The kernel matrix $K = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ is not positive semi-definite (exercise).
- **Polynomial kernel** is given as $k(x, x') = (1 + \langle x, x' \rangle)^p$ for $p \in \mathbb{N}$. We will verify that $k(x, x')$ is a kernel shortly.
- **Gaussian kernel** is given as $k(x, x') = \exp\{-\frac{1}{2\sigma^2} \|x - x'\|^2\}$. Here, σ^2 determines the width of the kernel. Large σ^2 corresponds to smoother kernel. We will verify that $k(x, x')$ is a kernel shortly. What happens when $\sigma^2 \downarrow 0$?

In the sequel, we discuss some key properties of kernels.

1. **Inner product:** A function of the form $k(x, x') = \langle \phi(x), \phi(x') \rangle$ is a kernel (See Theorem 53).
2. **Summation:** Summation of two kernels is a kernel $k(x, x') = k_1(x, x') + k_2(x, x')$. This can be seen by considering the summation two PSD kernels K_1 and K_2 associated with the kernels k_1 and k_2, respectively, and showing that it is in fact PSD.

$$\forall u \in \mathbb{R}^d, \quad \langle u, Ku \rangle = \langle u, (K_1 + K_2)u \rangle = \langle u, K_1 u \rangle + \langle u, K_2 u \rangle \geq 0.$$

3. **Elementwise product:** (Hadamard) product of two kernels is a kernel $k(x, x') = k_1(x, x') \cdot k_2(x, x')$. Because the kernel matrices K_1 and K_2 are PSD, and we can write the following eigenvalue decomposition.

$$K_1 = UDU^T = \sum_k d_k u_k u_k^T \quad \text{and} \quad K_2 = VBV^T = \sum_k b_k v_k v_k^T$$

Here, U and V are orthogonal matrices, and D and B are diagonal matrices with nonnegative entries $d_i, b_i \geq 0$. We write

$$[K_1]_{ij} = \sum_k d_k u_{ki} u_{kj} \quad \text{and} \quad [K_2]_{ij} = \sum_k b_k v_{ki} v_{kj}$$

$$\begin{aligned}[K]_{ij} = [K_1]_{ij}[K_2]_{ij} &= \left(\sum_k d_k u_{ki} u_{kj}\right)\left(\sum_l b_l v_{li} v_{lj}\right) \\ &= \sum_{kl} d_k b_l (u_{ki} v_{li})(u_{kj} v_{lj}) \\ K &= \sum_{kl} d_k b_l (u_k \circ v_l)(u_k \circ v_l)^T \succeq 0.\end{aligned}$$

Now, we can go back and verify that polynomial and Gaussian kernels are valid kernels. We start with the polynomial kernel $k(x, x') = (1 + \langle x, x' \rangle)^p$.

1. $\langle x, x' \rangle$ is a kernel by the **inner product** property.
2. 1 is the identity kernel, so $1 + \langle x, x' \rangle$ is kernel by the **summation** property.
3. Since $1 + \langle x, x' \rangle$ is a kernel, $(1 + \langle x, x' \rangle)^p$ is kernel by the **product** property.

Using these properties, we can also verify that the Gaussian kernel is a valid kernel. The trick is to write a Taylor's series expansion.

$$\begin{aligned}k(x, x') &= \exp\left\{ -\frac{\|x - x'\|^2}{2\sigma^2} \right\} \\ &= \underbrace{\exp\left\{ -\frac{\|x\|^2}{2\sigma^2} \right\} \cdot \exp\left\{ -\frac{\|x'\|^2}{2\sigma^2} \right\}}_{k_1(x,x')} \cdot \underbrace{\exp\left\{ \frac{\langle x, x' \rangle}{\sigma^2} \right\}}_{k_2(x,x')}\end{aligned}$$

Clearly k_1 above is a valid kernel by the **inner product** property with $\phi(x) = \exp\left\{ -\frac{\|x\|^2}{2\sigma^2} \right\}$. If k_2 is a valid kernel, the **product** property will ensure that Gaussian kernel is a valid kernel. But we have

$$k_2(x, x') = \exp\left\{ \frac{\langle x, x' \rangle}{\sigma^2} \right\} = \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{\langle x, x' \rangle}{\sigma^2} \right)^i \quad \text{by the Taylor series of } e^x.$$

Since k_2 is a sum of polynomial kernels, it is also a valid kernel.

12.2 Learning with kernels

Let's turn our attention to applications of kernels in machine learning. Suppose that we collected a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, and we would like to fit a function $y \approx f(x)$ using the function class $\mathcal{F}$ which is a RKHS. We consider the ℓ_2 regularized empirical risk minimization problem,

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2. \tag{12.1}$$

Theorem 60 (Representer theorem). *For the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, and a kernel $k(x, x')$, we define the set of functions*

$$\mathcal{V}_\mathcal{D} = \Big\{ f(x) = \sum_{i=1}^n \alpha_i k(x, x_i) : \alpha_i \in \mathbb{R} \text{ for } i = 1, ..., n \Big\}.$$

Then, $\hat{f}$ in (12.1) belongs to $\mathcal{V}_\mathcal{D}$, i.e. $\hat{f} \in \mathcal{V}_D \subset \mathcal{F}$.

Remark. Representer theorem has a remarkable algorithmic consequence. It tells us that minimizing over the entire RKHS $\mathcal{F}$ is equivalent to minimizing over $\mathcal{V}_\mathcal{D}$. This will reduce the empirical risk minimization problem (12.1) to a simple least squares problem over α_i's. We will revisit this after proving the theorem.

Proof.

- First, we note that in the definition of $\mathcal{V}_\mathcal{D}$, n is the number of samples and x_i's are input data, and both are fixed, whereas we recall from Moore-Aronszajn Theorem 59 that,

  $$\mathcal{F} = \Big\{ f(x) = \sum_{j=1}^m \alpha'_j k(x, x'_j), \quad \forall m \in \mathbb{N}, \forall \alpha'_j \in \mathbb{R}, \forall x'_j \in \mathbb{R}^d \Big\}.$$

  Also, we notice that $\mathcal{V}_\mathcal{D}$ is a subspace in $\mathcal{F}$ (exercise).

- We define the orthogonal complements of the subspace $\mathcal{V}_\mathcal{D}$ as

  $$\mathcal{V}_D^\perp = \{ f' \in \mathcal{F} : \ \langle f', f \rangle = 0 \ \forall f \in \mathcal{V}_D \}.$$

  A vector space is the summation of a subspace and its orthogonal complement. This enables us to write a function in RKHS $\mathcal{F}$ as the summation of a parallel and a orthogonal component (not union!). That is, $\forall f \in \mathcal{F}$, we can write

  $$f(x) = f^\parallel(x) + f^\perp(x)$$

  where $f^\parallel \in \mathcal{V}_\mathcal{D}$ and $f^\perp \in \mathcal{V}_\mathcal{D}^\perp$. In other words, projection of f on $\mathcal{V}_\mathcal{D}$ is $f^\parallel$, and on $\mathcal{V}_\mathcal{D}^\perp$ is $f^\perp$.

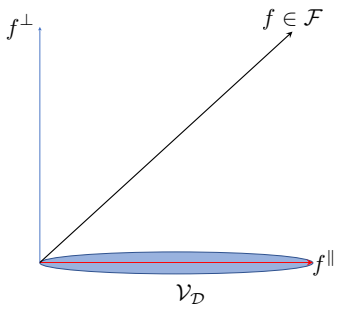


Figure 4: Decomposing RKHS into two subspaces.

- But notice that for $(x_i, y_i) \in \mathcal{D}$, by the reproducing property, we can write for $f \in \mathcal{F}$

  $$f^\perp(x_i) = \langle \underbrace{f^\perp}_{\in \mathcal{V}_\mathcal{D}^\perp}, \underbrace{k(x_i, \cdot)}_{\in \mathcal{V}_\mathcal{D}} \rangle = 0$$

where the last step follows since $k(x_i, \cdot) \in \mathcal{V}_\mathcal{D}$ and $f^\perp \in \mathcal{V}_\mathcal{D}^\perp$.

Therefore for $(x_i, y_i) \in \mathcal{D}$, we have $f(x_i) = f^\parallel(x_i) + f^\perp(x_i) = f^\parallel(x_i)$. This implies that the loss over f only depends on its projection onto $\mathcal{V}_\mathcal{D}$, i.e.

$$\ell(y_i, f(x_i)) = \ell(y_i, f^\parallel(x_i)).$$

Consequently, the training error of f only depends on its projection $f^\parallel$.

- For the regularizer, we have for every $f \in \mathcal{F}$

$$\|f\|_\mathcal{F}^2 = \|f^\parallel\|_\mathcal{F}^2 + \|f^\perp\|_\mathcal{F}^2.$$

- Combining these, we obtain that the minimization problem over f can be written as

$$\begin{aligned}\hat{f} &= \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) + \frac{\lambda}{2}\|f\|_\mathcal{F}^2, \\ &= \operatorname*{argmin}_{\substack{f = f^\parallel + f^\perp: \\ f^\parallel \in \mathcal{V}_\mathcal{D},\ f^\perp \in \mathcal{V}_\mathcal{D}^\perp}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f^\parallel(x_i)) + \frac{\lambda}{2}\|f^\parallel\|_\mathcal{F}^2 + \frac{\lambda}{2}\|f^\perp\|_\mathcal{F}^2.\end{aligned}$$

Since $f^\perp$ doesn't affect the training error, we might as well choose it to be zero so that the regularizer becomes smaller. Thus, the minimizer $\hat{f}$ can be obtained by just minimizing over $f^\parallel \in \mathcal{V}_\mathcal{D}$.

□

Remark. The above proof also holds for any loss function $\ell(\{x, y, f(x)\})$, and regularizer $r(\|f\|_\mathcal{F})$ where r is monotone and strictly increasing.

Example. [Squared error loss] We choose $\ell(y, f(x)) = \frac{1}{2}(y - f(x))^2$ and the empirical risk minimization in (12.1) reduces to

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{2}\sum_{i=1}^n (y_i - f(x_i))^2 + \frac{\lambda}{2}\|f\|_\mathcal{F}^2.$$

Here, $\mathcal{F}$ is a RKHS. Applying the representer theorem, we obtain that the above minimizer has to satisfy

$$\hat{f}(x) = \sum_{j=1}^n \alpha_j k(x, x_j),$$

where $\alpha_i \in \mathbb{R}$. Note that the only thing that is not known to us is α_j's, which we will use our data to learn. Concatenating α_j's, we define $\alpha = [\alpha_1, ..., \alpha_n]^\top \in \mathbb{R}^n$, and we can convert the original problem to a minimization over α.

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \frac{1}{2}\sum_{i=1}^n \Big(y_i - \sum_{j=1}^n \alpha_j k(x_i, x_j)\Big)^2 + \frac{\lambda}{2}\|f\|_\mathcal{F}^2$$

Recall the definition of kernel matrix $K_{ij} = k(x_i, x_j)$ and notice that

$$\begin{aligned}\|f\|_\mathcal{F}^2 = \langle f, f\rangle &= \Big\langle \sum_i \alpha_i k(\cdot, x_i), \sum_j \alpha_j k(\cdot, x_j)\Big\rangle \\ &= \sum_i \sum_j \alpha_i \alpha_j \langle k(\cdot, x_i), k(\cdot, x_j)\rangle \quad \text{(by prop of inner prod)} \\ &= \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) = \alpha^T K \alpha \quad \text{(by def of inner prod in RKHS)}\end{aligned}$$

For the training error, we have

$$\frac{1}{2}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{n}\alpha_j k(x_i, x_j)\Big)^2 = \frac{1}{2}\sum_{i=1}^{n}\Big(y_i - \langle K_i, \alpha\rangle\Big)^2 \quad K_i \text{ is the } i\text{-th row of } K$$
$$= \frac{1}{2}\|y - K\alpha\|_2^2.$$

Therefore, the problem reduces to

$$\hat{\alpha} = \underset{\alpha\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\|y - K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^T K\alpha \coloneqq \hat{R}(\alpha)$$

which can be easily solved by taking derivatives and solving for α

$$\nabla_\alpha \hat{R}(\alpha) = K(y - K\alpha) + \lambda K\alpha = 0$$
$$\implies \hat{\alpha} = (K + \lambda I_n)^{-1} y.$$

Note that the solution is not unique unless $K \succ 0$. If $\hat{\alpha}$ is a solution, so is $\hat{\alpha} + u$ where u belongs to the null space of K.

12.3 Maximum mean discrepancy (MMD)

In this section, we discuss another example of kernel methods, that is RKHS embeddings of probability distributions. We start with a few definitions.

Definition 61 (∞**-norm**). *For a function $f : \mathcal{X} \to \mathbb{R}$, infinity norm is given as $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. This defines the following metric,*

$$\|f - f'\|_\infty = \sup_{x\in\mathcal{X}} |f(x) - f'(x)|.$$

The above metric measures the worst case difference between two functions. It is often the case that we want to measure the difference between two probability distributions. For this, we have distance measures such as KL-divergence, total variation, Wasserstein distance etc. We should note that not all of these are proper metrics. The following is another way to measure distance between two probability distributions.

Definition 62 (**Maximum mean discrepancy (MMD)**). *Let p, q be probability distributions on $\mathcal{X}$. Define MMD as*

$$d_{\mathcal{F}}(p, q) = \sup_{f\in\mathcal{F}} \big|\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)]\big|.$$

The above definition seems like something that can be useful in practice. But as in any learning algorithm, we need to choose a reference function class.

- *How complex the function class (set of test functions) $\mathcal{F}$ should be so that the above metric is good, i.e., $d_{\mathcal{F}}(p, q) = 0$ if and only if $p = q$?*

At least one side is obvious, if $p = q$ then $d_{\mathcal{F}}(p, q) = 0$. For the other side, as a starter, we have that if $\mathcal{F}$ is a 1-Lipschitz continuous functions on $\mathcal{X}$ denoted by L_1, Monge-Kantorovich duality says the above MMD metric reduces to Wasserstein-1 distance. That is,

$$d_{L_1}(p, q) = \mathcal{W}_1(p, q) \triangleq \inf_{\substack{\text{couplings } (x,y)\\ x\sim p,\ y\sim q}} \mathbb{E}[\|x - y\|].$$

Another function class that proves the above metric useful is the class of bounded continuous functions on $\mathcal{X}$, which we denote by C_0.

Theorem 63 (Dudley's result on MMD)**.** *If the function class is the set of all bounded continuous functions $\mathcal{F} = C_0$, then $d_{C_0}(p, q) = 0$ if and only if $p = q$.*

But taking supremum over L_1 or C_0 may be too much to ask since these function classes are too complex. If we can find a good representation of, say C_0, then we may be able to come up with something useful.

In lieu of Dudley's result on MMD, Theorem 63, we will require the RKHS defined by its unique kernel to be representative of the space of bounded continuous functions.

Definition 64 (**Universal Kernel**)**.** *For the set C_0 of all bounded continuous functions $f : \mathcal{X} \to \mathbb{R}$, we call a kernel k* a universal kernel *if its RKHS $\mathcal{F}$ is dense in C_0.*

Note that $\mathcal{F}$ is dense in C_0 if for every function $f \in C_0$, $\forall \epsilon > 0$, there exists $f' \in \mathcal{F}$ such that

$$\|f - f'\|_\infty \leq \epsilon.$$

The notion of universality is exactly what we need from an RKHS $\mathcal{F}$ to be a good representation of C_0. Indeed, this property translates the desired feature of C_0 to RKHS.

Theorem 65 (Steinwart's theorem on unit RKHS ball)**.** *Define the unit ball centered at origin*

$$\mathcal{G} = \{f \in \mathcal{F} : \|f\|_\mathcal{F} \leq 1\}$$

where $\mathcal{F}$ is a RKHS of a universal kernel k, then $d_\mathcal{G}(p, q) = 0 \Leftrightarrow p = q$.

Remark.

- The ball doesn't need to be unit, that is, the radius of the ball can be arbitrary as long as it is non-zero.

Proof. One side is obvious as before. For the other side, we assume that $d_\mathcal{G}(p, q) = 0$ but $p \neq q$ and hope to achieve contradiction.

- If $p \neq q$, this implies by Theorem 63, that $d_{C_0}(p, q) = \epsilon$ for some $\epsilon > 0$. Hence, there exists a function $h \in C_0$ such that

  $$|\mathbb{E}_p[h(x)] - \mathbb{E}_q[h(y)]| = \epsilon,$$

  since C_0 is compact. h may not belong to $\mathcal{F}$.

- But since k is a universal kernel, $\mathcal{F}$ is dense in C_0 which implies that there exists $f \in \mathcal{F}$ such that $\|f - h\|_\infty \leq \epsilon/3$, which in turn implies that

  $$|\mathbb{E}_p[f(x)] - \mathbb{E}_p[h(x)]| \leq \frac{\epsilon}{3} \quad \text{and} \quad |\mathbb{E}_q[f(x)] - \mathbb{E}_q[h(x)]| \leq \frac{\epsilon}{3}. \tag{12.2}$$

  To see this, we can write

  $$\begin{aligned} |\mathbb{E}_p[f(x)] - \mathbb{E}_p[h(x)]| &= \left| \int [f(x) - h(x)] dp(x) \right| \\ &\leq \int |f(x) - h(x)| dp(x) \\ &\leq \int \sup_{x \in \mathcal{X}} |f(x) - h(x)| dp(x) \\ &= \int \|f - h\|_\infty dp(x) = \|f - h\|_\infty \leq \epsilon/3. \end{aligned}$$

  The other term can be bounded by following the same steps.

- We should be careful about that $f \in \mathcal{F}$ may not belong to $\mathcal{G}$. This is okay since we can update $h \to h/\|f\|_{\mathcal{F}}$ and $\epsilon \to \epsilon/\|f\|_{\mathcal{F}}$ so that $f \to f/\|f\|_{\mathcal{F}} \in \mathcal{G}$.
- We proceed with the triangle inequality,

$$\begin{aligned}\epsilon =& |\mathbb{E}_p[h(x)] - \mathbb{E}_q[h(y)]| \\ =& |\mathbb{E}_p[h(x)] \pm \mathbb{E}_p[f(x)] \pm \mathbb{E}_q[f(x)] - \mathbb{E}_q[h(y)]| \\ \leq& \underbrace{|\mathbb{E}_p[h(x)] - \mathbb{E}_p[f(x)]|}_{\leq \epsilon/3 \text{ by (12.2)}} + \underbrace{|\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)]|}_{d_{\mathcal{G}}(p,q)} + \underbrace{|\mathbb{E}_q[f(x)] - \mathbb{E}_q[h(y)]|}_{\leq \epsilon/3 \text{ by (12.2)}}\end{aligned}$$

  Notice that the first and last terms are smaller than $\epsilon/3$ and the summation of all three terms is equal to ϵ. This implies that the second term $d_{\mathcal{G}}(p, q)$ has to be larger than $\epsilon/3$ which contradicts with our initial assumption that $d_{\mathcal{G}}(p, q) = 0$.

□

Seems like the unit ball $\mathcal{G}$ is a representative function class to work with. We don't have access to the expectations, but we may be able to leverage some data and estimate the MMD metric. But how can we compute the MMD $d_{\mathcal{G}}(p, q)$ between two probability distributions p and q in practice? The answer is given by the reproducing property of the kernel k associated to its unique RKHS $\mathcal{F}$.

Recall the reproducing property of kernels that says the representer $R_x = k(x, \cdot)$ satisfies $\langle R_x, f\rangle = f(x)$. Thus, for $f \in \mathcal{F}$, we can write

$$\mathbb{E}_p[f(x)] = \mathbb{E}_p[\langle f, \underbrace{k(x, \cdot)}_{\text{random}}\rangle_{\mathcal{F}}] = \langle f, \underbrace{\mathbb{E}_p[k(x, \cdot)]}_{\triangleq \mu_p}\rangle_{\mathcal{F}} = \langle f, \mu_p\rangle_{\mathcal{F}}$$

where we defined the RKHS embedding of the probability distribution p as $\mu_p = \mathbb{E}_p[k(x, \cdot)]$. This tells us that expectations under the distribution p can be written as inner products. Therefore we can rewrite the MMD metric in its simple form

$$\begin{aligned}d_{\mathcal{G}}(p, q) &= \sup_{f \in \mathcal{G}} |\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(y)]| \\ &= \sup_{f \in \mathcal{G}} |\langle f, \mu_p - \mu_q\rangle_{\mathcal{F}}| \\ &= \|\mu_p - \mu_q\|_{\mathcal{F}},\end{aligned}$$

where the last equality follows from $\sup_{f:\|f\|_{\mathcal{G}} \leq 1} \langle f, \mu\rangle = \|\mu\|_{\mathcal{F}}$. It turns out that MMD between p and q was just the distance between their corresponding RKHS embeddings.

Now that we converted MMD to a simple norm, we can do a lot. We write,

$$\begin{aligned}d_{\mathcal{G}}(p, q)^2 &= \|\mu_p - \mu_q\|_{\mathcal{F}}^2 \\ &= \|\mu_p\|_{\mathcal{F}}^2 + \|\mu_q\|_{\mathcal{F}}^2 - \langle \mu_p, \mu_q\rangle_{\mathcal{F}} - \langle \mu_q, \mu_p\rangle_{\mathcal{F}}.\end{aligned} \tag{12.3}$$

Note that

$$\begin{aligned}\|\mu_p\|_{\mathcal{G}}^2 = \langle \mu_p, \mu_p\rangle &= \langle \mathbb{E}_p[k(x, \cdot)], \mathbb{E}_p[k(x, \cdot)]\rangle_{\mathcal{F}} \\ &= \mathbb{E}_p[\langle k(x, \cdot), k(x', \cdot)\rangle_{\mathcal{F}}] \quad \text{where} \quad x, x' \sim p \text{ are independent.} \\ &= \mathbb{E}_{p,p}[k(x, x')].\end{aligned}$$

Similarly, we have

$$\begin{aligned}\langle \mu_p, \mu_q \rangle_{\mathcal{G}} =& \langle \mathbb{E}_p[k(x,\cdot)], \mathbb{E}_q[k(y,\cdot)] \rangle_{\mathcal{F}} \\ =& \mathbb{E}_{p,q}[k(x,y)] \quad \text{where } x \sim p, y \sim q \text{ are independent.}\end{aligned}$$

Plugging this back in (12.3), we obtain

$$d_{\mathcal{G}}(p,q)^2 = \mathbb{E}_{p,p}[k(x,x')] + \mathbb{E}_{q,q}[k(y,y')] - \mathbb{E}_{p,q}[k(x,y')] - \mathbb{E}_{q,p}[k(y,x')] \tag{12.4}$$

where $x, x' \sim p$, $y, y' \sim q$ and all random variables are mutually independent.

Now assume that we have iid samples from two distributions $x_1, x_2, ..., x_n \sim p$ and $y_1, y_2, ..., y_n \sim q$. We cannot calculate the MMD expression given in (12.4), but we can estimate this using the sample mean estimator. That is, we can write the following U-statistic.

$$\widehat{d_{\mathcal{G}}(p,q)^2} = \frac{1}{\binom{n}{2}} \sum_{i<j} k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(y_i, x_j).$$

This is clearly an unbiased estimator of the squared MMD $d^2_{\mathcal{G}}$. It is also consistent, i.e. it converges to $d^2_{\mathcal{G}}$ in probability,

$$\widehat{d_{\mathcal{G}}(p,q)^2} \xrightarrow{p} d_{\mathcal{G}}(p,q)^2.$$

Looking at this value can give us an idea about how close the distributions p and q are.