11 Kernel Methods: Basics

Up until now, we have considered the supervised learning framework where we have data points $(y, x) \in \mathcal{Y} \times \mathcal{X}$ and a loss function $\ell((y, x), f)$. Much of the focus was on generalization properties of the empirical risk minimizers, and different measures of complexity for the function class at hand. In the sequel, we focus on kernel methods which have a lot of connections to previous setup, but we will barely scratch the surface here, so it may seem like quite disconnected at first.

In classical machine learning, it is often the case to consider minimizing some loss function over a mapped feature space $\phi : \mathcal{X} \to \Phi$, with $\ell((y, \phi(x)), f)$. For example, we have been considering linear functions as a popular example $\langle \theta, x \rangle$ where x is the set of features. If we only have a single feature, instead of fitting a 1-dimensional linear regression, we can use a polynomial transformation as a feature map, e.g. $\phi(x) = [1, x, x^2]$, which allows us to fit a degree-3 polynomial by simply using a linear regression. This can be easily generalized to higher dimensions, and there are several reasons such as the ability to represent non-linear dependencies in the data.

One concern is that by increasing the dimension, how much additional computation do we need?

Example. Let's turn to our canonical example, linear regression where we have data points $(y, x) \in \mathcal{Y} \times \mathcal{X}$ and a linear hypothesis class $\mathcal{F} = \{f(\cdot) = \langle \cdot, \theta \rangle,\ \theta \in \mathbb{R}^d\}$, with squared loss $\ell((y, \phi(x)), f) = (y - \langle \phi(x), \theta \rangle)^2$. Notice that there is a feature map $\phi : \mathcal{X} \to \mathbb{R}^d$ applied to the features $x \in \mathcal{X}$.

Using the easily derived closed form solution for the least squares problem, we write

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \theta, \phi(x_i) \rangle)^2 \\ &= (\Phi^\top \Phi)^{-1} \Phi^\top y \quad \text{where} \quad \Phi = \begin{bmatrix} —\phi(x_1)^\top— \\ \vdots \\ —\phi(x_n)^\top— \end{bmatrix}.\end{aligned}$$

Here, Φ takes the place of design matrix X. If we take the SVD of $\Phi = UDV^\top$, we can write $(\Phi^\top \Phi)^{-1} = (VDU^\top UDV^\top)^{-1} = VD^{-2}V^\top \triangleq Q^\top Q$ where $Q = D^{-1}V^\top$. Therefore,

$$\hat{\theta} = Q^\top Q \Phi^\top y.$$

For a new data point x the predicted value from the linear regression model will be

$$\begin{aligned}\hat{y} &= \langle \phi(x), Q^\top Q \Phi^\top y \rangle \\ &= \langle \underbrace{Q\phi(x)}_{\phi'(x)}, Q\Phi^\top y \rangle \quad \text{define} \quad \phi'(x) = Q\phi(x), \\ &= \sum_{i=1}^{n} \langle \phi'(x), \phi'(x_i) \rangle y_i,\end{aligned}$$

which shows that any new predicted value will be a weighted average of the response y_i's with n inner-product operations. The inner product $\langle \phi'(x), \phi'(x_i) \rangle$ should be understood as a similarity metric, i.e., if x is close to the data point x_i, the inner product will be large.

- Even though we possibly increased the dimension of the original features by applying a feature map, we observe that we only need to be able to efficiently compute the inner products $\langle \phi'(x), \phi'(x') \rangle$.

- We only need to know $k(x, x') = \langle \phi'(x), \phi'(x') \rangle$ which is termed as the "kernel" which allows us to efficiently work with high dimensional features (maps).
- There is no unique way of defining a kernel. For instance, if P is another orthogonal matrix, one can use $\psi'' = P\phi'$ which is also a valid kernel.

11.1 Basics of Hilbert Spaces

We will recall some of the basic definitions in this section.

Definition 46 (Hilbert Space)**.** *A Hilbert space $\mathcal{H}$ is a real (or complex) inner product space that is also a complete metric space with respect to the norm induced by its inner product.*

We have been using inner products throughout the lecture; thus, it is useful remind ourselves their formal definition. There are two important characteristics of an Hilbert space, its inner product and completeness. We will define inner products next, but very briefly, completeness of a space means if every Cauchy sequence of points in $\mathcal{H}$ has a limit that is also in $\mathcal{H}$. We will mostly focus on the inner product property.

Definition 47 (Inner product)**.** *An inner product is a function $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ which has the following three properties.*

1. *Symmetry: If $f, g \in \mathcal{H}$, then $\langle f, g \rangle = \langle g, f \rangle$.*
2. *Linearity: If $f, g, h \in \mathcal{H}$ and $a, b \in \mathbb{R}$, then $\langle af + bg, h \rangle = a\langle f, h \rangle + b\langle g, h \rangle$.*
3. *Non-negativity:*
   - *For all $f \in \mathcal{H}$, we have $\langle f, f \rangle \geq 0$.*
   - *Further, $\langle f, f \rangle = 0$ if and only if $f = 0$.*

We finally note that the norm defined by this inner product is: $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle}$.

Example. [Euclidean space] If we have vectors $u, v \in \mathbb{R}^d$, the standard inner product is given as $\langle u, v \rangle = \sum_i u_i v_i$ which defines the Euclidean norm as $\|u\| = \sqrt{\sum_i u_i^2}$.

Example. [Square integrable functions] Let's consider the square integrable functions on $[0, 1]$. That is,

$$L^2([0,1]) = \left\{ f : [0,1] \to \mathbb{R}, \text{ and } \int_0^1 f(x)^2 dx < \infty \right\}$$

with the inner product $\langle f, g \rangle = \int_0^1 f(x)g(x)dx$.

Definition 48 (Dual space)**.** *The dual space $\mathcal{H}^*$ of a Hilbert space $\mathcal{H}$ is the space of all continuous linear functions from the space $\mathcal{H}$ into $\mathbb{R}$. It carries a norm $\|\phi\|_* = \sup_{\|x\|_{\mathcal{H}}=1} |\phi(x)|$.*

This definition will be useful when in the main theorem, but before moving forward, we need to define what a linear function in this context means.

Definition 49 (Linear function)**.** *A function $f : \mathcal{X} \to \mathbb{R}$ is linear if for $x, x' \in \mathcal{X}$ and any $c \in \mathbb{R}$ it satisfies,*

$$f(x + y) = f(x) + f(y) \quad \textit{and} \quad f(cx) = cf(x).$$

It is important to highlight that the linear functions defined as above is different than what we normally refer to in machine learning. That is, $f(x) = ax + b$ is **not linear in** $\mathcal{X}$, but $f(x) = ax$ is linear. This can be easily verified by checking the conditions in the above definition.

Example. [Euclidean space] The dual space of Euclidean space $\mathbb{R}^d$ is given as

$$\mathcal{H}^* = \{\phi : \mathbb{R}^d \to \mathbb{R} \text{ where } \phi \text{ is linear and continuous}\}.$$

Intuitively, since $\phi(x)$ has to be linear and continuous, a linear function in $\mathbb{R}^d$ is given as

$$\phi(x) = \langle u, x \rangle,$$

for some u. Are there any other functions in $\mathcal{H}^*$?

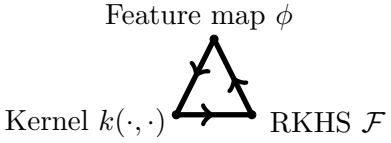
The answer to the above questions is given by the Riesz-Fréchet representation theorem, which is the main building block of what comes next.

Theorem 50 (Riesz-Fréchet representation theorem)**.** *For every element $f \in \mathcal{H}$, there is a unique element $\phi_f \in \mathcal{H}^*$ defined by $\phi_f(g) = \langle f, g \rangle$. Also, for every element $\phi \in \mathcal{H}^*$, there is a unique element $f_\phi \in \mathcal{H}$ such that $\phi(g) = \langle f_\phi, g \rangle$.*

Using this theorem, we can answer the question in the previous example. For every element in the Hilbert space $u \in \mathbb{R}^d$, there is a unique function $\phi \in \mathcal{H}^*$ defined as $\phi(x) = \langle u, x \rangle$. The converse is also true. Therefore the dual space is exactly those functions that can be written as $\phi(x) = \langle x, u \rangle$ where $u \in \mathbb{R}^d$.

11.2 Kernels: formal definitions

In this section, we will formally define kernels. Our objective is to complete the triangular relationship between the feature map ϕ, the kernel k, and the reproducing kernel Hilbert space (RKHS) to be denoted with $\mathcal{F}$ and defined later. We start with the feature map.



Definition 51 (Feature map)**.** *A feature map is a function from the input space $\mathcal{X}$ to a Hilbert space $\mathcal{H}$, i.e.,*

$$\phi : \mathcal{X} \to \mathcal{H}.$$

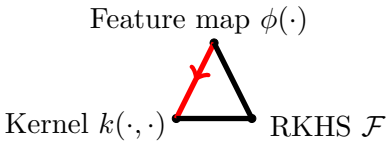
The following notation will be useful. When we define a function f from another function with two arguments $g(x, y)$, e.g. $f(x) = g(x, 3)\ \forall x$, we write this as $f(\cdot) = g(\cdot, 3)$.

Definition 52 (Kernel)**.** *A kernel is a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for any n points $x_1, ..., x_n \in \mathcal{X}$, the matrix defined as $K_{ij} = k(x_i, x_j)$ is positive semidefinite, i.e. $K \succeq 0$.*

Example. Linear kernel $k(x, x') = \langle x, x' \rangle$ is a kernel since for any $x_1, ..., x_n$, the matrix $K_{ij} = k(x_i, x_j)$ can be written as $K = XX^\top$ where X is a matrix with rows $x_i^\top$. It is positive semidefinite (psd) since

$$K \text{ is psd if } \forall u, \quad \langle u, Ku \rangle \geq 0,$$
$$\langle u, Ku \rangle = \langle u, XX^T u \rangle = \langle X^T u, X^T u \rangle = \|X^T u\|^2 \geq 0.$$

We will see more examples of kernels later. The following result connects the feature map to kernel.



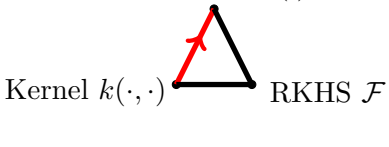
Theorem 53 (Feature map defines a kernel $[\phi(\cdot) \to k(\cdot,\cdot)]$)**.** *A feature map $\phi : \mathcal{X} \to \mathcal{H}$ defines a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.*

Proof. Let $k(x, x') = \langle \phi(x), \phi(x') \rangle$, then $\forall n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$, the kernel matrix is given as $K_{ij} = k(x_i, x_j)$. We show that this matrix is positive semi-definite, $\forall u \in \mathbb{R}^n$,

$$\begin{aligned}
\langle u, Ku \rangle &= \sum_{ij} u_i u_j K_{ij} \\
&= \sum_{ij} u_i u_j \langle \phi(x_i), \phi(x_j) \rangle \\
&= \left\langle \sum_i u_i \phi(x_i), \sum_j u_j \phi(x_j) \right\rangle = \left\| \sum_i u_i \phi(x_i) \right\|_2^2 \geq 0.
\end{aligned}$$

□

The following result connects the kernel to the feature map when the input space $\mathcal{X}$ is finite.



Theorem 54 (Kernel defines a feature map $[k(\cdot,\cdot) \to \phi(\cdot)]$)**.** *For every kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$.*

We prove the above theorem after introducing some key concepts. However, it is quite straightforward to prove it when the input space $\mathcal{X}$ is finite.

Proof. [for finite $\mathcal{X}$] Let $\mathcal{X} = \{x_1, \ldots, x_n\}$, and define the kernel matrix $K_{ij} = k(x_i, x_j)$. Since K is positive semidefinite, its eigen decomposition can be written as $K = UDU^\top \triangleq \Phi\Phi^\top$, therefore $\phi(x_i)^\top = u_i^\top D^{1/2}$ defines a feature map. □

Notice that the choice of feature map is not unique. That is, $\phi'(x) = Q\phi(x)$ also defines a feature map when Q is an orthogonal matrix.

The next section introduces a key concept.

11.3 Hilbert Space defined by the Reproducing Kernel

For the dataset (y_i, x_i) for $i = 1, ..., n$, such that $y_i \in \mathbb{R}$ and the function $\mathcal{F}$ consists of functions that belongs to $f \in L^2([0, 1])$, we consider the canonical ℓ_2-regularized least squares problem

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

where $\mathcal{F}$ is the set of functions that are in consideration. If we choose $\mathcal{F}$ as the entire $f \in L^2([0,1])$, this is too complex and result in overfitting. In this case, the minimizer of the above problem would be simply the function $f(x_i) = y_i$, and $f(x) = 0$ otherwise. This function has $\|f\|_{\mathcal{F}}^2 = \int_0^1 f(x)^2 dx = 0$ and also 0 training loss. The main problem here is that the space covers indicator functions of the form $f(x) = y_i \mathbb{1}_{\{x=x_i\}}$. Clearly, Hilbert spaces are too complex of a search space, so we need some sort of restriction on the space we work with.

Definition 55 (**Lipschitz functional**). *For a Hilbert space $\mathcal{H}$, we say $L : \mathcal{H} \to \mathbb{R}$ is a Lipschitz functional if $\exists M < \infty$,*

$$|L(h) - L(h')| \leq M\|h - h'\|_{\mathcal{H}} \quad \textit{for all} \quad h, h' \in \mathcal{H}.$$

Example. If the Hilbert space is the Euclidean space $\mathbb{R}^d$ with standard inner product, define the functional $L(h) = \langle h, u \rangle$ for some $u \in \mathbb{R}^d$. Then

$$|L(h) - L(h')| = |\langle u, h \rangle - \langle u, h' \rangle| \leq \underbrace{\|u\|}_{M} \|h - h'\|_{\mathcal{H}}.$$

Definition 56 (**Evaluation functional**). *For an Hilbert space $\mathcal{H}$ consisting of functions $h : \mathcal{X} \to \mathbb{R}$, for each $x \in \mathcal{X}$, we define the evaluation functional $L_x : \mathcal{H} \to \mathbb{R}$*

$$L_x(h) = h(x).$$

A little inspection reveals that evaluation functionals are indeed linear! Notice that for $h, h' \in \mathcal{H}$ and $c \in \mathbb{R}$

$$\begin{aligned} L_x(h + h') =& (h + h')(x) = h(x) + h'(x) = L_x(h) + L_x(h'), \\ L_x(ch) =& (ch)(x) = ch(x) = cL_x(h). \end{aligned}$$

Linearity property will be crucial.

Example. Consider the Euclidean input space $\mathcal{X} = \mathbb{R}^d$, and class of linear functions $\mathcal{H} = \{h_\theta(x) = \langle x, \theta \rangle,\ \theta \in \mathbb{R}^d\}$, then $L_x(h_\theta) = \langle x, \theta \rangle$. Linearity can be verified as well.

We are ready to define RKHS.

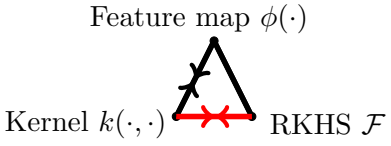
Definition 57 (Reproducing kernel Hilbert space (RKHS)). *An RKHS $\mathcal{F}$ is a Hilbert space over functions $f : \mathcal{X} \to \mathbb{R}$ such that $\forall x \in \mathcal{X}$, the evaluation functionals L_x are Lipschitz continuous.*

The constraint on the evaluation functional of a Hilbert space also restricts the function class. Notice that the reason for overfitting in the ℓ_2-regularized least squares example was the availability of indicator functions which allowed for interpolation. These indicators are not Lipschitz. For example, the problematic indicator function $f(x) = \mathbb{1}_{\{x=1\}} = L_x(f)$ violates the Lipschitz condition, hence doesn't belong to the RKHS.

One key observation was that evaluation functionals L_x are linear. Another one is that they are also continuous, which together imply that they belong to the dual space $\mathcal{F}^*$ (See Definition 48). Therefore, we can apply the second statement in the Riesz-Fréchet representation Theorem 50 and conclude that $\forall f \in \mathcal{F}$, $\exists R_x \in \mathcal{F}$ such that

$$f(x) = L_x(f) = \langle R_x, f \rangle.$$

This tells us that **function evaluations can be written as inner products**. We continue completing the triangular relationship between these key concepts. Next two results completes the following edge in the triangle.



Theorem 58 (Every RKHS defines a unique kernel $[\mathcal{F} \to k(\cdot,\cdot)]$)**.**

Proof.

- By the definition of RKHS, evaluation functionals L_x are Lipschitz (continuous) and linear, so
$$L_x \in \mathcal{F}^*.$$
- By Riesz-Fréchet representation Theorem 50, for L_x, there exists a unique $R_x \in \mathcal{F}$ such that
$$\forall f \in \mathcal{F}, \quad L_x(f) = \langle f, R_x \rangle = f(x). \tag{11.1}$$
The last equality is since L_x is an evaluation functional.
- R_x is called the *representer* and (11.1) is called the *reproducing property*.
- Since $\forall x$, the representer belongs to RKHS $R_x \in \mathcal{F}$, we can use the reproducing property on this functional as well. That is, $\forall x' \in \mathcal{X}$, and for the evaluation functional $L_{x'} \in \mathcal{F}^*$, there exists $R_{x'} \in \mathcal{F}$ such that
$$R_x(x') = L_{x'}(R_x) = \langle R_x, R_{x'} \rangle \triangleq k(x, x'),$$
where k is the kernel. This can be seen by noticing that the representer defines a feature map, i.e., $R_x = \phi(x)$; thus we can invoke Theorem 53, where we showed feature maps define a proper kernel function.

□

Remark. RKHS $\mathcal{F}$ defines a unique kernel $k(\cdot,\cdot)$ which is termed as the *reproducing kernel*. The reason for the name is
$$f(x) = L_x(f) = \langle f, R_x \rangle = \langle f, k(x,\cdot) \rangle,$$
which is where RKHSs get their name. The kernel can be transformed into being representer.

Theorem 59 (Moore-Aronszajn: Every kernel corresponds to a unique RKHS $[k(\cdot,\cdot) \to \mathcal{F}]$)**.** *For every kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a unique RKHS $\mathcal{F}$ with the reproducing kernel k.*

Proof.

- The basic idea is to use the reproducing kernel $k(x,\cdot)$ as a basis for the RKHS $\mathcal{F}$.
- Let $\forall n \in \mathbb{N}$, $\forall x_1, \cdots, x_n \in \mathcal{X}$, $\forall \alpha_1, \cdots, \alpha_n \in \mathbb{R}$, $\forall \theta_1, \cdots, \theta_n \in \mathbb{R}$,
$$f(x) = \sum_i \alpha_i k(x, x_i), \quad \text{and} \quad g(x) = \sum_i \theta_i k(x, x_i).$$

- Let $\mathcal{F}$ be the space composed of functions of the above form. $\mathcal{F}$ is vector space, but not necessarily complete.
- Define the function $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ as

$$\langle f, g \rangle = \sum_{ij} \alpha_i \theta_j k(x_i, x_j).$$

We show it is an inner product. Let $f, g, h \in \mathcal{F}$ and $a \in \mathbb{R}$

  - Symmetry: holds.
  - Linearity: for $\langle af + g, h \rangle = a\langle f, h \rangle + \langle g, h \rangle$.
  - Non-negativity: It is easy to show that $\langle f, f \rangle = \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) = \alpha^\top K \alpha \geq 0$ since k is a kernel. We also need to show $\langle f, f \rangle = 0$ if and only if $f = 0$. Here, $f = 0$ translates to $\alpha = 0$. It is clear that if $f = 0$, then $\langle f, f \rangle = 0$. For the other direction, we define $c(x)^\top = [k(x, x_1), \cdots, k(x, x_n)]^\top$, $\forall x \in \mathcal{X}$. The augmented kernel for a point $x \in \mathcal{X}$ is

$$K' = \begin{bmatrix} K & c(x) \\ c(x)^\top & k(x,x) \end{bmatrix}.$$

    We will prove this by contradiction. Assume that $\langle f, f \rangle = \alpha^\top K \alpha = 0$ but $f \neq 0$ (equivalently $\alpha \neq 0$). For a scalar $b \in \mathbb{R}$, let $u^\top = [\alpha, b]^\top$. Then

$$\begin{aligned} u^\top K' u =& \underbrace{\alpha^\top K \alpha}_{=0} + 2b\alpha^\top c(x) + b^2 k(x,x), \\ =& 2b\alpha^\top c(x) + b^2 k(x,x) \geq 0 \quad \text{since } K' \text{ is psd.} \end{aligned}$$

    But b can be any number, the only way to preserve the inequality for any b is when $\alpha = 0$. To see this, we investigate a function of the form $g(b) = b\xi_1 + b^2\xi_2$. We have $g(0) = 0$ and $g'(0) = \xi_1$. This means that unless $\xi_1 = 0$, the function g is either strictly increasing or decreasing at 0. Thus, one of $g(0 \pm \epsilon)$ for a small ϵ has to be negative.

- We showed that $\mathcal{F}$ is a Hilbert space. To show it is an RKHS, we need to prove that all its evaluation functionals are Lipschitz. We write $\forall f \in \mathcal{F}$

$$\begin{aligned} f(x) =& \sum_i \alpha_i k(x_i, x) \quad \text{by construction of } \mathcal{F} \\ =& \langle f, \underbrace{k(x, \cdot)}_{R_x} \rangle \quad \implies k(x, \cdot) \text{ is indeed the representer.} \end{aligned}$$

  This notation may seem confusing at first. Here, we have

$$k(x, \cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) = \underbrace{1}_{\alpha_1} \cdot k(\underbrace{x}_{x_1}, \cdot), \quad \alpha_i = 0, i > 1.$$

  For an evaluation functional L_x, for $f, g \in \mathcal{F}$, we have

$$\begin{aligned} |L_x(f-g)| = |\langle f - g, R_x \rangle| =& |\langle f - g, k(x, \cdot) \rangle| \\ \leq& \|f - g\|_{\mathcal{F}} \|k(x, \cdot)\|_{\mathcal{F}} \quad \text{by Cauchy-Schwartz} \\ =& \|f - g\|_{\mathcal{F}} \sqrt{k(x,x)} \quad \text{since } \|k(x, \cdot)\|_{\mathcal{F}}^2 = \langle k(x, \cdot), k(x, \cdot) \rangle = k(x,x). \end{aligned}$$

- To complete the proof, one needs to consider the completion of $\mathcal{F}$ by including all the limit points of $\mathcal{F}$. We skip this part.

□

Perhaps, the most important property we derived so far is that a function f in an RKHS $\mathcal{F}$ can be written as a linear combination of kernel evaluations

$$f(x) = \sum_i \alpha_i k(x, x_i) \quad \text{for some } x_i \in \mathcal{X}$$

where k is the unique kernel associated with the RKHS. This will help us reduce complex learning problems to least squares.

It is important to note that the above theorem also proves that if you have a kernel $k(x, x')$, you have a feature map $k(x, \cdot)$.