3 Reproducing Kernel Hilbert Space

I have previously mentioned that a kernel is a positive semi-definite function defined on the product $X \times X$ of a data set X and that this function will replace the Euclidean dot product on a new feature space. Let us now formalize the notion of a kernel.

Definition 3.1 (Kernel)**.** *Let X be a non-empty set. A function $\kappa : X \times X \to \mathbb{R}$ is called a kernel if there exist*

1. *a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ and*
2. *a feature map $\phi : X \to \mathcal{H}$ such that $\forall x, y \in X$*

$$\kappa(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} .$$

REMARK 3.1. *Because inner product is positive semi-definite on the image of X under ϕ, we see that a kernel has to be positive semi-definite. We will see later that the converse statement is also true; that is, any positive semi-definite function on X is a kernel (Theorem 3.1).*

REMARK 3.2. *A kernel may have more than one feature map representation.*

Example 3.1. *Consider $\kappa(x, y) = xy$ defined on $\mathbb{R} \times \mathbb{R}$. If we define $\phi_1 : \mathbb{R} \to \mathbb{R}$ by the identity map and $\phi_2 : \mathbb{R} \to \mathbb{R}^n$ by $\phi_2(x) = (\frac{x}{\sqrt{n}}, \ldots, \frac{x}{\sqrt{n}})^t$, then we see that*

$$\kappa(x, y) = xy = \langle \phi_1(x), \phi_1(y) \rangle_{\mathbb{R}} = \langle \phi_2(x), \phi_2(y) \rangle_{\mathbb{R}^n}$$

with respect to the standard inner product on $\mathbb{R}$ and $\mathbb{R}^n$.

REMARK 3.3. *Given any kernel κ, however, there is one special uniquely defined Hilbert space, called the reproducing kernel Hilbert space (RKHS), with the feature map given by the kernel $\kappa(\cdot, x)$ itself. The RKHS is a function space consisting of certain functions defined on X, and which functions are allowed in the RKHS is determined by the kernel function κ.*

More precisely,

Definition 3.2 (RKHS)**.** *Let X be a set and $\mathbb{R}^X$ the set of all functions from X to $\mathbb{R}$. A Hilbert space $\mathcal{H} \subset \mathbb{R}^X$ is called a Reproducing Kernel Hilbert Space (RKHS) on X over $\mathbb{R}$ if for all $x \in X$, the linear evaluation functional $E_x : \mathcal{H} \to \mathbb{R}$, defined by $f \mapsto f(x)$ for all $f \in \mathcal{H}$, is bounded and thus continuous.*

To see the connection between RKHS and a kernel, we define:

Definition 3.3 (Reproducing Kernel)**.** *Let $\mathcal{H} \subset \mathbb{R}^X$ be a Hilbert space of functionals on a non-empty set X. A function $\kappa : X \times X \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if*

1. *$\forall x \in X, \kappa(\cdot, x) \in \mathcal{H}$*
2. *$\forall x \in X, \forall f \in \mathcal{H}$, the evaluation functional $E_x(f) \equiv f(x) = \langle f, \kappa(\cdot, x) \rangle_{\mathcal{H}}$.*

REMARK 3.4. *From the above definition, a reproducing kernel κ has to satisfy $\kappa(y, x) = \langle \kappa(\cdot, y), \kappa(\cdot, x)\rangle_{\mathcal{H}}$, which implies that it is both symmetric and positive semi-definite. Moreover, a reproducing kernel κ is also a kernel in the sense of Definition 3.1, with the feature map given by $\phi(x) = \kappa(\cdot, x)$.*

REMARK 3.5. *By Cauchy's inequality, we immediately see that a Hilbert space having a reproducing kernel is a RKHS. In fact, we will see in this Chapter that a Hilbert space $\mathcal{H} \subset \mathbb{R}^X$ is a RKHS iff it has a reproducing kernel representing the evaluation functional $E_x, \forall x \in X$.*

This Chapter is thus about understanding the structure and properties of RKHS which takes a central place in modern machine learning. In particular, we will show

1. Every positive semi-definite function on a non-empty set X is a reproducing kernel of some RKHS on X.

2. The reproducing kernel of a RKHS is a kernel in the sense of Definition 3.1.

3. By Remark 3.1, a kernel is positive semi-definite.

In other words, we will show:

Theorem 3.1. *Let X be a non-empty set. The follow conditions on a function $\kappa : X \times X \to \mathbb{R}$ are equivalent:*

1. *κ is positive semi-definite;*

2. *κ is a reproducing kernel of a RKHS on X;*

3. *κ is a kernel.*

3.1 Finite Dimensional RKHS

To develop our intuition for RKHS, we will first study finite dimensional RKHS. Because any finite dimensional real vector space is a Hilbert space, we will postpone a rigorous definition of a Hilbert space; in this section, a Hilbert space will thus mean a finite dimensional inner product space V, which may be a subspace of a larger vector space $\mathbb{R}^m$. As in Chapter 1, we will identify $\mathbb{R}^m$ with the function space $\mathbb{R}^X$, where $X = \{x_1, \ldots, x_m\}$. The columns of a reproducing kernel will then arise as a possibly over-determined ordered spanning set of this subspace $V \subset \mathbb{R}^X$.

In finite dimensions, the evaluation map $E_i(v) \equiv v_i$ is picking out the image of $x_i \in X$ under $v \in \mathbb{R}^X$. We will organize these values of v into a column vector $(v_1, \ldots, v_m)^t$. Then, letting e_i be a column vector with 1 at the i-the location and 0 elsewhere, we can write

$$E_i(v) = e_i^t v\,, \forall v \in \mathbb{R}^X.$$

Given a subspace $V \subset \mathbb{R}^m$ endowed with an inner product, we would like to characterize V by describing how to represent the elements of V, the orientation of V inside $\mathbb{R}^m$, and the inner product on V. We would like this description to be unique and independent of an arbitrary choice of basis.

Definition 3.4 (Reproducing Kernel in Finite Dimensions)**.** *Let $V \subset \mathbb{R}^m$ be a vector space with an inner product $\langle \cdot, \cdot \rangle$. The reproducing kernel of V is an ordered set $K = (\kappa_1 \cdots \kappa_m) \in \mathbb{R}^{m\times m}$ of m vectors $\kappa_i \in \mathbb{R}^m$ such that each κ_i lies in V and satisfies the reproducing property $e_i^t v = \langle v, \kappa_i \rangle, \forall v \in V$.*

REMARK 3.6. *We should think of K as an $m \times m$ matrix.*

To see that a reproducing kernel can be found for a generic finite dimensional inner product space, we need the following lemma.

Lemma 3.1 (Finite Dimensional Riesz Representation Theorem)**.** *Let $(V, \langle \cdot, \cdot \rangle)$ be a finite dimensional real inner product space. Then, any linear functional $\phi \in V^*$ can be uniquely represented as $\langle \cdot, v \rangle$; that is, there exists a unique $v \in V$, such that $\forall w \in V, \phi(w) = \langle w, v \rangle$.*

Proof. Because ϕ and $\langle \cdot, v \rangle$ are both linear, we only need to show that there exists v such that they agree on the basis elements of V. Let $\{e_1, \dots, e_n\}$ be a basis of V. Then, consider the map $\psi : V \to \mathbb{R}^n$ defined by $\psi(v) = (\langle e_1, v \rangle, \langle e_2, v \rangle, \dots, \langle e_n, v \rangle)^t$, which is linear in v. We claim that ψ is injective; since V and $\mathbb{R}^n$ have the same dimension, this claim will also show that ψ is surjective. Suppose $\psi(v_1) = \psi(v_2)$. Then, $\langle e_i, v_1 - v_2 \rangle = 0$ for $i = 1, \dots, n$. Since $v_1 - v_2 \in V$, we can write $v_1 - v_2$ as a linear combination of e_i's; by bilinearity of inner product, $\langle e_i, v_1 - v_2 \rangle = 0$ thus implies that $\langle v_1 - v_2, v_1 - v_2 \rangle = \|v_1 - v_2\|^2 = 0$. The only zero-normed vector is the 0-vector, so $v_1 = v_2$. Thus, ψ is a bijective map, and we can solve for a unique solution v to $\psi(v) = (\phi(e_1), \dots, \phi(e_n))^t$. □

Alternate Proof. Let $\{e_1, \dots, e_n\}$ be an orthonormal basis of V, and define $v = \sum_{i=1}^n \phi(e_i) e_i$. Then, $\langle e_i, v \rangle = \phi(e_i)$ for $i = 1, \dots, n$, implying that $\phi = \langle \cdot, v \rangle$ as linear functionals. The uniqueness of v is proven as above. □

REMARK 3.7. *If V is infinite dimensional, then the statement is true only for continuous linear functionals.*

REMARK 3.8. *The map $\psi : V \to \mathbb{R}^n \cong V^*$ is an isometric isomorphism under the operator norm of linear functionals:*

$$\|\psi(v)\|_{V,1} = \sup_{\|w\|_V = 1} |\psi(v)(w)| = \sup_{\|w\|_V = 1} |\langle w, v \rangle| \leq \sup_{\|w\|_V = 1} \|w\|_V \|v\|_V = \|v\|_V.$$

The inequality is saturated by the unit vector w pointing in the direction of v. Hence,

$$\|\psi(v)\|_{V*} \equiv \|\psi(v)\|_{V,1} = \|v\|_V.$$

Theorem 3.2. *Any finite dimensional inner product space $V \subset \mathbb{R}^m$ has a unique reproducing kernel.*

Proof. Since each evaluation map E_i is a linear functional, Lemma 3.1 implies that it has a unique representer $\kappa_i \in V$. □

To learn how to explicitly construct the reproducing kernel of a finite dimensional inner product space, we will use the following theorem:

Theorem 3.3. *Let $(V, \langle \cdot, \cdot \rangle)$be an n-dimensional inner product space that is a subspace of $\mathbb{R}^m$. The following conditions are equivalent.*

1. *$K \in \mathbb{R}^{m \times m}$ is a reproducing kernel of V.*
2. *$K = \sum_{j=1}^n u_j u_j^t$, where u_j's form an orthonormal basis of V w.r.t. $\langle \cdot, \cdot \rangle$.*
3. *The columns κ_i of K span V and $K_{ij} = \langle \kappa_i, \kappa_j \rangle$.*

Proof. (1⇒2). Suppose $K = (\kappa_i \cdots \kappa_m)$ is a reproducing kernel of $V = \text{span}\{u_1, \ldots, u_n\}$, where u_i are orthonormal with respect to $\langle \cdot, \cdot \rangle$. Then, since $\kappa_i \in V$, we can write $\kappa_i = \sum_{j=1}^n \alpha_{ij} u_j$. The reproducing property of K implies that

$$(u_j)_i = \langle u_j, \kappa_i \rangle = \sum_{k=1}^n \alpha_{ik} \langle u_j, u_k \rangle = \alpha_{ij}.$$

That is, $\kappa_i = \sum_{j=1}^n u_j (u_j)_i = (\sum_{j=1}^n u_j u_j^t)_{:,i}$.
(2⇒3) Since each column of K is a linear combination of the basis elements u_j of V, the column span of K lies inside V. Since u_i are linearly independent, we can find n vectors w_j such that $w_i^t u_j = \delta_{ij}$. Thus, $K w_j = u_j$; i.e. all the basis elements of V are in the column span of K, and V thus lies inside the column span of K. Hence, V has to equal to the column span of K. Furthermore,

$$\langle \kappa_i, \kappa_j \rangle = \sum_{k=1}^n \sum_{\ell=1}^n \langle u_k (u_k)_i, u_\ell (u_\ell)_j \rangle = \sum_{k=1}^n (u_k)_i (u_k)_j = K_{ij}.$$

(3⇒1) Since the columns κ_i of K span V, any $v \in V$ can be written as $v = \sum_{i=1}^m \alpha_i \kappa_i$. Hence,

$$\langle v, \kappa_j \rangle = \sum_{i=1}^m \alpha_i \langle \kappa_i, \kappa_j \rangle = \sum_{i=1}^m \alpha_i \langle \kappa_i, \kappa_j \rangle = \sum_{i=1}^m \alpha_i (\kappa_i)_j = \left(\sum_{i=1}^m \alpha_i \kappa_i \right)_j = v_j.$$

That is, κ_j have the reproducing property $e_j^t v = \langle v, \kappa_j \rangle$ for $j = 1, \ldots, m$. □

REMARK 3.9. *The reproducing kernel of V thus encodes both a spanning set of V and the Gram matrix of these vectors.*

REMARK 3.10. *Defining a feature map $i \mapsto k_i$, the third condition of Theorem 3.3 implies that a reproducing kernel is indeed a kernel in the sense of Definition 3.1.*

REMARK 3.11. *Note that the expression $K = \sum_{j=1}^n u_j u_j^t$ is invariant under the orthogonal transformation $(u_1 u_2 \cdots u_n) \to (u_1 u_2 \cdots u_n) S, S \in O(n)$, and is thus independent of the specific choice of an orthonormal basis of V.*

REMARK 3.12. *Kernel captures both intrinsic and extrinsic geometry of RKHS embedded in a larger function space.*

EXERCISE 3.1. *Let $V = \mathbb{R}^m$ with an inner product $\langle v, w\rangle = v^t Q w$, where Q is a symmetric positive definite matrix. Then, show that $K = Q^{-1}$. Note that in this case, the orthonormal vectors u_j are conjugate w.r.t. to Q.*

Given a finite data set X of m elements, we have so far proved that any finite dimensional inner product space $V \subset \mathbb{R}^X \simeq \mathbb{R}^m$ has a unique $m \times m$ reproducing kernel K. We will now prove the converse; i.e. any kernel $K : X \times X \to \mathbb{R}$ uniquely defines a subspace $V \subset \mathbb{R}^X \simeq \mathbb{R}^m$ whose reproducing kernel is K. For this purpose, we will view K as an $m \times m$ positive semi-definite matrix. We need the following lemma to prove this statement.

Lemma 3.2. *Let $K \in \mathbb{R}^{m\times m}$ be a positive semi-definite matrix. Then, any vector $\alpha \in \mathbb{R}^m$ satisfying $\alpha^t K \alpha = 0$ satisfies $K\alpha = 0$.*

Proof. Because K is symmetric, it has m orthonormal eigenvectors v_i with non-negative eigenvalues λ_i, $i = 1, \dots, m$. Thus, any $\alpha \in \mathbb{R}^m$ can be written as $\alpha = \sum_{i=1}^m a_i v_i$, and we have

$$\alpha^t K \alpha = \sum_{i=1}^m \lambda_i a_i^2 = 0.$$

Since $\lambda_i \geq 0$, the last equality implies that $a_i = 0$ for any non-zero eigenvalue λ_i. Thus, α must be a linear combination of null eigenvectors only. □

We can now prove our claim.

Theorem 3.4. *Let $K \in \mathbb{R}^{m\times m}$ be a positive semi-definite matrix. Then, there exists a unique inner product subspace $V \subseteq \mathbb{R}^m$ whose reproducing kernel is K.*

Proof. Let us first prove the existence. First, we define V to be the column span of K. Then, any $v, w \in V$ can be written as $v = K\alpha$ and $w = K\beta$. We define the inner product on V as $\langle v, w\rangle \equiv \alpha^t K \beta$. Because the columns of K may be linearly dependent, we need to check that this definition of inner product does not depend on the representation of v and w as linear combinations of the columns of K. Suppose $v = K\alpha'$ and $w = K\beta'$. Then,

$$\begin{aligned} \alpha'^t K \beta' &= (\alpha' - \alpha)^t K \beta' + \alpha^t K \beta' \\ &= (\alpha' - \alpha)^t K \beta' + \alpha^t K(\beta' - \beta) + \alpha^t K \beta. \end{aligned}$$

But, $v = K\alpha = K\alpha'$ and $w = K\beta = K\beta'$ imply that $(\alpha' - \alpha)^t K = 0$ and $K(\beta' - \beta) = 0$. Hence, $\alpha'^t K \beta' = \alpha^t K \beta$. We next need to show that this inner product is positive definite on V. Since K is assumed to be positive semi-definite, any $v = K\alpha$ in V must satisfy

$$\langle v, v\rangle = \alpha^t K \alpha \geq 0.$$

Suppose $v = K\alpha$ satisfies $\langle v, v\rangle \equiv \alpha^t K \alpha = 0$; then, by Lemma 3.2, we must have $v = K\alpha = 0$. Hence, our definition defines a true inner product on V. Choosing $v = Ke_i$ and $w = Ke_j$, we have $\langle k_i, k_j\rangle = K_{ij}$. Thus, K is the reproducing kernel of V.

To see that our V above is unique, suppose V_2 is another inner product subspace in $\mathbb{R}^m$ whose reproducing kernel is K. Then, Theorem 3.3 implies that V_2 is the column span of K and thus equal to V as a vector space. Again, by Theorem 3.3, $\langle k_i, k_j\rangle_{V_2} = K_{ij} \equiv \langle k_i, k_j\rangle_V$.

Hence, since the inner products agree on the spanning sets, V_2 and V also have the same inner product. □

We have thus proved Theorem 3.1 in finite dimensions:

$$\begin{tikzcd}
 & \text{K is positive semi-definite} \arrow[dr, leftrightarrow] & \\
\text{K is a reproducing kernel} \arrow[ur, leftrightarrow] \arrow[rr, leftrightarrow] & & \text{K is a kernel}
\end{tikzcd}$$

We have also shown that:

There is a bijective correspondence between $m \times m$ positive semi-demifite matrices K and subspaces in $\mathbb{R}^m$ with K as the reproducing kernel.

3.1.1 Intrinsic vs. Extrinsic Coordinates

Given a finite data set $X = \{x_1, \ldots, x_m\}$, a vector $v \in V \subset \mathbb{R}^X$ can be represented by its external coordinates $(v_1, v_2, \ldots, v_m)$, which are the values that the function v takes on the points $x_1, x_2, \ldots, x_m$. If V is a proper subset of $\mathbb{R}^m$, then arbitrarily changing the coordinates will make v lie outside V. In this sense, the extrinsic coordinate system may not be good for parametrizing V. By contrast, viewing V as a RKHS, we have seen that any vector $v \in V$ can be written as $v = K\alpha$, where K is the reproducing kernel of V; importantly, any choice of $\alpha \in \mathbb{R}^m$ will leave v inside V, so the columns of K provide a redundant spanning set for constructing an intrisic coordinate system on V. The extrinsic and intrinsic coordinate systems are related by

$$v_i = \langle v, k_i \rangle = \sum_{j=1}^{m} \alpha_j \langle k_j, k_i \rangle = (K\alpha)_i.$$

3.1.2 Feature Map from a Reproducing Kernel

Let us recapitulate our initial motivation and our findings. We are given a data set X, and we have some notion of similarity $\kappa(x_i, x_j)$ between data points. This similarity measure generalizes the dot product in a linear space and may be a highly non-linear function of the data points; however, we would like our linear analysis techniques to be useful, so we search for a new data representation $\phi : X \to \mathcal{H}$, where $\mathcal{H}$ is an inner product space, such that $\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$. Even though there may be many such representations, there is a unique representation if we demand that $\mathcal{H}$ be a RKHS $\mathcal{H}_K$ on X with $K = (\kappa(x_i, x_j))$ as its reproducing kernel (we have proved this statement for X a finite set, but the same holds when X is not finite and $\mathcal{H}_K$ is an infinite dimensional Hilbert space.). By the reproducing property of K on $\mathcal{H}_K$, we have

$$\kappa(x_i, x_j) = \langle \kappa(\cdot, x_i), \kappa(\cdot, x_j) \rangle_{\mathcal{H}_K},$$

meaning that our desired feature map is $\phi(x_i) \equiv \kappa(\cdot, x_i)$.

When $X = \{x_1, \ldots, x_m\}$ is a finite set, we have shown that $\phi(x_i) \equiv \kappa_i$, the i-th column of the kernel matrix K. The transformed data points will lie in $V \subset \mathbb{R}^m$ which is spanned by the columns of K and is a RKHS on X with K as its reproducing kernel. The inner product on V is defined through

$$\langle v, w \rangle = \alpha^t K \beta,$$

where $v = K\alpha$ and $w = K\beta$.

3.1.3 Interpolation Problem

Can we reconstruct an unknown function if we are given its value at a finite set of points? This seemingly impossible interpolation or imputation task becomes feasible if we restrict that the search space is a RKHS and that the ideal solution has the smallest norm among all possible solutions. We will study this problem in finite dimensions here, but the formalism easily carries over to infinite dimensions.

Let $(V, \langle\cdot,\cdot\rangle)$ be a subspace of $\mathbb{R}^m$ endowed with an inner product. Suppose we are given s extrinsic coordinates $v_{i_1}, v_{i_2}, \ldots, v_{i_s}$ of a vector $v \in V$. That is, we are given the values $v_{i_\ell} = E_{x_{i_\ell}}(v) \equiv v(x_{i_\ell})$, $\ell = 1, \ldots, s$, of the evaluation maps, viewing v as a function defined on the data set $X = \{x_1, \ldots, x_m\}$. Note that any vector $w \in V$ that lies in the joint null space $\mathcal{N} \subset V$ of $E_{x_{i_\ell}}$, $\ell = 1, \ldots, s$, can be added to a particular solution v without changing the specified values; i.e., $E_{x_{i_\ell}}(v + w) = E_{x_{i_\ell}}(v)$. Hence, in general, the solution space will be an affine subspace in V, and there does not exist a unique solution to the interpolation problem. To find the best solution that agrees with our intuition of parsimony or simplicity being desirable, we will thus impose that our solution has the smallest norm among all possible solutions.

This constraint on the norm can be understood as follows: suppose our tentative solution v is not orthogonal to some $w \in \mathcal{N} \subset V$; then, for any $t \in \mathbb{R}$, we have

$$\|v - tw\|^2 = \langle v - tw, v - tw\rangle = \|v\|^2 - 2t\langle v, w\rangle + t^2\|w\|^2.$$

Hence, we can always choose a sufficiently small t such that the t^2 term is dominated by the term linear in t, and we can choose the sign of t such that the linear term is negative. Thus, we can find another solution $v - tw$ that is smaller than v. As a result, the minimum norm constraint implies $v \in \mathcal{N}^\perp$.

Because any finite dimensional inner product space is a RKHS, we can rephrase the null space conditions $E_{x_{i_\ell}}(w) = 0$ as

$$\forall w \in \mathcal{N}, \langle w, k_{i_\ell}\rangle = 0,\ \ \ell = 1, \ldots, s,$$

where k_{i_ℓ} is the i_ℓ-th column of the reproducing kernel matrix of $(V, \langle\cdot,\cdot\rangle)$. Let $\mathcal{K} \subset V$ denote the subspace spanned by k_{i_ℓ}, $\ell = 1, \ldots, s$. The null space condition thus implies that

$$\mathcal{N} = \mathcal{K}^\perp.$$

But, all finite dimensional spaces are closed, so we have $\mathcal{N}^\perp = (\mathcal{K}^\perp)^\perp = \mathcal{K}$. Thus, the minimum norm solution v lies in $\mathcal{K}$; i.e.

$$v = \sum_{p=1}^{s} \alpha_p k_{i_p}.$$

To find the coefficients α_p, we use the constraints $E_{x_{i_\ell}}(v) = v_{i_\ell}$:

$$v_{i_\ell} = \sum_{p=1}^{s} \alpha_p E_{x_{i_\ell}}(k_{i_p}) = \sum_{p=1}^{s} K_{i_\ell i_p}\, \alpha_p, \ \ \ell = 1, \dots, s.$$

We thus need to solve the linear equation

$$\boxed{u = L\alpha}, \tag{3.1}$$

where $u = (v_{i_1}, v_{i_2}, \dots, v_{i_s})^t$ is the vector of known values of v, and $L = (K_{i_\ell i_p})_{\ell,p=1}^{s}$ is the submatrix of the kernel matrix K corresponding to the given samples $x_{i_1}, \dots, x_{i_s}$.

REMARK 3.13. *In principle, the submatrix L will be invertible if K is strictly positive definite. Even is L is invertible, however, it might be ill conditioned, and numerical calculation of the inverse might be unstable. We should thus use the Moore-Penrose pseudo-inverse instead.*

REMARK 3.14. *In case the observed values were contaminated by noise, the amount of data needed to interpolate the data without overfitting would increase. One might also try penalized regression to estimate α.*

REMARK 3.15. *The same equation 3.1 holds when V is an infinite dimensional RKHS. In this case, the full interpolated function is*

$$v = \sum_{p=1}^{s} \alpha_p K(\cdot, x_{i_p}),$$

$u = (v(x_{i_1}), v(x_{i_2}), \dots, v(x_{i_s}))^t$ *and* $L = (K(x_{i_\ell}, x_{i_p}))_{\ell,p=1}^{s}$.

3.2 Infinite Dimensional RKHS

3.2.1 Hilbert Space

A quintessential assumption in quantum mechanics is that any physical state can be expressed as a convergent series in the orthonormal eigenfunctions of the Hamiltonian:

$$\psi(x) = \sum_{k=1} \alpha_k \psi_k(x). \tag{3.2}$$

Conversely, any such valid expansion should describe a physical state. Establishing the mathematical formalism that allowed us to understand the meaning of (3.2) and what it means for an expansion to be valid was one of the great accomplishments of physics in the early 20th century. This success is largely attributable to John von Neumann who introduced the theory of Hilbert space, thereby unifying Heisenbeg's matrix formulation of QM with Schroedinger's wave function formulation into a single mathematical framework.

To define "valid" expansions, consider the sequence $\{f_n\}_{n=1}^{\infty}$, where $f_n = \sum_{k=1}^{n} \alpha_k \psi_k$. In QM, a valid expansion corresponds to the case where the partial sums become closer and closer to each other for sufficiently large n's. This concept is formalized by defining the notion of a Cauchy sequence, which characterizes a sequence that should converge:

Definition 3.5 (Cauchy Sequence)**.** *Let (S, d) be a metric space. A sequence $\{f_n\}_{n=1}^{\infty}$ of elements $f_n \in S$ is called Cauchy if $\forall \epsilon > 0$, $\exists N(\epsilon) > 0$ such that for any $n, m \geq N(\epsilon)$, $d(f_n, f_m) < \epsilon$.*

REMARK 3.16. *Any convergent sequence is clearly Cauchy, but the converse may not be true if the set is missing some limit points.*

Example 3.2. *Consider the set $\mathbb{Q}$ of all rational numbers. The sequence of rational numbers $f_1 = 3, f_2 = 3.1, f_3 = 3.14, f_4 = 3.141, f_5 = 3.1415, \ldots$ that represent truncated decimal expansions of π is Cauchy, but converges to $\pi \notin \mathbb{Q}$.*

Definition 3.6 (Complete Metric Space)**.** *A metric space is said to be complete if any Cauchy sequence in the space converge to a point in the space.*

Example 3.3. *The set $\mathbb{R}$ of all real numbers is complete, but $\mathbb{Q}$ is not.*

REMARK 3.17. *Given a metric space there is a unique completion of the space by adding the limit points of all Cauchy sequences in the space. See "Some notes on completions" by John Loftin.*

EXERCISE 3.2. *Let $\ell^2_{(0)}$ be the set of all sequences $\{f_n\}_{n=1}^{\infty}$, $f_n \in \mathbb{R}$, where only a finite number of f_n's are non-zero. Show that $\ell^2_{(0)}$, equipped with the inner product*

$$\langle \{f_n\}, \{g_n\} \rangle = \sum_{n=1}^{\infty} f_n g_n \,,$$

is not complete.

Definition 3.7 (Banach Space)**.** *A complete normed vector space is called a Banach space.*

Definition 3.8 (Hilbert Space)**.** *A complete inner product space is called a Hilbert Space.*

Example 3.4 (ℓ^2 Hilbert Space)**.** *The set ℓ^2 of all sequences $\{f_n\}_{n=1}^{\infty}$, $f_n \in \mathbb{R}$, such that $\|\{f_n\}\|_2 \equiv \sqrt{\sum_{n=1}^{\infty} f_n^2} < \infty$ is a Hilbert space with the inner product*

$$\langle \{f_n\}, \{g_n\} \rangle = \sum_{n=1}^{\infty} f_n g_n \,.$$

This space is obtained by adding all limit points to the $\ell^2_{(0)}$ in Exercise 3.2.

Example 3.5 (ℓ^p Banach Space)**.** *The set ℓ^p, $p \geq 1$, of all sequences $\{f_n\}_{n=1}^{\infty}$, $f_n \in \mathbb{R}$, such that $\|\{f_n\}\|_p \equiv (\sum_{n=1}^{\infty} |f_n|^p)^{1/p} < \infty$ is a Banach space. As in the finite dimensional case, it is not a Hilbert space if $p \neq 2$.*

REMARK 3.18. *In infinite dimensions, the ℓ_p and $\ell_{p'}$ norms are inequivalent when $p \neq p'$. For example, the Harmonic sequence*

$$x = \left\{1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, \dots\right\}$$

is not bounded in the ℓ_1-norm, but $\|x\|_p = \zeta(p)^{1/p} < \infty$, for $p > 1$, where ζ is the Riemann ζ-function.

Example 3.6 (L^2 Hilbert Space)**.** *The set $L^2([a,b])$ of all square integrable functions on $[a,b] \subset \mathbb{R}$, i.e. the functions satisfying*

$$\int_a^b |f(x)|^2 \, dx < \infty \, ,$$

is a Hilbert space with the inner product

$$\langle f, g \rangle = \int_a^b f(x) g(x) \, dx \, .$$

Quantum wave functions are elements of this space.

3.2.2 Separable Hilbert Spaces

Definition 3.9 (Separable Hilbert Space)**.** *A Hilbert space $\mathcal{H}$ is called separable if it has a countable orthonormal basis.*

REMARK 3.19. *This orthonormal basis $\{e_n\}_{n=1}^{\infty}$ is not a Hamel basis of an infinite dimensional Hilbert space $\mathcal{H}$, but it allows us to write any vector $f \in \mathcal{H}$ as*

$$f = \sum_{n=1}^{\infty} \alpha_n e_n \, , \alpha_n \in \mathbb{R},$$

where the equality is understood as a limit of partial sums. In other words, the span of $\{e_n\}_{n=1}^{\infty}$ is dense in $\mathcal{H}$.

REMARK 3.20. *In QM, an orthonormal basis typically consists of the eigenvectors of a self-adjoint Hamiltonian; or, to be more precise, eigenvectors of a related compact self-adjoint integral operator.*

Theorem 3.5 (Separable Hilbert Space)**.** *Every separable real Hilbert space is isometrically isomorphic to either $\mathbb{R}^n$, for some $n \geq 0$, or to ℓ^2.*

REMARK 3.21. *In general, two Hilbert spaces having orthonormal bases with the same cardinality are isomorphic.*

This theorem tells us that $L^2[a,b]$ and ℓ^2 can be identified as Hilbert spaces; you can think of ℓ^2 as consisting of the Fourier coefficients in the expansion of a square integrable function in terms of an orthonormal basis. It is this theorem that helped unify the approaches of Heisenberg's matrix mechanics on ℓ^2 and Schrödinger's wave mechanics on $L^2[a,b]$. Most Hilbert spaces that you will encounter in your lives will be separable, so we will assume the case in our discussion.

3.2.3 Infinite Dimensional RKHS

We have proved that any linear map defined on a finite dimensional vector space is bounded and, thus, continuous. In particular, all linear functionals on a finite dimensional vector space are bounded. When the domain is not finite dimensional, a general linear map is not bounded. So, when defining the dual space of a Hilbert space, we need to specify whether we impose the condition of continuity or not.

Definition 3.10 (Topological Dual of a Hilbert Space)**.** *Let $\mathcal{H}$ be a Hilbert space. The topological dual $\mathcal{H}^*$ of $\mathcal{H}$ is the vector space of all continuous linear functionals on $\mathcal{H}$.*

Definition 3.11 (Algebraic Dual of a Hilbert Space)**.** *Let $\mathcal{H}$ be a Hilbert space. The algebraic dual $\mathcal{H}'$ of $\mathcal{H}$ is the vector space of all linear functionals on $\mathcal{H}$.*

Clearly, $\mathcal{H}^* \subseteq \mathcal{H}'$. If $\mathcal{H}$ is finite dimensional, then $\mathcal{H}^* \cong \mathcal{H}$ via the finite dimensional Riesz Representation Theorem (Lemma 3.1), and $\mathcal{H}^* = \mathcal{H}'$ because any linear map is bounded on a finite dimensional vector space. If $\mathcal{H}$ is infinite dimensional, however, there always exists an unbounded linear functional, and $\mathcal{H}^*$ is strictly smaller than $\mathcal{H}'$. For example, pick a countably infinite set $\{e_n\}_{n=1}^{\infty}$ of linearly independent unit vectors in $\mathcal{H}$ and define $\phi(e_n) = n$. Defining the action of ϕ on the remaining elements of a Hamel basis containing $\{e_n\}_{n=1}^{\infty}$ to be 0 and extending the action of ϕ to entire $\mathcal{H}$ by linearity, we see that ϕ is an unbounded linear functional.

For a generic infinite dimensional inner product space V, we still have $V \subseteq V^*$ via the map $V \ni v \mapsto \langle \cdot, v \rangle \in V^*$, since the inner product is bounded by the Cauchy inequality. If V is also complete, then V is a Hilbert space, and we again have $V \cong V^*$. Let us prove this statement for separable Hilbert spaces:

Lemma 3.3 (Riesz Representation Theorem)**.** *Let $\mathcal{H}$ be a separable Hilbert space. Then, any $\phi \in \mathcal{H}^*$ can be represented by a unique $v \in \mathcal{H}$; i.e. $\forall f \in \mathcal{H}, \phi(f) = \langle f, v \rangle$. Thus, $\mathcal{H}^* \cong \mathcal{H}$.*

Proof. The uniqueness can be proved just as in the finite dimensional case (see the proof of Lemma 3.1). To see the existence, let $\{e_i\}_{i=1}^{\infty}$ be an orthonormal basis of $\mathcal{H}$, and consider

$$v = \sum_{i=1}^{\infty} \phi(e_i) e_i.$$

It is easily seen that $\phi(e_n) = \langle e_n, v \rangle$. So, by linearity, we see that $\phi(f) = \langle f, v \rangle$ for any $f \in \mathcal{H}$. We thus only need to show that $v \in \mathcal{H}$, i.e. $\|v\|^2 = \langle v, v \rangle_{\mathcal{H}} < \infty$. For this purpose, define $f_N = \sum_{i=1}^{N} \phi(e_i)e_i$. Then, since ϕ is bounded,

$$|\phi(f_N)| = \left|\sum_{i=1}^{N} \phi(e_i)^2\right| \leq \|\phi\|\|f_N\|_{\mathcal{H}} = \|\phi\|\sqrt{\sum_{i=1}^{N} \phi(e_i)^2} < \infty.$$

Hence, for all $N > 0$, $\|f_N\|_{\mathcal{H}} \leq \|\phi\|$. Since $\|f_N\|_{\mathcal{H}}$ is a non-decreasing function of N and each $\|f_N\|_{\mathcal{H}}$ is bounded by $\|\phi\|$, we see that $\lim_{N\to\infty} \|f_N\|_{\mathcal{H}}$ exists and is finite. □

REMARK 3.22. *It is important to note that the proof of the theorem requires that the inner product space is complete and that ϕ is bounded.*

REMARK 3.23. *The Riesz Representation Theorem can be extended to the case where $\mathcal{H}$ is a non-separable Hilbert space. In this case, note that the kernel $\ker(\phi)$ of $\phi \in \mathcal{H}^*$ is closed, because ϕ is continuous and $\ker(\phi)$ is its inverse image of the closed set $\{0\}$. This property of $\ker(\phi)$ allows us to decompose $\mathcal{H} = \ker(\phi) \oplus (\ker(\phi))^\perp$. If $\phi = 0$, then its representer is just the zero vector $v = 0$. If $\phi \neq 0$, then $(\ker(\phi))^\perp$ is one-dimensional; let z be any non-zero element in $(\ker(\phi))^\perp$; the unique representer of ϕ is $v = \phi(z)z/\|z\|^2$, as the action of $\langle \cdot, v\rangle$ on the direct sum $\ker(\phi) \oplus (\ker(\phi))^\perp$ is exactly the same as that of ϕ.*

We can now define a RKHS of arbitrary dimension. The definition of an infinite dimensional RKHS supersedes the definition in finite dimensions:

Definition 3.12. *An inner product space $\mathcal{H}$ is called a RKHS*

1. *if it is a Hilbert space of real functions on some non-empty set X (i.e $\mathcal{H} \subset \mathbb{R}^X$ and it is a Hilbert space); and,*
2. *for all $x \in X$, the evaluation linear functional $E_x : \mathcal{H} \to \mathbb{R}$ at x, defined by $f \mapsto f(x)$, is bounded and, thus, continuous (i.e. $E_x \in \mathcal{H}^*$).*

Definition 3.13 (Reproducing Kernel of a Hilbert Space)**.** *A function $K : X \times X \to \mathbb{R}$ is called a reproducing kernel of a Hilbert space $\mathcal{H} \subset \mathbb{R}^X$ on X if*

1. *$\forall x \in X$, $K(\cdot, x) \in \mathcal{H}$, and*
2. *$\forall x \in X, \forall f \in \mathcal{H}$, $E_x(f) \equiv f(x) = \langle f, K(\cdot, x)\rangle_{\mathcal{H}}$.*

Theorem 3.6. *Every RKHS on X has a unique reproducing kernel.*

Proof. The theorem follows from the Riesz Representation Theorem (Lemma 3.3), and the kernel $K(\cdot, x)$ evaluated at $x \in X$ is the unique representer of the evaluation map E_x at x. □

REMARK 3.24. *Choose $f = K(\cdot, y)$ in the above definition, we see that $K(x, y) = \langle K(\cdot, y), K(\cdot, x)\rangle_{\mathcal{H}}$, which implies that K is symmetric and positive semi-definite.*

An infinite dimensional version of Theorem 3.4 also holds:

Theorem 3.7 (Moore-Aronszajn)**.** *Let $X \subset \mathbb{R}^n$, and let $K : X \times X \to \mathbb{R}$ be finitely positive semi-definite; i.e. given any finite set $\{x_1, \dots, x_m\} \subset X$, $(K(x_i, x_j))$ is a positive semi-definite matrix. Then, there exists a unique RKHS $\mathcal{H} \subset \mathbb{R}^X$ with reproducing kernel K.*

Sketch of Proof. The proof is similar to that of Theorem 3.4. The main idea is that we will parametrize a subspace of $\mathbb{R}^X$ using the intrinsic coordinate system arising from taking linear combinations of $K(\cdot, x)$, endow this space with an inner product, and then complete this space into a Hilbert space.

Let $\mathcal{H}_0 = \operatorname{span}\{k(\cdot, x)\}_{x \in X}$. Any $f \in \mathcal{H}_0$ can then be written as

$$f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i)$$

for some set of points $\{x_1, \dots, x_n\}$. This step is in complete analogy with representing $v = K\alpha$ in the finite dimensional case. For any $f, g \in \mathcal{H}_0$, where $f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i)$ and $f = \sum_{j=1}^{m} \beta_j K(\cdot, x_j)$, define their inner product as

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i K(x_i, x_j) \beta_j.$$

We can check that this inner product is well defined on $\mathcal{H}_0$. Taking the set of all equivalence classes of Cauchy sequences on $\mathcal{H}_0$ then yields the unique RKHS. □

3.2.4 L^2 Space is not a RKHS

The $L^2[a, b]$ space of square integrable "functions" on $[a, b] \subset \mathbb{R}$ is not a RKHS. Let us see why this is the case using a concrete example. Let $[a, b] = [-1, 1]$. On this interval, the evaluation map E_x of a function $f \in L^2([-1, 1]$ is usually written in physics as $E_x(f) = \int_a^b f(y)\delta(x - y)dy$, where $\delta(x - y)$ is the Dirac delta function. It may thus appear that we can write $E_x(f) = \langle f, \delta(x - \cdot) \rangle_{L^2}$ and that the Riesz representation theorem still holds. However, one important point to notice is that the delta function is not square integrable.

Example 3.7. *For $n \in \mathbb{N}$, consider the function*

$$f_n(x) = \begin{cases} n - n^2|x|, & -1/n \le x \le 1/n \\ 0, & \text{otherwise} \end{cases}$$

on the interval $[-1, 1]$. In the limit $n \to \infty$, this function behaves likes the Dirac delta function centered at $x = 0$, but its squared norm

$$\int_{-1}^{1} |f_n(x)|^2 \, dx = \frac{2n}{3}$$

diverges as $n \to \infty$. Thus, the delta function is not in $L^2[-1, 1]$.

You might think that it's contradictory that f_n behaves likes a Dirac delta function for large n, but its "limit" is not in $L^2[-1,1]$, apparently violating the fact the $L^2[-1,1]$ is complete. This paradox is resolved by noticing that the sequence is actually not Cauchy:

EXERCISE 3.3. *Show that the sequence $\{f_n\}$ in Example 3.7 is not a Cauchy sequence in $L^2[-1,1]$.*

Example 3.8. *For $n \in \mathbb{N}$ and $0 < \alpha < \frac{1}{2}$, consider the function*

$$f_n(x) = \begin{cases} n^\alpha - n^{1+\alpha}|x|, & -1/n \le x \le 1/n \\ 0, & \textit{otherwise} \end{cases}$$

on the interval $[-1,1]$. Its ℓ_2-norm is

$$\|f_n\|_2^2 = \frac{2}{3n^{1-2\alpha}}.$$

The evaluation map at $x = 0$, $E_0(f_n) = n^\alpha = \sqrt{\frac{3n}{2}}\|f_n\|_2$, cannot satisfy

$$|E_0(f_n)| \le C\|f_n\|_2$$

for any constant C; hence, it is not bounded.

Strangely, $f_n \to 0$ in ℓ_2-norm, and the sequence is Cauchy. But, the fact that $E_0(0) = 0$ seems to contradict the fact that $E_0(f_n)$ diverges as $n \to \infty$. This paradox can be resolved by noting that the elements of $L^2[-1,1]$ is not a function, but an equivalence class of functions that may differ on a set of measure 0. Thus, the evaluation map is ill-defined on any L^2 space, further highlighting the difference between L^2 and RKHS.

3.2.5 ℓ^2 Space is a RKHS

Since $\ell^2 \subset \mathbb{R}^{\mathbb{N}}$, we can define an evaluation map E_i, for $i \in \mathbb{N}$, as

$$E_i(f) = f_i$$

for any $f = \{f_1, f_2, \ldots\} \in \ell^2$. Since $\sum_{i=1}^\infty |f_i|^2 < \infty$, we have

$$|E_i(f)| = f_i \le \sqrt{\sum_{i=1}^\infty |f_i|^2} = \|f\|_2 < \infty,$$

which implies that E_i is bounded. Its inner product representer is $e_i = (0, 0, \ldots, 0, 1, 0, 0, \ldots)$, having 1 at the i-th location and 0 elsewhere:

$$E_i(f) = \langle f, e_i \rangle_2 .$$

REMARK 3.25. *The fact that $L^2[a,b]$ is not a RKHS, while ℓ^2 is, may appear to contradict Theorem 3.5, which states that the two Hilbert spaces are isometrically isomorphic. This paradox can be resolved by noting that the reproducing kernel property imposes an extra structure on a Hilbert space that may not be respected by the isomorphism map. For one* *thing, the definition of RKHS assumes that the elements of RKHS are functions from some non-empty set X to $\mathbb{R}$ and that for all $x \in X$, there exists a continuous evaluation functional E_x. This underlying set X is not preserved under the isomorphism map identifying $L^2[a,b]$ with ℓ^2. Furthermore, we have already seen in Example 3.8 that the elements of $L^2[a,b]$ are not functions, but an equivalence class of functions, so an evaluation map is not even well defined on $L^2[a,b]$.*

3.2.6 The Momentum Operator in QM is Unbounded

Let $L^2([0,1])$ denote the space of all square integrable functions on the interval $[0,1]$. In units where $\hbar = 1$, the momentum operator is

$$p = -i\frac{\partial}{\partial x}.$$

Not all functions in $L^2([0,1])$ are differentiable, so p is defined on the subspace $C^1([0,1]) \subset L^2([0,1])$ of functions with a continuous first derivative. In particular, $x^n \in C^1([0,1])$ for $n > 1/2$ and $\|x^n\|_2 = \frac{1}{\sqrt{2n+1}}$. But,

$$\|p(x^n)\|_2 = \|nx^{n-1}\| = \frac{n}{\sqrt{2n-1}}.$$

Hence,

$$\frac{\|p(x^n)\|_2}{\|x^n\|_2} = n\sqrt{\frac{2n+1}{2n-1}} \approx n, \text{ for } n \gg 1.$$

Thus, there does not exist a constant $C > 0$ such that $\|p(x^n)\|_2 / \|x^n\|_2 \leq C$ for all n; i.e., p is unbounded.

3.3 Kernel Construction

Given some initial kernels, we can construct new kernels using the following approach:

Theorem 3.8. *Let X be a non-empty data set. Let K_1 and K_2 be kernels defined on $X \times X$. Then, the following functions are kernels on $X \times X$:*

1. *$K(x,y) = K_1(x,y) + K_2(x,y)$.*

2. *$K(x,y) = \alpha K_1(x,y)$ for any positive real constant α.*

3. *$K(x,y) = K_1(x,y)K_2(x,y)$.*

4. *$K(x,y) = f(x)f(z)$, where $f : X \to \mathbb{R}$ is any function.*

Proof. Let $\{x_1, \ldots, x_m\}$ denote a generic set of m samples from X. Then, we need to show that $(K(x_i,x_j))$ is a positive semi-definite matrix.

(1,2) The first two statements are obvious $a^t(K_1 + K_2)a = a^tK_1a + a^tK_2a \geq 0$ and $a^t(\alpha K_1)a = \alpha a^t(K_1)a \geq 0$.
(3) This elementwise multiplication of two matrices is called the Hadamard or Schur product. The tensor product $K_1 \otimes K_2$ of two kernel matrices has eigenvectors that are tensor products

of the eigenvectors of K_1 and K_2 with eigenvalues that are products of the corresponding eigenvalues of K_1 and K_2. Hence, $K_1 \otimes K_2$ is positive semi-definite. The Schur product represents a submatrix of this tensor product and is thus positive semidefinite.

(4) $a^t K a = \sum_{ij} a_i f(x_i) f(x_j) a_j = \left(\sum_i a_i f(x_i) \right) \left(\sum_j a_j f(x_j) \right) \geq 0.$ □

Corollary 3.1. *Let X be a non-empty data set, and let K_1 be a kernel defined on $X \times X$. Then, the following functions are kernels on $X \times X$:*

1. *$K(x, y) = p(K_1(x, y))$, where p is a polynomial with positive coefficients.*
2. *$K(x, y) = \exp(K_1(x, y))$.*
3. *$K(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2} \right)$, where we have assumed that the data set X is a subset of $\mathbb{R}^n$.*

Proof. The first two statements follow directly from Theorem 3.8 by using mathematical induction. To check the third statement, notice that

$$\exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2} \right) = \left[\exp\left(-\frac{\|x\|^2}{2\sigma^2} \right) \exp\left(-\frac{\|y\|^2}{2\sigma^2} \right) \right] \exp\left(\frac{\langle x, y \rangle}{\sigma^2} \right),$$

which is a product of two kernels. □

3.4 Mercer Kernel

The notion of kernel has a close connection to the Green's function of a differential operator in physics; more precisely, it can be viewed as the integral kernel of the integral operator that is "inverse" of the given differential operator. Utilizing this connection, some kernels can be expanded in terms of the eigenfunctions of a compact self-adjoint integral operator. This expansion depends on the following spectral theorem:

Theorem 3.9 (Hilbert-Schmidt Theorem)**.** *Let T be a compact self-adjoint operator on a separable Hilbert space $\mathcal{H}$. Then, the set $\{\phi_n\}_{n=1}^\infty$ of eigenfunctions of T is a complete orthonormal basis of $\mathcal{H}$, and the corresponding eigenvalues $\lambda_n \to 0$ as $n \to \infty$.*

If T is compact, but not self-adjoint, then we still have:

Theorem 3.10. *Let $T : \mathcal{H} \to \mathcal{H}$ be a compact operator on a separable Hilbert space $\mathcal{H}$. Then, there exist not necessarily complete orthonormal sets $\{\psi\}_{n=1}^N$ and $\{\phi\}_{n=1}^N$ and positive real numbers $\{\lambda_n\}_{n=1}^N$, called singular values, satisfying $\lambda_n \to 0$, such that*

$$T = \sum_{n=1}^N \lambda_n \phi_n \langle \psi_n, \cdot, . \rangle$$

(Here, N may be finite or infinite, depending on T.)

REMARK 3.26. *This is the singular value decomposition of a compactor operators on an infinite dimensional Hilbert space.*

The following theorem tells us how to expand a compact self-adjoint operator using its complete orthonormal eigenbasis:

Theorem 3.11. *Let T be a compact self-adjoint operator on a separable Hilbert space $\mathcal{H}$. Then, there exists a complete orthonormal set $\{\phi\}_{n=1}^{\infty}$ of eigenfunctions of T and a set $\{\lambda_n\}_{n=1}^{\infty}$ of corresponding real eigenvalues satisfying $\lambda_n \to 0$, such that*

$$T = \sum_{n=1}^{\infty} \lambda_n \phi_n \langle \phi_n, \cdot \rangle. \tag{3.3}$$

Finally, the following theorem states when a kernel can be viewed as an integral kernel of a compact self-adjoint integral operator and thus be expanded in terms of its complete orthonormal eigenbasis:

Theorem 3.12 (Mercer). *Let X be a compact subset of $\mathbb{R}^n$. Let $K : X \times X \to \mathbb{R}$ be a symmetric continuous positive semi-definite function. Then, for all $x, y \in X$,*

$$\boxed{K(x,y) = \sum_n \lambda_n \phi_n(x) \phi_n(y)},$$

where ϕ_n and $\lambda_n \geq 0$ are the eigenfunctions and corresponding eigenvalues of the integral operator $T : L^2(X) \to L^2(X)$ defined as

$$T(f)(x) = \int_X K(x,y) f(y) dy.$$

The sum is uniformly convergent on $X \times X$.

Ideas behind the proof. Since K is symmetric, T is easily seen to be self-adjoint with respect to the L^2 inner product. The continuity of K implies that T is a compact operator. Hence, T is a compact self-adjoint operator on $L^2(X)$ and has a complete basis of eigenfunctions that can expand T as in (3.3), which is satisfied by the proposed kernel expansion. The non-negativity of eigenvalues follows from the fact that K is positive semi-definite. □

REMARK 3.27. *Note that we can rewrite the kernel as*

$$K(x,y) = \sum_n \tilde{\phi}_n(x)\, \tilde{\phi}_n(y),$$

where $\tilde{\phi}_n = \sqrt{\lambda_n}\, \phi_n$ are orthonormal w.r.t. the inner product

$$\langle \phi_n, \phi_m \rangle = \frac{\delta_{mn}}{\lambda_n}.$$

This expression is the infinite-dimensional analogue of the outer product expansion of a kernel matrix in Theorem 3.3. The RKHS corresponding to the Mercer kernel consists of functions *that have a finite norm with respect to this inner product; that is, the allowable functions f in the RKHS are subsets of $L^2(X)$,*

$$f = \sum_n a_n \phi_n,$$

such that $(a_n/\sqrt{\lambda_n}) \in \ell^2$.

REMARK 3.28. *The function K defining the integral operator T is called the integral kernel of T or the Green's function of the differential operator L that is related to T via*

$$L(u) = f \Longleftrightarrow u = T(f)$$

Example 3.9 (Heat Kernel on S^2)**.** *The formal solution to the heat equation*

$$\partial_t \phi(x,t) = \Delta \phi(x,t)$$

with an initial condition $\phi(x,0) = \phi_0(x)$ is given by

$$\phi(x,t) = e^{t\Delta}\phi_0(x) = \int_{S^2} K(x,y;t)\,\phi_0(y)\,dy \equiv T(\phi_0),$$

where $K(x,y;t)$ is the heat kernel. The eigenfunctions of $e^{t\Delta}$ are the spherical harmonics $Y_{\ell m}$ on S^2 with eigenvalues $e^{-\ell(\ell+1)t}$.

The expansion of K in terms of these eigenfunctions is

$$K(x,y;t) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} e^{-\ell(\ell+1)t}\, Y_{\ell m}(x) Y_{\ell m}(y)^*.$$

Note that $K(x,y;t)$ is the Green's function of the differential operator $e^{-t\Delta}$ satisfying

$$e^{-t\Delta}\phi(x,t) = \phi_0(x),$$

which has the integral version

$$\phi(x,t) = T(\phi_0(x)).$$

3.5 Linear Regression

3.5.1 Estimation of Parameters

Let $(X_1, \dots, X_n) \in \mathbb{R}^n$ be input random variables and $Y \in \mathbb{R}$ an output random variable, with a joint distribution $P(Y, X_1, \dots, X_n)$. We are interested in predicting Y given $X_1, \dots, X_n$ with a function $f(X_1, \dots, X_n)$, and the definition of the "best predictor" depends on what kind of prediction error we want to minimize. If we are to minimize the expected square prediction error defined as $E_{Y,X_1,\dots,X_n}[(Y - f(X_1, \dots, X_n))^2]$, then the minimizer is

$$f(x_1, \dots, x_n) = E_Y[Y | X_1 = x_1, \dots, X_n = x_n].$$

In linear regression, we approximate $E_Y[Y|X_1 = x_1, \ldots, X_n = x_n]$ with a linear function of X_i, i.e. $f(x_1, \ldots, x_n) = \beta_0 + \beta_1 x_1 + \cdots \beta_n x_n$. If $Y|X_1, \ldots, X_n$ is normal, then for m observation $(y_i, x_{i1}, \ldots, x_{in}), i = 1, \ldots, m$, the maximum likelihood estimate of β_j is obtained by maximizing

$$\log \prod_i p(y_i|x_{i1}, \ldots, x_{in}) \propto -\sum_i (y_i - E[Y|x_{i1}, \ldots, x_{in}])^2 = -\sum_i (y_i - f(x_{i1}, \ldots, x_{in}))^2.$$

REMARK 3.29. *Note that in regression analysis, the values of predictive variables are given to you, so you condition the response variable on these given values, which you can treat as numbers instead of random variables. Thus, we only need to specify the probability distribution of the response variable.*

Equivalently, we need to minimize

$$L = \sum_{i=1}^{m} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_n x_{in})^2 \tag{3.4}$$

over $\beta_0, \ldots, \beta_n$. This estimation process is called the least square estimate, which amounts to maximum likelihood estimate for normally distributed Y given $X_1, \ldots, X_n$. We can uniquely solve for β when the input variables are not co-linear. The problem is easier to solve in a matrix form. Let

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}.$$

We will assume that $m > n+1$ and that $\boldsymbol{X}$ has rank $n+1$; i.e., the $n+1$ features are not colinear. As is commonly done, we will assume that given $\boldsymbol{X}$, the conditional distribution of $\boldsymbol{Y}$ is characterized by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{3.5}$$

where $\boldsymbol{e}$ is a column vector of i.i.d. normal random variables with mean 0 and variance σ^2. In (3.5), $\boldsymbol{X}$ are $\boldsymbol{\beta}$ non-random, while $\boldsymbol{Y}$ and $\boldsymbol{e}$ are multivariate normal random variables:

$$\boxed{\begin{aligned} \boldsymbol{Y}|\boldsymbol{X} &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_{m\times m}) \\ \boldsymbol{e} &\sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{m\times m}) \end{aligned}} \tag{3.6}$$

where $\boldsymbol{I}_{m\times m}$ is an $m \times m$ identity matrix; we will drop the subscript $m \times m$ when it is clear from the discussion. Then, (3.4) becomes

$$L = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Differentiating with respect to $\boldsymbol{\beta}$ yields the following condition that the MLE estimate $\hat{\boldsymbol{\beta}}$ has to satisfy:

$$\boldsymbol{X}^t(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0. \tag{3.7}$$

If $\boldsymbol{X}^t\boldsymbol{X}$ is invertible, i.e. $\boldsymbol{X}$ has a full column rank, then

$$\boxed{\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y}.} \tag{3.8}$$

Even though $\boldsymbol{\beta}$ in (3.5) is non-random, its MLE estimate $\hat{\boldsymbol{\beta}}$ is a multivariate normal random variable. The conditional mean of $\hat{\boldsymbol{\beta}}$ is

$$E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^tE[\boldsymbol{Y}|\boldsymbol{X}] = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

so $\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$.

The conditional covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\operatorname{Cov}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t \operatorname{Cov}[\boldsymbol{Y}|\boldsymbol{X}]\,\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2.$$

To summarize, we have

$$\boxed{\hat{\boldsymbol{\beta}}\,|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1})}$$

Example 3.10. *For the case of univariate regression, i.e. $n = 1$, the $m \times 2$ data matrix $\boldsymbol{X}$ has the form*

$$\boldsymbol{X} = \begin{pmatrix} \mathbb{1} & x \end{pmatrix}, \quad x \in \mathbb{R}^m,$$

where $\mathbb{1}$ is the m-dimensional column vector containing 1 in all entries, as seen the definition of the mean-centering matrix $J = I_{m\times m} - \frac{1}{m}\mathbb{1}\mathbb{1}^t$. The Gram matrix has the form

$$\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} m & m\bar{x} \\ m\bar{x} & \|x\|_2^2 \end{pmatrix},$$

where $\bar{x} = \sum_{i=1}^m x_i/m$ is the sample mean of the predictor variable. To simplify notation, let y denote the column response vector $\boldsymbol{Y}$.

Denoting the sample mean of the response variable as $\bar{y} = \sum_{i=1}^m y_i/m$, the regression *coefficients are*

$$\begin{aligned}\begin{pmatrix}\hat\beta_0\\ \hat\beta_1\end{pmatrix} &= \frac{1}{m\|x\|_2^2 - m^2\bar{x}^2}\begin{pmatrix}\|x\|_2^2 & -m\bar{x}\\ -m\bar{x} & m\end{pmatrix}\begin{pmatrix}m\bar{y}\\ x^t y\end{pmatrix}\\ &= \frac{1}{m\|Jx\|_2^2}\begin{pmatrix}m\bar{y}\|x\|_2^2 - m\bar{x}\,x^t y\\ m(Jx)^t(Jy)\end{pmatrix}\\ &= \frac{1}{\|Jx\|_2^2}\begin{pmatrix}\bar{y}\|Jx\|_2^2 - \bar{x}(Jx)^t(Jy)\\ (Jx)^t(Jy)\end{pmatrix}\end{aligned}$$

We thus get

$$\hat\beta_1 = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}, \quad \textit{and} \quad \hat\beta_0 = \bar{y} - \hat\beta_1\bar{x},$$

where

$$\mathrm{Cov}(X,Y) \equiv \frac{(Jx)^t(Jy)}{m-1} = \frac{\sum_{i=1}^m x_i y_i - m\left(\frac{1}{m}\sum_{i=1}^m x_i\right)\left(\frac{1}{m}\sum_{i=1}^m y_i\right)}{m-1},$$

and

$$\mathrm{Var}(X) \equiv \frac{\|Jx\|_2^2}{m-1} = \frac{\sum_{i=1}^m x_i^2 - m\left(\frac{1}{m}\sum_{i=1}^m x_i\right)^2}{m-1}.$$

When there are only two samples, i.e. $m = 2$, the linear regression describes a straight line through the two points (x_1, y_1) and (x_2, y_2). Note that we can rewrite $\hat\beta_1$ in terms of the empirical Pearson correlation coefficient $r_{X,Y}$ as

$$\hat\beta_1 = r_{X,Y}\sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}.$$

3.5.2 Predicted Values and Residuals

We denote the predicted values as $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Using (3.8), we can rewrite $\hat{\boldsymbol{Y}}$ as

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y},$$

where we define the "hat" operator $\boldsymbol{H}$ as

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t.$$

Note that the rank of $\boldsymbol{H}$ is equal to the rank of $\boldsymbol{X}$.

From the definition of $\boldsymbol{H}$, we can easily check

Proposition 3.1. *$\boldsymbol{H}$ is symmetric, and it is also idempotent, i.e. $\boldsymbol{H}^2 = \boldsymbol{H}$. Thus, $(\boldsymbol{I} - \boldsymbol{H})$ is also idempotent.*

and

Proposition 3.2. *$\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$, or equivalently, $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = 0$.*

Hence, $\boldsymbol{H}$ is the orthogonal projection operator onto the subspace spanned by the columns of $\boldsymbol{X}$, and we thus have

$$\hat{\boldsymbol{Y}}^t(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = 0. \tag{3.9}$$

Definition 3.14 (Residuals)**.** *The residuals $\hat{\boldsymbol{e}}$, defined as $\hat{\boldsymbol{e}} := \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, are the estimates of the error $\boldsymbol{e}$.*

Because $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, we can rewrite the residuals as

$$\hat{\boldsymbol{e}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{e}.$$

Similar to (3.6), we have

Theorem 3.13. *The predicted values and residuals are multivariate normal random variables distributed as*

$$\boxed{\begin{aligned} \hat{\boldsymbol{Y}}\,|\boldsymbol{X} &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H})\,, \\ \hat{\boldsymbol{e}}\,|\boldsymbol{X} &\sim \mathcal{N}(0, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))\,. \end{aligned}}$$

Proof. Using Proposition 3.2, we see that $E[\hat{\boldsymbol{Y}}|\boldsymbol{X}] = \boldsymbol{H}E[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta}$. Since $\boldsymbol{H}$ is symmetric and idempotent, $\operatorname{Cov}[\hat{\boldsymbol{Y}}|\boldsymbol{X}] = \operatorname{Cov}[\boldsymbol{H}\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{H}\operatorname{Cov}[\boldsymbol{Y}|\boldsymbol{X}]\boldsymbol{H}^t = \sigma^2\boldsymbol{H}$. Similarly, $E[\hat{\boldsymbol{e}}|\boldsymbol{X}] = (\boldsymbol{I} - \boldsymbol{H})E[\boldsymbol{e}] = 0$, and $\operatorname{Cov}[\hat{\boldsymbol{e}}|\boldsymbol{X}] = (\boldsymbol{I} - \boldsymbol{H})\operatorname{Cov}[\boldsymbol{e}](\boldsymbol{I} - \boldsymbol{H})^t = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$. □

Proposition 3.3. *The residuals $\hat{\boldsymbol{e}}$ are independent of the predicted values $\hat{\boldsymbol{Y}}$ and the regression coefficients $\hat{\boldsymbol{\beta}}$.*

Proof. Since $E[\hat{\boldsymbol{e}}|\boldsymbol{X}] = E[\boldsymbol{e}] = 0$, we have

$$\begin{aligned} \operatorname{Cov}[\hat{\boldsymbol{e}}, \hat{\boldsymbol{Y}}|\boldsymbol{X}] &= E[\hat{\boldsymbol{e}}\hat{\boldsymbol{Y}}^t|\boldsymbol{X}] - E[\hat{\boldsymbol{e}}|\boldsymbol{X}]\,E[\hat{\boldsymbol{Y}}^t|\boldsymbol{X}] = E[\hat{\boldsymbol{e}}\hat{\boldsymbol{Y}}^t|\boldsymbol{X}] \\ &= E[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{e}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e})^t\boldsymbol{H}|\boldsymbol{X}] \\ &= E[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{e}\boldsymbol{e}^t\boldsymbol{H}|\boldsymbol{X}] \\ &= (\boldsymbol{I} - \boldsymbol{H})\sigma^2\boldsymbol{H} \\ &= 0. \end{aligned}$$

Since $\hat{\boldsymbol{e}}$ and $\hat{\boldsymbol{Y}}$ are jointly multivariate normal, the fact that $\hat{\boldsymbol{e}}$ and $\hat{\boldsymbol{Y}}$ are uncorrelated implies that they are independent.

From (3.7), we see that

$$\boldsymbol{X}^t\hat{\boldsymbol{Y}} = \boldsymbol{X}^t\boldsymbol{Y}.$$

Thus, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\hat{\boldsymbol{Y}}$. Since $\hat{\boldsymbol{e}}$ and $\hat{\boldsymbol{Y}}$ are independent, we thus see that $\hat{\boldsymbol{e}}$ and $\hat{\boldsymbol{\beta}}$ are also independent. □

We thus have the variance decomposition

$$\operatorname{Cov}[\boldsymbol{Y}|\boldsymbol{X}] = \operatorname{Cov}[\hat{\boldsymbol{Y}}|\boldsymbol{X}] + \operatorname{Cov}[\hat{\boldsymbol{e}}|\boldsymbol{X}]. \tag{3.10}$$

Using the conditional variance formula

$$\mathrm{Cov}[Y] = \mathrm{Cov}_X[E[Y|X]] + E_X[\mathrm{Cov}_Y[Y|X]],$$

the variance decomposition formula (3.10) yields

$$\mathrm{Cov}[\boldsymbol{Y}] = \mathrm{Cov}[\hat{\boldsymbol{Y}}] + \mathrm{Cov}[\hat{\boldsymbol{e}}],$$

where Cov in the last line includes variation over $\boldsymbol{X}$. A corresponding expression for sampled values is the following sum of squares formula:

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS},$$

where the total sum of squares (TSS), explained sum of squares (ESS), and residual sum of squares (RSS) are

$$\mathrm{TSS} = \sum_{i=1}^{m}(y_i - \bar{y})^2, \quad \mathrm{ESS} = \sum_{i=1}^{m}(\hat{y}_i - \bar{y})^2, \text{ and } \quad \mathrm{RSS} = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2.$$

The goodness-of-fit is assessed by computing the R^2 value defined as

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}}.$$

EXERCISE 3.4. *Show that* $\bar{\hat{y}} = \bar{y}$.

3.5.3 Ridge Regression and Bayesian Approach

Let us assume that we have mean centered the response and predictive variables. One potential problem of direct least square minimization or maximum likelihood approach is that of overfitting to outliers. To ameliorate this problem, we can add to the cost function a penalty for the magnitude of regression coefficients:

$$L_\alpha = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \alpha\|\beta\|_2^2. \tag{3.11}$$

This approach is known as the ridge regression.

REMARK 3.30. *Ridge regression is equivalent to adding a normal prior* $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2/\alpha)$ *and computing the maximum a posteriori (MAP) estimate of* $\boldsymbol{\beta}$. *For small* α, *the variance of this prior distribution is large, i.e. the distribution becomes flat. In the opposite limit of large* α, *the variance shrinks to 0, pulling the posterior distribution of* $\boldsymbol{\beta}$ *towards 0.*

To minimize (3.11), we differentiate it with respect to $\boldsymbol{\beta}$ and set it equal to 0:

$$\frac{\partial L_\alpha}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^t(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\alpha\boldsymbol{\beta} = 0,$$

from which we get

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X} + \alpha\boldsymbol{I}_{n\times n})^{-1}\boldsymbol{X}^t\boldsymbol{Y}. \tag{3.12}$$

REMARK 3.31. *Notice that the only different between this expression and* (3.8) *is the shift* $\boldsymbol{X}^t\boldsymbol{X} \to \boldsymbol{X}^t\boldsymbol{X} + \alpha\boldsymbol{I}_{n\times n}$.

3.6 Kernel Regression

Using Theorem A.5, (3.12) can be rewritten as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{X}^t(\boldsymbol{X}\boldsymbol{X}^t + \alpha\boldsymbol{I}_{m\times m})^{-1}\boldsymbol{Y}, \tag{3.13}$$

and

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{X}^t(\boldsymbol{X}\boldsymbol{X}^t + \alpha\boldsymbol{I}_{m\times m})^{-1}\boldsymbol{Y}.$$

Given a new point $x = (x_1 \cdots x_n)$ of predictive variables, our prediction is

$$\hat{y} = x\boldsymbol{X}^t(\boldsymbol{X}\boldsymbol{X}^t + \alpha\boldsymbol{I}_{m\times m})^{-1}\boldsymbol{Y}.$$

In kernel regression, we use the "kernel trick" to replace the $m \times m$ Gram matrix $\boldsymbol{X}\boldsymbol{X}^t$, consisting of pairwise dot products of the feature vectors of m samples, with a kernel matrix $\boldsymbol{K} = (\kappa(x_i, x_j))$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{K}(\boldsymbol{K} + \alpha\boldsymbol{I}_{m\times m})^{-1}\boldsymbol{Y}$$

and, given a new value x of predictive variables, our prediction is

$$\hat{y} = \kappa(x, \cdot)(\boldsymbol{K} + \alpha\boldsymbol{I}_{m\times m})^{-1}\boldsymbol{Y}\,.$$

where $\kappa(x, \cdot)$ is the row vector $(\kappa(x, x_1) \cdots \kappa(x, x_m))$.

Figure 3.1 illustrates that increasing α from 0.01 to 1.0 makes the regression curve smoother; however, increasing α even further introduces large errors in the prediction. This phenomenon captures the well-known trade-off between variance and bias of a learning algorithm. Reducing variance (making the curve smoother) often comes at the cost of increasing bias (misfit). Typically, α and other parameters in the kernel are chosen via cross-validation. Figure 3.1 illustrates how the width parameter σ in the Gaussian RBF controls how many neighbors to consider in the local regression. This parameter and α can be tuned to minimize the validation error

$$\|\boldsymbol{Y}_{\text{validation}} - \hat{\boldsymbol{Y}}_{\text{validation}}\|_2^2.$$

REMARK 3.32. *In kernel regression, the response and predictive variables do not need to be mean centered across samples.*

REMARK 3.33. *Kernel regression can be viewed as a nonlinear local regression that utilizes information from nearby samples, as assessed by the kernel evaluation, to predict the response.*

REMARK 3.34. *You may be wondering, "where is the estimation step in kernel regression?" Note that the step of estimating* $\hat{\boldsymbol{\beta}}$ *in linear regression is automatically built into the kernel regression formalism, and the MLE estimation formula* (3.13) *is exactly what has allowed the "kernel trick." In kernel regression, the training data are thus used to construct the kernel matrix and to learn the behavior of* y *as a local function of* x.

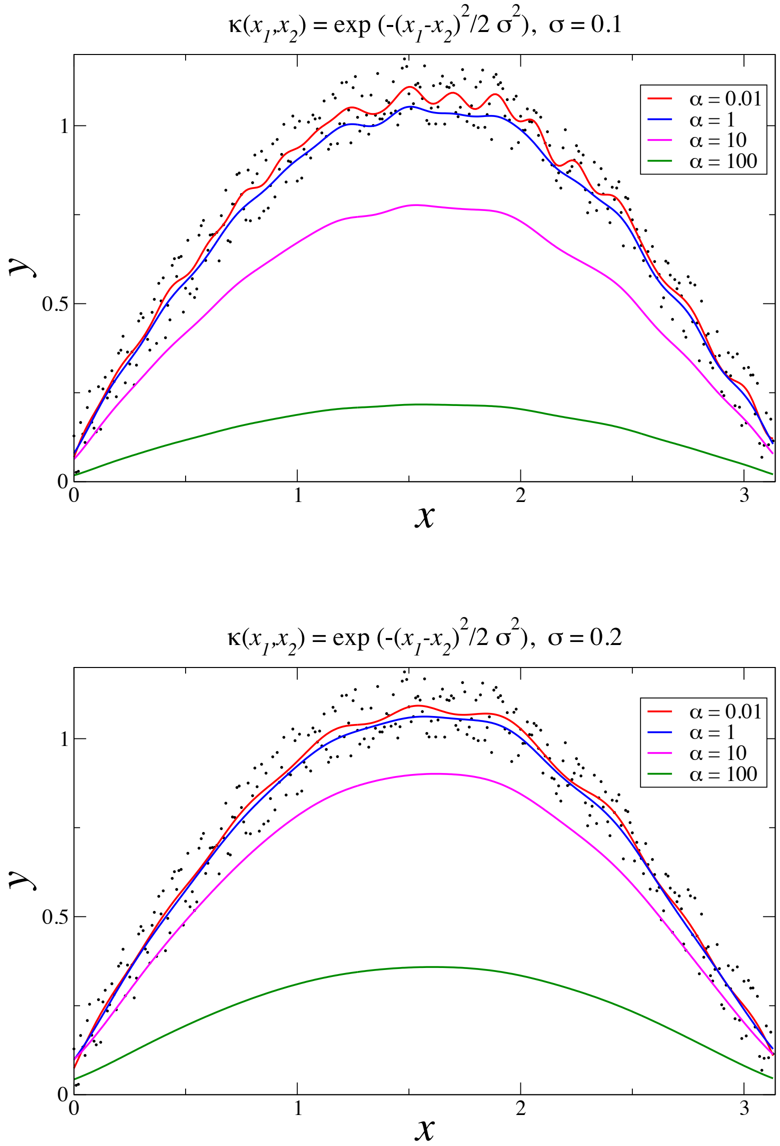

Figure 3.1: Kernel regression of a noisy sine curve using the Gaussian RBF kernel.