CHAPTER 3

Building function spaces from data

In the previous chapter we saw that certain special classes of nonlinear spaces could be linearized by unfolding, such as the case of a developable manifold. However, these methods to not apply to topologically or geometrically nontrivial spaces. In this chapter our goal is to find a way to perform 'linear' analysis on the nonlinear space. One way to do this is to move from the manifold to the space of functions defined on the manifold.

In fact, most practical goals of data analysis actually involve functions on the data. For example, interpolation, uncertainty quantification, and forecasting all involve representing certain classes of functions or operators on function spaces. The disadvantage of function spaces is that the spaces of interest are often infinite-dimensional, so we are restricted to representing a finite dimensional projection of the desired function space. The goal then becomes to construct finite dimensional function spaces that are adapted to the data set and converge to a meaningful infinite dimensional function space in the limit as the number of data points goes to infinity. This is a more far reaching perspective on the role of kernels in data science than the embedding perspective taken by KPCA.

3.1 Using a kernel to define a function space

Consider a finite data set $X = \{x_i\}_{i=1}^N \subset A$ contained in a measure space (A, μ) with measure μ. One could consider the function space composed of all real valued functions defined on X. However, this function space is purely combinatorial, and does not account for any relationships between the data points. Moreover, the data set X is only a finite sample, and we would like to extend our analysis beyond this data set to unobserved points in A. We would like to assume as little as possible about the data set X and the type of functions on X which may be of interest.

In order to represent relationships between data points and connect the observed data to unobserved points we will use a kernel function.

Definition 3.1 For data $x_i \in A$, a *kernel function* is a function $k : A \times A \to \mathbb{R}$ on pairs of points. A kernel function is symmetric if $k(x_i, x_j) = k(x_j, x_i)$. The matrix $\mathbf{K}_{ij} = k(x_i, x_j)$ associated to a kernel function k and a data set $\{x_i\}_{i=1}^N$ is called the *kernel matrix*. For brevity we often refer to a kernel function as a *kernel*.

The kernel function k is assumed to be known explicitly and computable on any pairs of points, although it may also depend on parameters which will be 'tuned' as part of a learning algorithm. The effect of the kernel is to spread the influence of each discrete data point to an area surrounding the point. Note that the use of a kernel puts an implicit assumption of smoothness on the process that is producing the data set.

Example 3.1 [Prototypical kernel] A crucial example of a kernel $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is given by

$$k(x, y) = \exp\left(-\frac{||x - y||^2}{2\delta}\right),$$

which is closely related to the both the Gaussian density function and Green's function for a Brownian diffusion.

The prototypical kernel can be generalized to any metric space by substituting the distance between x and y for $||x - y||$. The parameter δ is called the *bandwidth* of the kernel. The exponential function can also be

replaced with any shape function that decays as the distance between points increases. We will see in Chapter 7 that the exponential kernel function is the prototypical example of a *local* kernel, so-called because only nearby points have strong interactions due to the exponential decay.

Definition 3.2 When the set A containing the data set is a metric space, a kernel function that decays as a function of the distance between points is called a *radial basis function* (RBF) kernel. If the kernel function decays faster than any polynomial in the distance we call it a *local* kernel.

A kernel function allows us to consider functions of the form $k_{x_i} = k(\cdot, x_i) : A \to \mathbb{R}$ which are defined on all of A. Notice that for the prototypical kernel above, each k_{x_i} is a bump function localized near x_i, and for nearby points these bump functions will have a large overlap. We can now consider the function space of all possible linear combinations of these functions

$$\mathscr{H} \equiv \text{span}\{k_{x_i}\}_{i=1}^N = \left\{ \sum_{i=1}^N c_i k(\cdot, x_i) \,:\, c \in \mathbb{R}^N \right\}.$$

By definition the functions k_{x_i} are a spanning set for $\mathscr{H}$, but depending on the form of the kernel function and the locations of the data points, the functions k_{x_i} may not be linearly independent and typically they will not be orthogonal. A given function $f \in \mathscr{H}$ may have multiple representations as linear combinations of the k_{x_i} functions due to the relationships of the data points as measured by the kernel k. This is rare when the kernel function is nonlinear, and for the prototypical kernel happens only if a data point is repeated.

For a function $f \in \mathscr{H}$, we will denote the coefficients by the vector $\tilde{f}$, in other words, $\tilde{f}$ is defined by

$$f = \sum_{i=1}^N \tilde{f}_i k_{x_i}.$$

We will denote the values of f at the data points x_j by the vector $\mathbf{f}$, meaning that

$$\mathbf{f}_j \equiv f(x_j) = \sum_{i=1}^N \tilde{f}_i k(x_j, x_i) = \sum_{i=1}^N \mathbf{K}_{ji} \tilde{f}_i = \left(\mathbf{K}\tilde{f}\right)_j = \tilde{f}^T \mathbf{k}_{x_j}. \tag{3.1}$$

This equation shows that $\mathbf{f} = \mathbf{K}\tilde{f}$, so if the kernel matrix $\mathbf{K}$ is full rank the function $f \in \mathscr{H}$ is completely determined by its values on the data set since we can compute the coefficients by $\tilde{f} = \mathbf{K}^{-1}\mathbf{f}$. If $\mathbf{K}$ is not invertible, or if $f \notin \mathscr{H}$ but we know the values f on the data set, then we can define a function in $\mathscr{H}$ with coefficients $\tilde{f} = \mathbf{K}^{\dagger}\mathbf{f}$ which will agree with f on the data set, but may not be equal to f elsewhere.

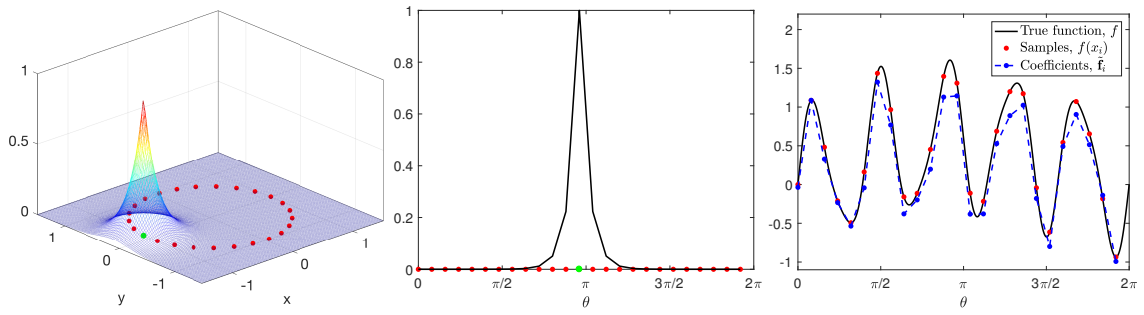
Example 3.2 [Data on $S^1 \subset \mathbb{R}^2$] Consider $N = 25$ evenly-spaced points $\theta_i = 2\pi(i-1)/25$ on $S^1 \subset \mathbb{R}^2$ with the embedding $x_i = h(\theta_i)$ given by

$$h(\theta) = (\cos(\theta), \sin(\theta))^T,$$

shown as the red and green points in the (x, y)-plane in Fig. 3.1(a).

We will use the kernel function $k(x_i, x_j) = \exp\left(-||x_i - x_j||^2/(2\delta^2)\right)$ with bandwidth $\delta = 0.3$. The function $k_{x_{13}} : \mathbb{R}^2 \to \mathbb{R}$ is shown in Fig. 3.1(a). It is a bump function centered at x_{13} which is highlighted in green. In order to make Fig. 3.1(a) we generated a grid of points $\{y_j\}_{j=1}^{N_y} \subset [-1.5, 1.5]^2$ and evaluated the kernel function $k_{x_{13}}(y_j)$ at the grid points. The kernel restricted to the circle is also shown as a function of θ given by $k_{x_{13}} \circ h$ in Fig. 3.1(b).

Next we consider the function $f(\theta) = 0.6\sin(\theta/2) + 0.8\sin(5\theta) + 0.2\sin(9\theta)$ defined on the circle. We sample this function at the data points, setting $\mathbf{f}_i = f(\theta_i)$ which is a 25×1 vector. In Fig. 3.1(c) we show the function f (black) and the discrete samples (red points). We then computed the 25×25 kernel matrix $\mathbf{K}$ and the function coefficients $\tilde{f} = \mathbf{K}^{-1}\mathbf{f}$ which are shown as the blue points. Notice that the function coefficients $\tilde{f}$ are highly correlated with the function samples $\mathbf{f}$. This is due to the shape of the kernel function which is narrow and centered on the point x_i and roughly approximates a delta function shape.



(a) (b) (c)

Fig. 3.1: (a) The function $k_{x_i} = k(\cdot, x_i) : \mathbb{R}^2 \to \mathbb{R}$ defined by the kernel at the data point x_i (green point), the full data set includes the red points and the green point. (b) The function $k_{x_i} \circ h : [0, 2\pi) \to \mathbb{R}$ (black curve) where $h : [0, 2\pi) \to \mathbb{R}^2$ is the embedding of the unit circle into the plane. The inverse images of the data points $h^{-1}(x_i)$ under the embedding are shown by the red and green points on the x-axis. (c) The periodic function $f(\theta) = 0.6\sin(\theta/2) + 0.8\sin(5\theta) + 0.2\sin(9\theta)$ (black curve) is evaluated on the inverse images of the data set (red points) and the function coefficients $\tilde{f} = \mathbf{K}^{-1}\mathbf{f}$ are plotted as points $(\theta_i, \tilde{f}_i)$ (blue, dashed curve).

In the next section we will place additional assumptions on the kernel function which allow us to define an inner product on $\mathscr{H}$. We can then construct an orthonormal basis for $\mathscr{H}$. The orthonormal basis will be used to interpolate functions on $\mathscr{H}$.

3.2 An orthonormal basis for $\mathscr{H}$

In order to define an orthonormal basis for $\mathscr{H}$ we first need to define an inner product on $\mathscr{H}$. Since $\mathscr{H}$ is finite dimensional, this inner product will make $\mathscr{H}$ into a Hilbert space. We will then define a natural operator on $\mathscr{H}$ that is self-adjoint with respect to the inner product, which yields an orthonormal basis of eigenfunctions of this operator. These properties require that we restrict our attention to a special class of kernels that are symmetric and positive semidefinite.

3.2.1 An inner product on $\mathscr{H}$

A symmetric kernel function gives rise to symmetric kernel matrices, and if those matrices are also positive semidefinite then they can be interpreted as Gram matrices with respect to an appropriate inner product. If a kernel function gives rise to a positive semidefinite matrix for any finite data set then that kernel function is called positive semidefinite.

Definition 3.3 A symmetric kernel function is called *positive semidefinite* if for all finite data sets $\{x_i\}_{i=1}^N \subset A$, $\mathbf{K}$ is a positive semidefinite matrix.

For a symmetric and positive semidefinite kernel function k, we can equip the associated function space $\mathscr{H} = \text{span}\{k_{x_i}\}$ with a natural inner product defined by

$$\langle k_{x_i}, k_{x_j} \rangle_{\mathscr{H}} \equiv k(x_i, x_j) = \mathbf{K}_{ij} \tag{3.2}$$

and extended linearly to the rest of $\mathscr{H}$. If $f, g \in \mathscr{H}$, then the inner product is given by the kernel matrix

$$\langle f, g \rangle_{\mathscr{H}} = \sum_{i,j=1}^{N} \tilde{f}_i \langle k_{x_i}, k_{x_j} \rangle \tilde{g}_j = \tilde{f}^\top \mathbf{K} \tilde{g}. \tag{3.3}$$

The finite dimensional linear space $\mathscr{H}$ together with the inner product $\langle\cdot,\cdot\rangle_{\mathscr{H}}$ defines a Hilbert space of function defined on the entire set A.

Fact 1. For $1 \le i \le N$, $\langle f, k_{x_i}\rangle_{\mathscr{H}} = f(x_i)$.

Proof. By definition of the inner product, $\langle f, k_{x_i}\rangle_{\mathscr{H}} = \tilde{f}^T \mathbf{K}\, \tilde{k}_{x_i} = \tilde{f}^T \mathbf{k}_{x_i} = f(x_i)$, where we use (3.1) for the last equality.

Fact 2. For $x \in A$, define the coefficient vector $\tilde{k}_x = \mathbf{K}^{-1}[k_{x_1}(x), \ldots, k_{x_N}(x)]^T$. Then $\langle f, k_x\rangle = f(x)$.

Proof. It suffices to check the equality for f in the spanning set $\{k_{x_i}\}$, for which

$$\langle k_{x_i}, k_x\rangle = \tilde{k}_{x_i}^T \mathbf{K} \tilde{k}_x = [0,0,\ldots,1,\ldots,0]\cdot[k_{x_1}(x),\ldots,k_{x_N}(x)]^T = k_{x_i}(x).$$

If we further assume that the kernel function k is bounded then $\mathscr{H}$ has an additional structure and is known as a reproducing kernel Hilbert space.

Definition 3.4 A Hilbert space $(\mathscr{H}, \langle\cdot,\cdot\rangle_{\mathscr{H}})$ of real valued functions $f : A \to \mathbb{R}$ is called a *reproducing kernel Hilbert space* (RKHS) if the evaluation functional $L_x : f \mapsto f(x)$ is bounded for all $x \in A$. The *reproducing kernel* of a RKHS is the map $k : A \to \mathscr{H}$ defined by $k : x \mapsto k_x$ where $\langle f, k_x\rangle_{\mathscr{H}} = L_x(f) = f(x)$. Recall that k_x is the Riesz representative of the functional L_x which exists and is unique by the Riesz representation theorem.

Due to our assumption that the kernel function k is bounded, the evaluation functionals

$$L_x : f \mapsto f(x) = \sum_{j=1}^{N} \tilde{\mathbf{f}}_j k(x, x_j)$$

are all bounded so $\mathscr{H}$ is a RKHS. To find the reproducing kernel of this RKHS notice that

$$\mathbf{f}_j = f(x_j) = \sum_{i=1}^{N} \tilde{\mathbf{f}}_i k(x_j, x_i) = \sum_{i=1}^{N} \tilde{\mathbf{f}}_i \langle k_{x_i}, k_{x_j}\rangle_{\mathscr{H}} = \left\langle \sum_{i=1}^{N} \tilde{\mathbf{f}}_i k_{x_i}, k_{x_j} \right\rangle_{\mathscr{H}} = \langle f, k_{x_j}\rangle_{\mathscr{H}}$$

meaning that evaluating a function $f \in \mathscr{H}$ at a point x_j is equivalent to computing the inner product of f with k_{x_j}. Thus the reproducing kernel of $\mathscr{H}$ is defined on the data points by $x_i \mapsto k_{x_i} \equiv k(\cdot, x_i)$ making use of the kernel function k. For a general point $x \in A$ the reproducing kernel takes x to the function $k_x \in \mathscr{H}$ defined by

$$k_x = \sum_{i,j=1}^{N} \left(\tilde{\mathbf{k}_x}\right)_i k_{x_i} \tag{3.4}$$

where the i-th coefficient of k_x is given by

$$(\tilde{\mathbf{k}_x})_i = \sum_{j=1}^{N} k(x, x_j)\mathbf{K}_{ij}^{-1}. \tag{3.5}$$

To see that k_x satisfies $\langle f, k_x\rangle_{\mathscr{H}} = f(x)$ we compute

$$\langle f, k_x\rangle_{\mathscr{H}} = \tilde{\mathbf{f}}^\top \mathbf{K}\tilde{\mathbf{k}_x} = \sum_{i,j,s=1}^{N} \tilde{\mathbf{f}}_s \mathbf{K}_{si} k(x, x_j)\mathbf{K}_{ij}^{-1} = \sum_{j=1}^{N} \tilde{\mathbf{f}}_j k(x, x_j) = f(x)$$

where we simplify $\sum_{i=1}^{N} \mathbf{K}_{si}\mathbf{K}_{ij}^{-1} = \delta_{sj}$ and then sum over $j = s$. Similarly, notice that plugging a data point x_s into (3.4) we recover $k_{x_s} = k(\cdot, x_s)$.

Definition 3.5 [Discrete RKHS] The space $\mathscr{H} = \text{span}\{k_{x_i}\}$ associated to a data set $\{x_i\}_{i=1}^N \subset A$ by a kernel function k is a RKHS with reproducing kernel $x \mapsto k_x = \sum_{i,j=1}^{N} k(x, x_j)\mathbf{K}_{ij}^{-1} k_{x_i}$ and $\mathscr{H}$ is called the *discrete RKHS* associated to the data set by the kernel function.

Example 3.3 [Reproducing kernel for data on $S^1 \subset \mathbb{R}^2$] Using the data set $\{\mathbf{x}_i\}_{i=1}^N$ with $N = 25$ and the exponential kernel function from example 3.2 we computed the reproducing kernel based at 4 different points in the plane as shown in Fig. 3.2. Notice that the reproducing kernel based at the data point $\mathbf{x}_{13}$ is shown in Fig. 3.1(a).

In order to plot the reproducing kernel centered at non-data points we first use (3.5) to find the vector of coefficients $\tilde{\mathbf{k}}_x$ for the given point $x \in \mathbb{R}^2$. We evaluate the kernel function $(\mathbf{k}_x)_j = k_{x_j}(x) = k(x, x_j)$ to form the $N \times 1$ vector $\mathbf{k}_x$ and we multiply by the inverse of the kernel matrix to find the $N \times 1$ vector of coefficients $\tilde{\mathbf{k}}_x = \mathbf{K}^{-1}\mathbf{k}_x$. The coefficients $\tilde{\mathbf{k}}_x$ define the reproducing kernel k_x by (3.4). Next, we generated a grid of points $\{y_j\}_{j=1}^{N_y} \subset [-1.5, 1.5]^2$ and evaluated the kernel function

$$\mathbf{K}^{yx}_{ji} = k(y_j, x_i) = k_{x_i}(y_j)$$

to form the $N_y \times N$ matrix K^{yx} whose columns are the bump functions k_{x_i} centered at the data points. Finally, in order evaluate the reproducing kernel k_x at the grid points y_j we simply multiply

$$k_x(y_j) = \left(\mathbf{K}^{yx}\tilde{\mathbf{k}}_x\right)_j = \sum_{i=1}^{N} \mathbf{K}^{yx}_{ji} \left(\tilde{\mathbf{k}}_x\right)_i = \sum_{i=1}^{N} \left(\tilde{\mathbf{k}}_x\right)_i k_{x_i}(y_j)$$

which is exactly (3.4).

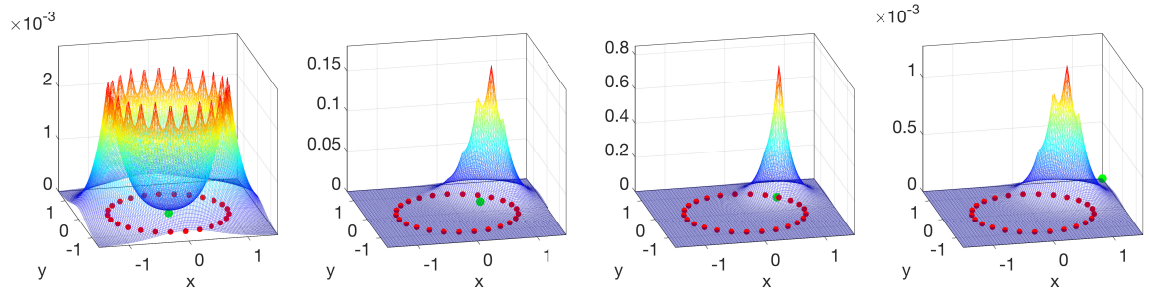


Fig. 3.2: For the data set from example 3.2 (red points) the reproducing kernel $k_x : \mathbb{R}^2 \to \mathbb{R}$ is drawn based at four different points $x \in \mathbb{R}^2$ (green point). Each reproducing kernel is a weighted sum of the reproducing kernels k_{x_i} centered on the data points, which leads to the shape with peaks near data points. Notice that for points far from the circle the reproducing kernel decays very rapidly for this choice of kernel function.

3.2.2 A basis of eigenfunctions for $\mathscr{H}$

In addition to the kernel matrix $\mathbf{K}$ defining an inner product (3.2) it also defines an operator $K : \mathscr{H} \to \mathscr{H}$. Notice that if $f = \sum_{i=1}^N \tilde{\mathbf{f}}_i k_{x_i}$ then the matrix $\mathbf{K}$ will map the coefficient vector $\tilde{\mathbf{f}}$ to another coefficient vector $\mathbf{K}\tilde{\mathbf{f}}$ given by

$$\left(\mathbf{K}\tilde{\mathbf{f}}\right)_j = \sum_{i=1}^{N} \mathbf{K}_{ji}\tilde{\mathbf{f}}_i = \sum_{i=1}^{N} k(x_j, x_i)\tilde{\mathbf{f}}_i = f(x_j).$$

which is simply the evaluation of the original function on the data points, so $\mathbf{K}\tilde{\mathbf{f}} = \mathbf{f}$. We now think of the vector $\mathbf{f}$ as defining the coefficients of the image of f under the operator K. In other words the function $Kf \in \mathscr{H}$ is defined by $Kf = \sum_{j=1}^N f(x_j)k_{x_j}$ so that

$$(Kf)(x_i) = \sum_{j=1}^{N} k(x_i, x_j)f(x_j).$$

The previous equation shows that that the j-th coefficient of the function Kf is $\mathbf{f}_j = f(x_j)$, so we see that $\widetilde{(Kf)} = \mathbf{f}$. We call K the kernel operator.

Definition 3.6 For a data set $X = \{x_i\}_{i=1}^N \subset A$ and a kernel function k with associated RKHS $\mathscr{H}$, we define the *kernel operator* $K : \mathscr{H} \to \mathscr{H}$ by $Kf(x) = \sum_{i=1}^N k(x,x_i) f(x_i)$.

Also, notice that if $g = \sum_{i=1}^N \tilde{\mathbf{g}}_i k_{x_i}$ then

$$\langle Kf, g\rangle_{\mathscr{H}} = \sum_{i,j=1}^N \mathbf{f}_j \langle k_{x_j}, k_{x_i}\rangle \tilde{\mathbf{g}}_i = \mathbf{f}^\top \mathbf{K}\tilde{\mathbf{g}} = \tilde{\mathbf{f}}^\top \mathbf{K}^\top \mathbf{K}\tilde{\mathbf{g}} = \langle f, Kg\rangle_{\mathscr{H}}$$

which proves the following result.

Lemma 3.7 The kernel operator K is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{\mathscr{H}}$.

We also have that

$$\langle Kf, g\rangle_{\mathscr{H}} = \mathbf{f}^T \mathbf{K}\tilde{\mathbf{g}} = \mathbf{f}^T \mathbf{g} = \mathbf{g}^T \mathbf{f} = \langle f, Kg\rangle_{\mathscr{H}}$$

so that $\langle Kf, g\rangle_{\mathscr{H}} = \mathbf{f}^\top \mathbf{g} = \sum_{i=1}^N f(x_i) g(x_i)$ and we will return to this later.

Since $\mathscr{H}$ is finite dimensional, the operator K is compact in addition to being self-adjoint so the spectral theorem states that $\mathscr{H}$ has an orthonormal basis $\{\phi_\ell\}_{\ell=1}^N$ of eigenfunctions $K\phi_\ell = \lambda_\ell \phi_\ell$. Notice that if $f = \sum_{i=1}^N \tilde{\mathbf{f}}_i k_{x_i}$ satisfies $\mathbf{K}\tilde{\mathbf{f}} = \lambda\tilde{\mathbf{f}}$ then

$$\widetilde{Kf} = \mathbf{f} = \mathbf{K}\tilde{\mathbf{f}} = \lambda\tilde{\mathbf{f}}$$

so $Kf = \sum_j \widetilde{Kf}_j \mathbf{k}_{x_j} = \sum_j \lambda \tilde{\mathbf{f}}_j \mathbf{k}_{x_j} = \lambda f$ meaning that f is an eigenfunction of the operator K with eigenvalue λ. Notice that since we assume that the kernel matrix $\mathbf{K}$ is symmetric and full rank, there are N independent eigenvectors, and the corresponding eigenfunctions must span the N-dimensional Hilbert space $\mathscr{H}$. The above discussion indicates that we can find the eigenfunctions ϕ_ℓ and eigenvalues λ_ℓ of the operator K simply by finding the eigenvectors $\mathbf{v}_\ell$ and eigenvalues λ_ℓ of the kernel matrix $\mathbf{K}$.

Since $\mathbf{K}$ is symmetric it has an eigendecomposition $\mathbf{K} = \mathbf{V}\Lambda\mathbf{V}^\top$ where $\Lambda_{\ell\ell} = \lambda_\ell$ is a diagonal matrix of eigenvalues and the ℓ-th column of V is the corresponding eigenvector $\mathbf{v}_\ell$. Note that by symmetry the eigenvalues are all real and since $\mathbf{K}$ is positive semi-definite they are all non-negative and we will assume that the eigenvalues are decreasing $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$. The eigenfunctions are given by $\phi_\ell = \lambda_\ell^{-1/2} \sum_{i=1}^N \mathbf{V}_{i,\ell} k_{x_i}$ where $\mathbf{V}$ is the matrix whose columns are the eigenvectors of $\mathbf{K}$. In other words, the vector representation of the eigenfunction ϕ_ℓ is

$$\tilde{\phi}_\ell = \lambda_\ell^{-1/2} \mathbf{v}_\ell.$$

The factor $\lambda_\ell^{-1/2}$ is included to insure the correct normalization of the eigenfunctions in order to obtain an orthonormal basis. We assume that the eigenvectors are normalized so that $\mathbf{v}_\ell^\top \mathbf{v}_\ell = 1$, so by (3.3) we find that

$$||\phi_\ell||_{\mathscr{H}}^2 \equiv \langle \phi_\ell, \phi_\ell\rangle_{\mathscr{H}} = \frac{1}{\lambda_\ell} \mathbf{v}_\ell^\top \mathbf{K}\mathbf{v}_\ell = \mathbf{v}_\ell^\top \mathbf{v}_\ell = 1.$$

Notice that the factors $\lambda_\ell^{-1/2}$ combine to cancel the factor λ_ℓ arising from the presence of the kernel matrix in the inner product. Moreover, since $\tilde{\phi}_\ell = \lambda_\ell^{-1/2} \mathbf{v}_\ell$ we can evaluate ϕ_ℓ at a point x_i by

$$\phi_\ell(x_i) = \left(\mathbf{K}\tilde{\phi}_\ell\right)_i = \lambda_\ell \left(\tilde{\phi}_\ell\right)_i = \lambda_\ell^{1/2} \left(\mathbf{v}_\ell\right)_i.$$

This means that the columns of the matrix $\mathbf{V}\Lambda^{1/2}$ contain the evaluation of the ℓ-th eigenfunction on the data set, so $\phi_\ell(x_i) = \left(\mathbf{V}\Lambda^{1/2}\right)_{i,\ell}$. The above discussion proves the following lemma.

Lemma 3.8 The Hilbert space $\mathscr{H}$ associated to a kernel function k and a data set $\{x_i\}_{i=1}^N \subset A$ has an orthonormal basis $\{\phi_\ell\}_{\ell=1}^N$ given by the eigenfunctions $K\phi_\ell = \lambda_\ell \phi_\ell$ of the kernel operator K. The eigenfunctions are given by

$$\phi_\ell = \sum_{i=1}^N \lambda_\ell^{-1/2} (\mathbf{v}_\ell)_i k_{x_i} \qquad \text{and} \qquad \phi_\ell(x_i) = \lambda_\ell^{1/2} (\mathbf{v}_\ell)_i$$

where $\mathbf{K}v_\ell = \lambda_\ell v_\ell$ are the eigenvectors of the kernel matrix.

Finally, if the kernel matrix is not full rank, then the rank determines the dimension of $\mathscr{H}$ and the eigenfunctions with non-zero eigenvalues form an orthonormal basis.

Algorithm 3.1 Kernel based function space.

Inputs: Data set $\{x_i\}_{i=1}^N \subset A$, bounded, symmetric, and positive semidefinite kernel function $k : A \times A \to \mathbb{R}$
Outputs: Eigenfunction coefficients $\{\tilde{\phi}_\ell\}_{\ell=1}^N$ and eigenvalues λ_ℓ of the kernel operator $K : \mathscr{H} \to \mathscr{H}$.

1. Compute the kernel matrix $\mathbf{K}_{ij} = k(x_i, x_j)$
2. Compute the eigendecomposition $\mathbf{K} = \mathbf{V}\Lambda\mathbf{V}^\top$
3. The diagonal entries λ_ℓ of Λ are the eigenvalues of the kernel operator.
4. Define $\tilde{\Phi} = \mathbf{V}\Lambda^{-1/2}$, where $\tilde{\phi}_\ell$ is the ℓ-th column of $\tilde{\Phi}$.
5. Optionally, define $\Phi = \mathbf{V}\Lambda^{1/2}$ where $\phi_\ell(x_i) = \Phi_{i\ell}$.

Example 3.4 [Eigenfunctions for the discrete RKHS of example 3.2] Using Algorithm 3.1 we compute the eigenfunction $\phi_\ell(x_i)$ of the kernel operator evaluated on the data points $\{x_i\}_{i=1}^{25} \subset S^1 \subset \mathbb{R}^2$ from example 3.2.

All $N = 25$ eigenfunctions are plotted in Fig. 3.3 as functions of θ_i by composing each eigenfunction ϕ_ℓ with the embedding function h introduced in example 3.2. The eigenfunctions are shown in order of decreasing eigenvalues. Notice that the first few eigenfunctions approximate sine and cosine functions of various frequencies and phase shifts. This is not a coincidence and will be explained in Chapter 5.

Example 3.5 [Smoothing a noisy function] The goal of this example is to recover a function f from noisy samples $\mathbf{f}_i^s = \mathbf{f}_i + \nu_i = f(x_i) + \nu_i$ where ν_i are independent identically distributed Gaussian random variables with mean zero and standard deviation σ.

Since the columns of $\tilde{\Phi}$ are orthogonal, the inner products

$$\hat{\mathbf{f}}^s = \tilde{\Phi}^\top \mathbf{f}^s = \tilde{\Phi}^\top \mathbf{f} + \tilde{\Phi}^\top \nu$$

are corrupted by independent mean zero random variables $\omega_\ell = \left(\tilde{\Phi}^\top \nu\right)_\ell$ with covariance matrix

$$\mathbb{E}[\omega\omega^\top] = \tilde{\Phi}^\top \mathbb{E}[\nu\nu^\top]\tilde{\Phi} = \tilde{\Phi}^\top(\sigma I)\tilde{\Phi}^\top = \sigma\Lambda^{-1/2}\mathbf{V}^\top\mathbf{V}\Lambda^{-1/2} = \sigma\Lambda^{-1}.$$

We can now denoise the function f by eliminating coefficients which are below the noise floor. One method is to set a coefficient $\hat{\mathbf{f}}_j^s$ equal to zero if we find $|\hat{\mathbf{f}}_j^s| \leq \frac{3\sigma}{\lambda_j}$ so we define the denoised coefficients

$$\hat{\mathbf{f}}_j^d = \begin{cases} \hat{\mathbf{f}}_j^s & \text{if } |\hat{\mathbf{f}}_j^s| > \frac{3\sigma}{\lambda_j} \\ 0 & \text{else} \end{cases}$$

and we can then reconstruct the denoised samples $\mathbf{f}^d = \Phi\hat{\mathbf{f}}^d$. The above criterion states that we assume a coefficient is noise unless it is greater in magnitude than three standard deviations, meaning we are 99.8%

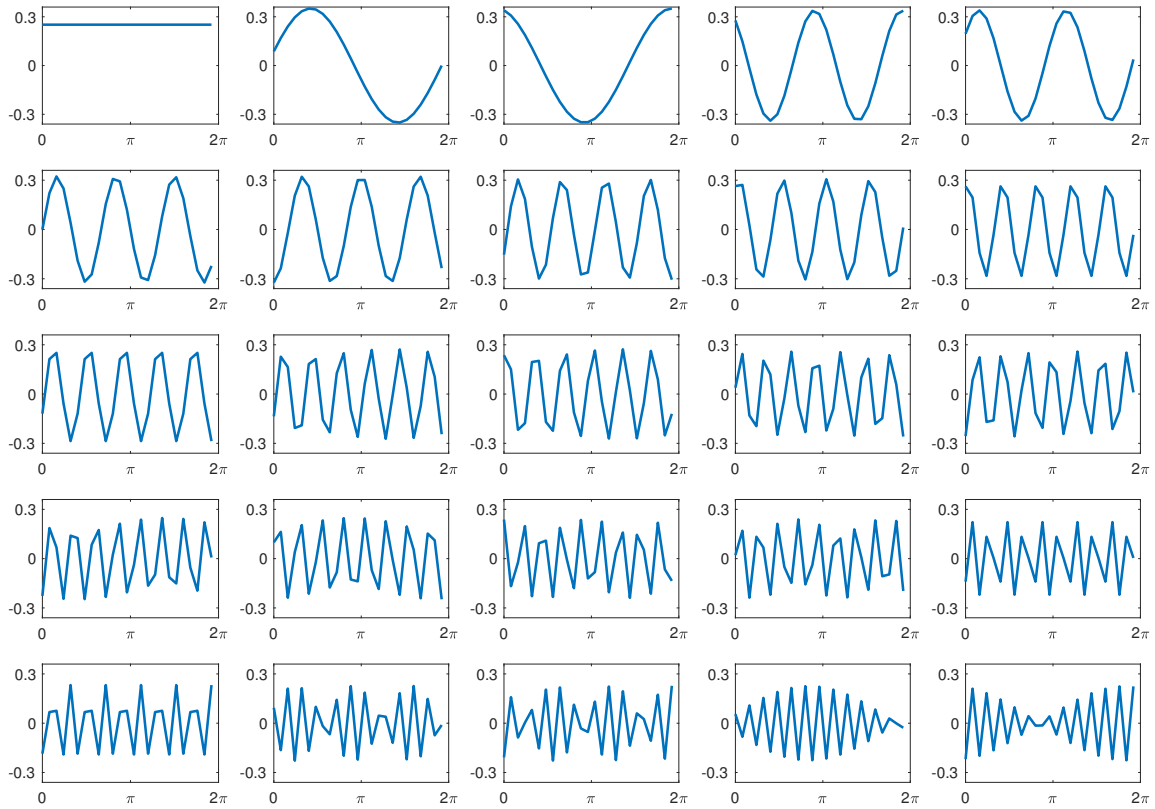


Fig. 3.3: Eigenfunctions of the kernel operator for the $N = 25$ points on the unit circle embedded in the plane from example 3.2 shown as functions of the intrinsic variable θ_i by composing with the embedding function h. The eigenfunctions $\phi_\ell \circ h : [0, 2\pi) \to \mathbb{R}$ are shown above in order of decreasing eigenvalue.

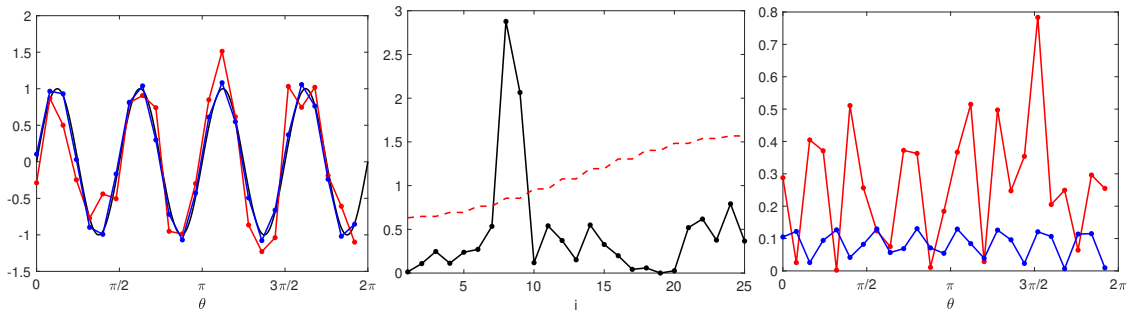
confident that the coefficient contains information beyond the noise. In Fig. 3.4(a) we show a function $f \circ h(\theta) = \sin(4\theta)$ (black) along with noisy samples $\mathbf{f}^s$ (red) and the denoised function (blue). The sample coefficients $\hat{\mathbf{f}}^s$ are shown in Fig. 3.4(b) (black, solid) along with the three standard deviation noise floor $3\sigma/\lambda_j$ (red, dashed). Notice that the dominant coefficients reveal f to be well approximated as a linear combination of the 8-th and 9-th eigenfunctions. Finally, in Fig. 3.4(c) we show the original error $|\mathbf{f}^s_i - f(x_i)| = |\nu_i|$ (red) between the noisy samples and the the true values as well as the error $|\mathbf{f}^d_i - f(x_i)|$ (blue) between the denoised samples and the true values.

3.3 Nyström extension and interpolation

Since $\phi(x_i) = \lambda_\ell \left(\tilde{\phi}_\ell\right)_i$ we have the following expression for the function ϕ_ℓ

$$\phi_\ell = \sum_{i=1}^{N} (\tilde{\phi}_\ell)_i k_{x_i} = \frac{1}{\lambda_\ell} \sum_{i=1}^{N} \phi_\ell(x_i) k_{x_i} \tag{3.6}$$

which is known as the *Nyström extension* since it extends the eigenfunction ϕ_ℓ from its values on the data set $\phi_\ell(x_i)$ to a function on all of A. Since the functions ϕ_ℓ form an orthonormal basis we can introduce the inner products $\hat{\mathbf{f}}_\ell \equiv \langle f, \phi_\ell \rangle_{\mathscr{H}}$. The notation $\hat{\mathbf{f}}_\ell$ is chosen to remind the reader of the Fourier coefficients of a function and the vector $\hat{\mathbf{f}}$ can be computed by



(a) (b) (c)

Fig. 3.4: (a) The function $f \circ h(\theta) = \sin(4\theta)$ (black) on S^1 along with the noisy samples $\mathbf{f}_i^s$ (red) and the denoised samples $\mathbf{f}_i^d$. (b) The coefficients $\hat{\mathbf{f}}_j^s$ (black, solid) of the noisy samples for $j = 1, \ldots, 25$ and the three standard deviation noise floor $3\sigma/\lambda_j$ (red, dashed). (c) The errors $|\mathbf{f}_i^s - f(x_i)|$ for the noisy samples (red) and $|\mathbf{f}_i^d - f(x_i)|$ for the denoised samples.

$$\hat{\mathbf{f}}_\ell \equiv \langle f, \phi_\ell \rangle_{\mathscr{H}} = \mathbf{f}^\top \tilde{\phi}_\ell = \frac{1}{\lambda_\ell} \sum_{i=1}^{N} \phi_\ell(x_i) f(x_i).$$

Since the eigenfunctions ϕ_ℓ are an orthonormal basis for $\mathscr{H}$ we have the representation

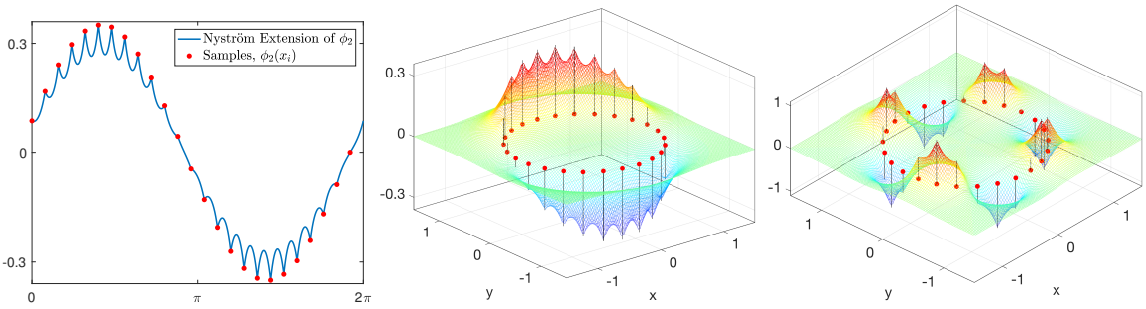
$$f = \sum_{\ell=1}^{N} \hat{\mathbf{f}}_\ell \phi_\ell = \sum_{j,\ell=1}^{N} \hat{\mathbf{f}}_\ell \frac{1}{\lambda_\ell} \phi_\ell(x_j) k_{x_j} = \sum_{i,j,\ell=1}^{N} f(x_i) \frac{1}{\lambda_\ell^2} \phi_\ell(x_i) \phi_\ell(x_j) k_{x_j}$$

which is the Nyström extension of the function f based on its values $f(x_i)$ on the data set.

The kernel functions can also be represented in the basis ϕ_ℓ as

$$k_{x_i} = \sum_{\ell=1}^{N} \langle k_{x_i}, \phi_\ell \rangle_{\mathscr{H}} \phi_\ell = \sum_{\ell=1}^{N} \phi_\ell(x_i) \phi_\ell$$

so that $k(x_i, x_j) = \sum_{\ell=1}^{N} \phi_\ell(x_i) \phi_\ell(x_j)$. The reproducing kernel can also be represented as $k_x(y) = \sum_{\ell=1}^{N} \phi_\ell(x) \phi_\ell(y)$ but notice that this is not necessarily equal to the kernel function $k(x, y)$ when x, y are not in the data set.



(a) (b) (c)

Fig. 3.5: (a) We show the Nyström extension of the first nontrivial eigenfunction ϕ_2 to the unit circle (blue) along with the values of $\phi_2(x_i)$ on the data set highlighted (red points). (b) We further extend ϕ_2 to $[-1.5, 1.5]^2 \subset \mathbb{R}^2$ using the Nyström extension. (c) We extend the function defined by the coefficients $\tilde{\mathbf{f}}^d$ from example 3.5 to $[-1.5, 1.5]^2$.

Example 3.6 [Interpolating functions] Using the data set $\{x_i\}_{i=1}^{25}$ with $N = 25$ from example 3.2 and the grid points $\{y_j\} \subset [-1.5, 1.5]^2$ we apply the Nyström extension to extend functions from the data set to the grid points.

Using the $N \times 1$ coefficients $\tilde{\phi}_2$ contained in the second column of $\tilde{\Phi}$ along with the matrix $\mathbf{K}^{yx}_{ji} = k(y_j, x_i)$ we compute the Nyström extension $\tilde{\phi}_2(y_j) = \left(\mathbf{K}^{yx}\tilde{\phi}_2\right)_j$. To define the Nyström extension of the samples $\mathbf{f}^d$ we use the coefficients $\hat{\mathbf{f}}^d$ to compute $f^d(y_j) = \left(\mathbf{K}^{yx}\Phi\Lambda^{-1}\hat{\mathbf{f}}^d\right)_j$. These extensions are shown in Fig. 3.5. Notice that the Nyström extensions with this kernel function decays quickly away from the unit circle. While this may seem to be due to the exponential decay of the kernel function, we will see in Chapter 5 that it can also be attributed to failure to consider the density of the sample points.

We now consider truncated representations of a function f using only m of the N coefficients from $\hat{f}$ so that f is approximated by $\tilde{f} \equiv \sum_{\ell=1}^{m} \hat{\mathbf{f}}_\ell \phi_\ell$. The following theorem shows that $\tilde{f}$ is the best approximation of f in $\mathscr{H}$ and bounds the relative error between the two.

Theorem 3.9 [Optimal Approximation in $\mathscr{H}$] Consider approximating functions $f \in \mathscr{H}$ by their projection $\tilde{f} = \Pi_{\tilde{\mathscr{H}}} f$ onto an m-dimensional subspace $\tilde{\mathscr{H}} \subset \mathscr{H}$, then the worst-case approximation errors

$$\max_{f \in \mathscr{H} - \{0\}} \left\{ \frac{\sum_{i=1}^{N} (f(x_i) - \tilde{f}(x_i))^2}{\sum_{i=1}^{N} f(x_i)^2} \right\} \quad \text{and} \quad \max_{f \in \mathscr{H} - \{0\}} \frac{||f - \tilde{f}||^2_{\mathscr{H}}}{||f||^2_{\mathscr{H}}}$$

are simultaneously minimized by choosing the subspace $\tilde{\mathscr{H}} = \text{span}(\{\phi_1, ..., \phi_m\})$ and with this choice they are bounded above by λ^2_{m+1} and λ_{m+1} respectively.

Proof Let $u_{m+1}, ..., u_N$ a basis for the orthogonal complement of $\tilde{\mathscr{H}}$ and let U be the $N \times (N-m)$ matrix whose ℓ-th column is $\tilde{\mathbf{u}}_{m+\ell}$. By the relation $\sum_{i=1}^{N} f(x_i)g(x_i) = \langle f, Kg \rangle_{\mathscr{H}}$ we have

$$\sum_{i=1}^{N} (f(x_i) - \tilde{f}(x_i))^2 = \left\langle f - \tilde{f}, K(f - \tilde{f}) \right\rangle_{\mathscr{H}} = \sum_{\ell,s=m+1}^{N} \langle f, u_\ell \rangle_{\mathscr{H}} \langle f, u_s \rangle_{\mathscr{H}} \langle u_\ell, Ku_s \rangle_{\mathscr{H}} = \mathbf{f}^\top \mathbf{U}\mathbf{U}^\top \mathbf{K}\mathbf{U}\mathbf{U}^\top \mathbf{f}$$

where the third equality follows from $f - \tilde{f} = \sum_{\ell=m+1}^{N} \langle f, u_\ell \rangle_{\mathscr{H}} u_\ell$. Since $\sum_{i=1}^{N} f(x_i)^2 = \mathbf{f}^\top \mathbf{f}$ we find that the maximum relative error is

$$\max_{f \in \mathscr{H} - \{0\}} \left\{ \frac{\sum_{i=1}^{N} (f(x_i) - \tilde{f}(x_i))^2}{\sum_{i=1}^{N} f(x_i)^2} \right\} = \max_{\mathbf{f} \in \mathbb{R}^N - \{0\}} \left\{ \frac{\mathbf{f}^\top \mathbf{U}\mathbf{U}^\top \mathbf{K}\mathbf{U}\mathbf{U}^\top \mathbf{f}}{\mathbf{f}^\top \mathbf{f}} \right\}$$

and the maximum is achieved when $\mathbf{f}$ is the eigenvector of $\mathbf{U}\mathbf{U}^\top \mathbf{K}\mathbf{U}\mathbf{U}^\top$ with the largest eigenvalue. Since $\mathbf{K} = \mathbf{V}\Lambda\mathbf{V}^\top$ we find that

$$\mathbf{U}\mathbf{U}^\top \mathbf{K}\mathbf{U}\mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top \mathbf{V}\Lambda^2 \mathbf{V}^\top \mathbf{U}\mathbf{U}^\top$$

so that the columns of $\mathbf{V}^\top \mathbf{U}\mathbf{U}^\top$ are the eigenvectors of $\mathbf{U}\mathbf{U}^\top \mathbf{K}\mathbf{U}\mathbf{U}^\top$. Since $\mathbf{V}^\top \mathbf{U}\mathbf{U}^\top$ has $N - m$ columns, the associated eigenvalues will be $N - m$ of the diagonal entries of Λ^2. In order the minimize the maximum relative error, we must choose the columns of U to span the same space as the $N - m$ eigenvectors of $\mathbf{K}$ which have the smallest eigenvalues, meaning

$$\text{span}(\{\tilde{\mathbf{u}}, ..., \tilde{\mathbf{u}}_N\}) = \text{span}(\{\mathbf{v}_{m+1}, ..., \mathbf{v}_N\}).$$

The maximum relative error is then minimized when f is projected onto the orthogonal complement of this space, which is $\text{span}(\{\phi_1, ..., \phi_m\})$. With this choice of projection, the maximum error is the largest eigenvalue associated to an eigenfunction from the orthogonal complement, which is λ^2_{m+1}. Similarly, for the second form of approximation error, we have

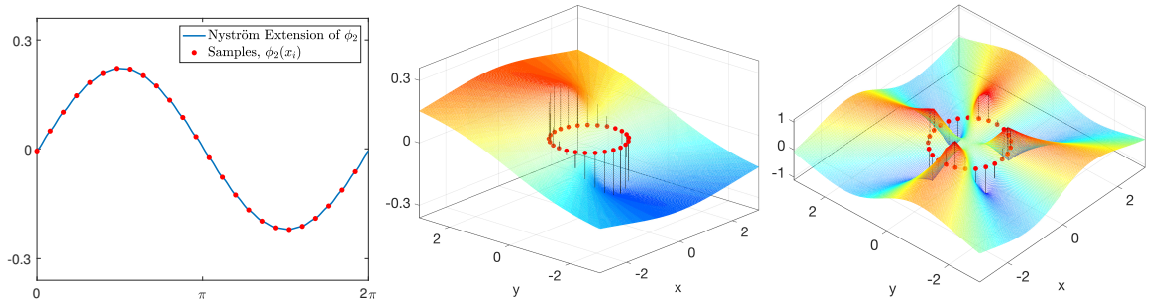
$$||f - \tilde{f}||^2_{\mathscr{H}} = \sum_{\ell=m+1}^{N} \langle f, u_\ell \rangle^2_{\mathscr{H}} = \sum_{\ell=m+1}^{N} \tilde{\mathbf{f}}^\top \mathbf{K}\tilde{\mathbf{u}}_\ell \tilde{\mathbf{u}}_\ell^\top \mathbf{K}\tilde{\mathbf{f}} = \tilde{\mathbf{f}}^\top \mathbf{V}\Lambda\mathbf{V}^\top \mathbf{U}\mathbf{U}^\top \mathbf{V}\Lambda\mathbf{V}^\top \tilde{\mathbf{f}} = \tilde{\mathbf{g}}^\top \Lambda^{1/2} \mathbf{V}^\top \mathbf{U}\mathbf{U}^\top \mathbf{V}\Lambda^{1/2} \tilde{\mathbf{g}}$$

where $\tilde{\mathbf{g}} \equiv \Lambda^{1/2}\mathbf{V}^\top\tilde{\mathbf{f}}$ so that $||f||^2_{\mathscr{H}} = \tilde{\mathbf{f}}^\top\mathbf{K}\tilde{\mathbf{f}} = \tilde{\mathbf{g}}^\top\tilde{\mathbf{g}}$. The maximum relative error is then obtained when $\tilde{\mathbf{g}}$ is the eigenvector of $\Lambda^{1/2}\mathbf{V}^\top\mathbf{U}\mathbf{U}^\top\mathbf{V}\Lambda^{1/2}$ with the largest eigenvalue, and this is minimized when $u_\ell = \phi_\ell$ as above and the corresponding maximum error is λ_{m+1}. □

So far we have considered the practical implications of using a kernel to extend and interpolate functions based on a finite data set, leading the the theory of the finite RKHS. However, while all data sets are necessarily finite, it is still necessary to consider the limit as the number of data points approaches infinity. For example, if we replace the kernel k used in the examples above with the following normalized kernel

$$\bar{k}(x,y) = \frac{k(x,y)}{\sum_{i,j=1}^N k(x,x_i)k(y,x_j)} \tag{3.7}$$

we obtain much smoother and more natural Nyström extensions as shown in Fig. 3.6 (compare to Fig. 3.5 which differs only in the kernel used).



(a) (b) (c)

Fig. 3.6: Using the data set from Ex. 3.2 and the normalized kernel function $\bar{k}$ from (3.7), we show: (a) The Nyström extension of the first nontrivial eigenfunction ϕ_2 to the unit circle (blue) along with the values of $\phi_2(x_i)$ on the data set highlighted (red points). (b) We further extend ϕ_2 to $[-3,3]^2 \subset \mathbb{R}^2$ using the Nyström extension. (c) We extend the function defined by the coefficients $\tilde{\mathbf{f}}^d$ from example 3.5 to $[-3,3]^2$. Notice that the normalized kernel gives smooth extensions compared to Ex. 3.6

The difference between the kernels k and $\bar{k}$ is best understood by examining their behavior in the large data limit. We will demonstrate this limit in Chapter 5 but we first need to introduce some basic concepts from differential geometry in Chapter 4. However, before we turn to these modern geometric results, we first introduce the infinite dimensional counterpart of the finite RKHS theory which is based on Mercer's theorem.

3.4 Mercer's theorem and the limit of large data

We now generalize the theory of finite RKHS to infinite dimensions using a countable data set $\{x_i\}_{i=1}^\infty$ combined with a positive semidefinite kernel function. Recall that the theory of MDS says that for a Gram matrix G there are coordinates $\{x_i\} \subset \mathbb{R}^N$ with $x_i^T x_j = G_{ij}$. The kernel function k is an infinite generalization of a Gram matrix, and in order to find the analog of an eigendecomposition of k we need to consider the integral operator associated with k which is

$$(T_k f)(x) = \int k(x,y)f(y)\,d\mu(y). \tag{3.8}$$

Notice that the integral in the definition of T_k is simply the natural generalization of the summation which occurs when the kernel operator K of a finite RKHS is applied to a function $f(x_i) = \mathbf{f}_i$ as an operator

$$K(f)(x_i) = \mathbf{Kf} = \sum_{j=1}^{N} \mathbf{K}_{ij}\mathbf{f}_j \sum_{j=1}^{N} k(x_i,x_j)f(x_j).$$

Moreover, if the data x_i are sampled from the measure μ the Monte-Carlo quadrature formula says that

$$\mathbf{E}\left[\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}k(x,x_j)f(x_j)\right] = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[k(x,x_j)f(x_j)] = \int k(x,y)f(y)\,d\mu(y)$$

assuming that we can exchange the limit and the expectation. For T_k to be well defined we require a A to be a measure space with measure μ. We can now state the infinite dimensional generalization of MDS which is called Mercer's theorem.

Theorem 3.10 (Mercer's theorem) *For a measure space (A,μ) a measurable positive semidefinite kernel function $k \in L^2_{\mu\otimes\mu}(A\times A)$ has representation*

$$k(x,y) = \sum_{\ell=1}^{\infty}\lambda_\ell v_\ell(x)v_\ell(y) \tag{3.9}$$

where $T_k v_\ell = \lambda_\ell v_\ell$ are the eigenfunctions and eigenvalues of T_k and $\lambda_\ell \geq 0$. Setting

$$\phi_\ell = \lambda_\ell^{1/2} v_\ell \qquad \text{and} \qquad \hat{f}_\ell = \lambda_\ell^{-1}\langle f,\phi_\ell\rangle_{L^2(A,\mu)} = \lambda_\ell^{-1/2}\langle f, v_\ell\rangle_{L^2(A,\mu)}$$

these eigenfunctions define the RKHS associated to the kernel k

$$\mathscr{H} = \left\{\sum_{\ell=1}^{\infty}\hat{f}_\ell\phi_\ell : \sum_{\ell=1}^{\infty}\hat{f}_\ell^2 < \infty\right\} = \left\{\sum_{\ell=1}^{\infty}a_\ell v_\ell : \sum_{\ell=1}^{\infty}\frac{a_\ell^2}{\lambda_\ell} < \infty\right\}$$

which is also the domain of $T_k : L^2(A,\mu) \to \mathscr{H}$. The inner product associated to $\mathscr{H}$ is

$$\langle f,g\rangle_{\mathscr{H}} \equiv \left\langle f, T_k^{-1}g\right\rangle_{L^2(A,\mu)} = \sum_{\ell=1}^{\infty}\frac{1}{\lambda_\ell}\langle f,v_\ell\rangle_{L^2(A,\mu)}\langle g,v_\ell\rangle_{L^2(A,\mu)} = \sum_{\ell=1}^{\infty}\hat{f}_\ell\hat{g}_\ell \tag{3.10}$$

so that $\hat{f}_\ell = \langle f,\phi_\ell\rangle_{\mathscr{H}}$.

Notice that the inner product $\langle\cdot,\cdot\rangle_{\mathscr{H}}$ directly generalizes the finite RKHS inner product $\tilde{\mathbf{f}}^\top\mathbf{K}\tilde{\mathbf{g}} = \tilde{\mathbf{f}}^\top\mathbf{K}\mathbf{K}^{-1}\mathbf{K}\tilde{\mathbf{g}} = \mathbf{f}^\top\mathbf{K}^{-1}\mathbf{g}$ which is the natural Euclidean inner product of $\mathbf{f}$ with the inverse operator $\mathbf{K}^{-1}$ applied to $\mathbf{g}$. The decomposition of the kernel in Mercer's theorem is the infinite dimensional generalization of the eigendecomposition of $\mathbf{K} = \mathbf{V}\Lambda\mathbf{V}^\top$ used in the finite RKHS.

One interpretation of Mercer's theorem is related to the embedding interpretation of KPCA and defines the following feature map.

Definition 3.11 The *feature map* Φ associated to a measurable positive semidefinite kernel function is

$$\Phi(x) = (\phi_1(x),\phi_2(x),...,\phi_\ell(x),...) \tag{3.11}$$

where $\phi_\ell = \sqrt{\lambda_\ell}v_\ell$. The feature map takes A to the space of square summable sequences since $\Phi(x)_\ell = \langle\phi_\ell,k_x\rangle_{\mathscr{H}}$ and the range of Φ is called the *feature space*.

The feature map is defined so that $k(x,y) = \sum_{i=1}^{\infty}\phi_i(x)\phi_i(y) = \langle\Phi(x),\Phi(y)\rangle$ which means that the 'MDS-like' coordinates of a kernel function are infinite dimensional. Of course, for any finite data set, we will find a finite dimensional kernel matrix which is also a Gram matrix, and this will give us a finite dimensional approximation of the feature map. This process is called Kernel PCA as introduced in the previous chapter. However, using the eigenfunctions ϕ_ℓ simply to map the space A into the Hilbert space $\mathscr{H}$ does not make use of the full power of these eigenfunctions to interpolate, denoise, and approximate functions on A as demonstrated above.

The RKHS implied by Mercer's theorem is ultimately insufficient to describe the limiting behavior of kernels such as $\overline{k}$ introduced in (3.7) which adapt to the data set and change as the size of the data set increases. In fact, even the prototypical kernel from Ex. 3.1 changes since the bandwidth parameter δ typically decreases as the size of the data set increases. Understanding the limiting behavior of these changing kernels requires asymptotic analysis of the kernel function with respect to the bandwidth parameter, and this is carried out in Chapter 5. For data lying near an embedded manifold, the analysis will reveal that local kernels are intimately connected with the Laplace-Beltrami operator, which is described in Chapter 4. The following examples will motivate this connection.

3.4.1 Mercer's theorem examples

In general Mercer's theorem is difficult to apply to non-compact domains since even kernels with fast decay may not be integrable on $A \times A$.

Example 3.7 [Heat kernel on $\mathbb{R}$ is not a Mercer kernel] Consider the heat kernel

$$k(t,x,y) = \frac{1}{\sqrt{4\pi t}} \exp(-(x-y)^2/(4t))$$

defined on the real line. Notice that $\int\int k(t,x,y)^2\,dx\,dy = \infty$ so this kernel does not satisfy Mercer's theorem.

We now turn to a compact domain and an example which will motivate the kernels studied in the remainder of the book.

Example 3.8 [Heat kernel on S^1] Consider the heat kernel

$$k(t,\theta,\tau) = \frac{1}{c\sqrt{t}} \exp(-\min\{|\theta-\tau|, 2\pi - |\theta-\tau|\}^2/(4t))$$

defined on the unit circle parameterized by $\theta \in [0,2\pi)$ with $c = \frac{1}{\sqrt{t}} \int_{-\pi}^{\pi} e^{-u^2/4t} du$.

The heat kernel is the solution of the heat equation $\frac{\partial k}{\partial t} = \Delta k$ where $\Delta = \frac{\partial^2}{\partial \theta^2}$ is the Laplacian operator on the circle. The eigenfunctions of the associated integral operator are the sine and cosine functions

$$v_{2\ell}(\theta) = \sin(\ell\theta) \qquad v_{2\ell+1}(\theta) = \cos(\ell\theta).$$

since

$$\begin{aligned} T_k v_{2\ell}(\theta) &= \int_0^{2\pi} k(t,\theta,\tau) v_{2\ell}(\tau)\,d\tau = \frac{1}{c\sqrt{t}} \int_{\theta-\pi}^{\theta+\pi} e^{-\frac{|\theta-\tau|^2}{4t}} \sin(\ell\tau)\,d\tau = \frac{1}{c\sqrt{t}} \int_{-\pi}^{\pi} e^{-\frac{u^2}{4t}} \sin(\ell(u+\theta))\,du \\ &= \frac{1}{c\sqrt{t}} \int_{-\pi}^{\pi} e^{-\frac{u^2}{4t}} (\sin(\ell u)\cos(\ell\theta) + \cos(\ell u)\sin(\ell\theta))\,du = \frac{\sin(\ell\theta)}{c\sqrt{t}} \int_{-\pi}^{\pi} e^{-\frac{u^2}{4t}} \cos(\ell u)\,du \\ &= \frac{\sin(\ell\theta)}{c\sqrt{t}} \int_{-\pi}^{\pi} \mathrm{Re}\left(e^{-\frac{u^2}{4t} + i\ell u}\right) du = e^{-t\ell^2} \sin(\ell\theta) = e^{-t\ell^2} v_{2\ell}(\theta) \end{aligned} \tag{3.12}$$

(where the last steps follow from completing the square inside the exponential, integrating the resulting Gaussian and then taking the real part). Similarly $T_k v_{2\ell+1}(\theta) = e^{-t\ell^2} v_{2\ell+1}(\theta)$ so the eigenvalues are $\lambda_{2\ell} = \lambda_{2\ell+1} = e^{-t\ell^2}$. The RKHS associated to the heat kernel on the circle is then

$$\mathscr{H} = \left\{ \sum_{\ell=0}^{\infty} a_\ell v_{2\ell} + b_\ell v_{2\ell+1} : \sum_{\ell=0}^{\infty} (a_\ell^2 + b_\ell^2) e^{t\ell^2} < \infty \right\}$$

and since this requires the coefficients a_ℓ, b_ℓ to have exponential decay, we find that every function in $\mathscr{H}$ is infinitely differentiable. Notice that in the limit $t \to 0$ the kernel approaches the delta function $k(t,\theta,\tau) \to \delta(\theta-\tau)$ and the associated RKHS approaches $L^2(S^1)$.

Classical Fourier analysis shows that the eigenfunction v_i form an orthonormal basis for $L^2(S^1)$, but what makes this basis so useful in applications? The answer is that the Fourier basis is the smoothest possible orthonormal basis on the unit circle, and this fact is the key to generalizing the Fourier basis to general nonlinear spaces.

Lemma 3.12 The i-th Fourier function v_i minimizes the 'roughness' functional

$$R(f) \equiv \left|\left|\frac{df}{d\theta}\right|\right|^2_{L^2(S^1)} = \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{df}{d\theta}\right)^2 d\theta$$

among all functions f which are orthogonal to the first $i-1$ Fourier functions and with $||f||_{L^2(S^1)} = 1$.

Proof Using integration by parts we can rewrite

$$R(f) = \left|\left|\frac{df}{d\theta}\right|\right|^2_{L^2(S^1)} = \left\langle \frac{df}{d\theta}, \frac{df}{d\theta}\right\rangle_{L^2(S^1)} = \left\langle f, -\frac{d^2f}{d\theta^2}\right\rangle_{L^2(S^1)} = \langle f, -\Delta f\rangle_{L^2(S^1)}$$

which shows that the operator $\Delta f \equiv \frac{d^2f}{d\theta^2}$ is a symmetric negative semidefinite linear operator. Notice that $\Delta v_i = \xi_i v_j$ and $\xi_i = -\text{ceil}(i/2)^2$. Writing $f = \sum_{i=0}^\infty a_i v_i$ and using the fact that $\{v_i\}$ is an orthonormal basis for $L^2(\mathscr{M})$ we find

$$R(f) = -\sum_{i=0}^{\infty} a_i^2 \xi_i.$$

Since $1 = ||f||_{L^2(S^1)} = \sum_{i=0}^\infty a_i^2$ and since the eigenvalues $\xi_i = -\text{ceil}(i/2)^2$ decrease monotonically, the minimizer of $R(f)$ has $a_0 = 1$ and all other coefficients equal to zero. This shows that v_0 is the minimizer of R with $R(v_0) = \xi_0^2$. In order to find an orthonormal basis which minimizes R, we require the next minimizer to be orthogonal to ξ_0 which implies that $a_0 = 0$. Again, by monotonicity of the eigenvalues ξ_i the second minimizer must have $a_1 = 1$ and all other coefficients equal to zero, which shows the second minimizer must be v_1. Continuing in this fashion, the i-th minimizer must be orthogonal to the previous $i-1$ minimizers, and by monotonicity of the eigenvalues must be v_j. □

The key to the proof of Lemma 3.12 was finding the eigenfunctions of the operator $\Delta f = \frac{d^2f}{d\theta^2}$ which captured the notion of roughness on the manifold. The property which makes the Fourier functions so useful is that they are eigenfunctions of this operator, which implies they are the smoothest functions on the circle. To find a generalized version of the Fourier basis on an arbitrary manifold, we will need to generalize the operator Δf. This generalization is called the Laplace-Beltrami operator and it is introduced in the next chapter.

3.5 Symmetric positive definite kernels

In this section we collect facts about symmetric positive definite matrices and kernels. As an example, Fact 5 establishes that the prototypical kernel introduced in Section 3.1 is a symmetric positive semidefinite kernel.

Although we state most of the facts in the context of matrices, the reader can convert most of the statements to kernel functions, and use the same proofs with appropriate changes.

Definition 3.13 The Hadamard product of two $m \times n$ matrices $A = \{A_{ij}\}$ and $B = \{B_{ij}\}$ is the matrix $A \circ B$ with ijth entry equal to $A_{ij}B_{ij}$.

Fact 1. (Schur product theorem.) If A and B are symmetric positive definite (resp., symmetric positive semidefinite) $n \times n$ matrices, then $A \circ B$ is positive definite (resp.,, positive semidefinite).

Proof. Let $A = \sum_{i=1}^{n} \mu_i a_i a_i^T$ and $B = \sum_{j=1}^{n} \nu_j b_j b_j^T$ be the eigenvector decompositions of A and B, where $\mu_i, \nu_j > 0$ by positive definiteness. Then

$$A \circ B = \sum_{i,j} \mu_i \nu_j (a_i a_i^T) \circ (b_j b_j^T) = \sum_{i,j} \mu_i \nu_j (a_i \circ b_j)(a_i \circ b_j)^T$$

and for $x \in \mathbb{R}^n$,

$$x^T (A \circ B) x = \sum_{i,j} \mu_i \nu_j x^T (a_i \circ b_j)(a_i \circ b_j)^T x = \sum_{i,j} \mu_i \nu_j [x^T (a_i \circ b_j)]^2 \geq 0.$$

This proves positive semi-definiteness. To show $A \circ B$ is positive definite, we want to show that if $x \neq 0$, then at least one of the $x^T (a_i \circ b_j)$ terms is nonzero.

Since B is positive-definite, the eigenvector expansion implies $\sum_j \nu_j x^T b_j \cdot b_j^T x = x^T B x > 0$, so there exists a j such that $b_j^T x \neq 0$, which implies $b_j \circ x \neq 0$. Since A is positive definite, $\sum_i \mu_i (b_j \circ x)^T a_i \cdot a_i^T (b_j \circ x) = \sum_i \mu_i (b_j \circ x)^T A (b_j \circ x) > 0$, so there exist an i such that $x^T (a_i \circ b_j) = a_i^T (b_j \circ x) \neq 0$. □

Fact 2. Let $A_1, A_2, A_3, \ldots$ be a sequence of positive definite (resp., positive semidefinite) $n \times n$ matrices, and $c_i > 0$. If the series $\sum_i c_i A_i$ converges to the matrix A, then A is positive definite (resp., positive semidefinite).

Fact 3. If A is a positive definite matrix (resp., positive semidefinite), then $\exp^\circ(A) \equiv \sum_i A^{\circ i}/i!$ is positive definite (resp., positive semidefinite). (Here $A^{\circ i}$ denotes the i-fold Hadamard product $A \circ \cdots \circ A$. This differs from the usual matrix exponential.)

Proof. Each term in the infinite sum is positive definite (resp., positive semidefinite) by Fact 1. The limiting matrix has the same property, by Fact 2. □

Fact 4. If A is a symmetric positive definite $n \times n$ matrix and B is a nonsingular $n \times n$ matrix, then $B^T AB$ and $\exp^\circ(B^T AB)$ are symmetric positive definite. If A is positive semidefinite and B is $n \times N$, then $B^T AB$ and $\exp^\circ(B^T AB)$ are symmetric positive semidefinite.

Proof. If $x \neq 0$ and B is a nonsingular square matrix, then $Bx \neq 0$, and $x^T B^T ABx = (Bx)^T ABx > 0$. Therefore, $B^T AB$ is symmetric positive definite, and Fact 3 implies $\exp^\circ(B^T AB)$ is also. If B is not square, $x^T B^T ABx = (Bx)^T ABx \geq 0$ and so $B^T AB$ and $\exp^\circ(B^T AB)$ are symmetric positive semidefinite. □

Fact 5. Let $\delta > 0$ and let K be the $N \times N$ matrix formed by $K_{ij} = k(x_i, x_j)$ where $k(x,y) = \exp\left(-\frac{||x-y||^2}{2\delta}\right)$ is the prototypical kernel and $\{x_i\}_{i=1}^N \subset \mathbb{R}^n$. Then K is symmetric positive semidefinite.

Proof. Let A be the $n \times N$ data matrix $[x_1, \ldots, x_N]/\sqrt{\delta}$ and let B be the diagonal matrix with entries $\exp^\circ(-||x_1||^2/(2\delta)), \ldots, \exp(-||x_N||^2/(2\delta))$. Since $||x_i - x_j||^2 = ||x_i||^2 - 2x_i^T x_j + ||x_j||^2$, $K = B \exp^\circ(A^T A/\delta) B$. Thus K is positive semidefinite by applying Fact 4 first to $\exp(A^T A)$ and then to $B \exp^\circ(A^T A/\delta) B$.