

## CHAPTER BA

---

### Basic approximation theory

---

The approximation of a given function with a Fourier type expansion will be based on the concept of orthogonal projections in an inner product space. Let us first review the abstract setting.

Let  $X$  denote a vector space over the set  $\mathcal{F}$  of real or complex numbers. In most applications  $X$  is the space of  $n$ -tuples of real or complex numbers ( $R_n$  and  $C_n$  resp.), or, as in this text, a suitably defined function space.

In order to judge the quality of an approximation  $PF$  to a given vector  $F$  we need to be able to quantify the “size” of the vector  $F - PF$ . For this task we need a norm.

**Definition BA 1:** Let  $X$  be a vector space. Then a norm  $\| \cdot \|$  on  $X$  is a function from  $X$  to the non-negative real number with the properties that for any  $x, y \in X$  and any number  $\alpha$

- i)  $\|x\| \geq 0$  and  $\|x\| = 0$  if and only if  $x = 0$
- ii)  $\|\alpha x\| = |\alpha| \|x\|$
- iii)  $\|x + y\| \leq \|x\| + \|y\|$ .

The Euclidean length of a vector is a typical, but by no means the only, example of a norm in  $R_n$ . For functions we will usually choose some integral average (i.e. a mean square value) of the function as its size. In general there are many different norms on a vector space.

Our goal will be to find the approximation  $PF$  to the given  $F \in X$  in a well defined subspace  $M \subset X$ . In this text  $M$  will always be a finite dimensional subspace generated by  $N$  linearly independent vectors in  $X$ . We recall:

**Definition BA 2:** The vectors  $\{\varphi_1, \dots, \varphi_N\}$  are linearly independent if

$$(BA.1) \quad \sum_{j=1}^N \alpha_j \varphi_j = 0$$

holds if and only if  $\alpha_i = 0$  for  $i = 1, \dots, N$ .

We note that this definition assures that no one vector  $\varphi_k$  can be expressed as a linear combination of the remaining vectors  $\{\varphi_1, \dots, \varphi_{k-1}, \varphi_{k+1}, \dots, \varphi_N\}$ . We now assume that

$\{\varphi_1, \dots, \varphi_N\}$  is a set of  $N$  linearly independent vectors and that  $M$  is the subspace of all linear combinations of these vectors. In other words,

$$M = \text{span}\{\varphi_1, \dots, \varphi_N\} = \left\{ m : m = \sum_{j=1}^N \alpha_j \varphi_j \right\}.$$

We will now state (loosely) our approximation problem:

**Problem:** Given  $F \in X$  and  $M = \text{span}\{\varphi_1, \dots, \varphi_N\}$ , find the “best” approximation  $PF$  in  $M$  to the vector  $F$ .

Since  $PF \in M$  it must have the form

$$PF = \sum_{j=1}^N \alpha_j \varphi_j,$$

where, of course, the  $\{\alpha_j\}$  are as yet unknown. If we define the so-called residual vector

$$(BA.2) \quad r = F - PF$$

then  $r$  is clearly a function of the coefficients  $\{\alpha_1, \dots, \alpha_N\}$ . The best approximation to  $F$  is that linear combination in  $M$  for which  $\|r\|$  is smallest. In other words, we need to find the minimizer  $\{\hat{\alpha}_1, \dots, \hat{\alpha}_N\}$  which minimizes the function  $\|r\|$ . The method for actually minimizing

$$h(\alpha_1, \dots, \alpha_N) = \|r\|$$

depends strongly on the choice of the norm imposed on  $X$ . Many norms are not differentiable because they involve absolute values so that the tools of calculus cannot be applied to minimize  $h$ . We shall avoid these complications by assuming that  $X$  is an inner product space and that the norm is “induced” by the inner product.

**Definition BA 3:** An inner product on a vector space  $X$  is a function  $\langle \cdot, \cdot \rangle$  defined on pairs of vectors  $x, y$  such that

- i)  $\langle x, y \rangle = \overline{\langle y, x \rangle}$
- ii)  $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$  for any real or complex number  $\alpha$
- iii)  $\langle x + w, y \rangle = \langle x, y \rangle + \langle w, y \rangle$  for any  $w \in X$
- iv)  $\langle x, x \rangle \geq 0$  and  $\langle x, x \rangle = 0$  if and only if  $x = 0$ .

Note that if the vector space is complex then generally

$$\langle x, y \rangle = \overline{\langle y, x \rangle} \neq \langle y, x \rangle$$

and

$$\langle x, \alpha y \rangle = \overline{\langle \alpha y, x \rangle} = \overline{\alpha \langle y, x \rangle} = \bar{\alpha} \langle x, y \rangle.$$

For real vectors the conjugation has no effect. The complex dot product on  $C_n$  defined by

$$\langle x, y \rangle = \sum_{j=1}^n x_j \bar{y}_j$$

is one example of an inner product. Again, there are infinitely many choices for an inner product on a vector space.

For most of our applications a real inner product will suffice, but complex Fourier series will require a complex inner product. Since the general approach is unchanged in the complex setting we shall pay attention to the ordering of the vectors and write all formulas consistently for a complex inner product.

It is well known that an inner product defines a norm on the vector space  $X$  if for any  $x$  we write

$$\|x\| = \langle x, x \rangle^{1/2}.$$

We shall now show that in this norm the best approximation  $PF$  to  $F$  is the so-called orthogonal projection of  $F$  onto  $M$ . We recall

**Definition BA 4:** Two vectors  $x, y \in X$  are orthogonal with respect to an inner product  $\langle \cdot, \cdot \rangle$  if

$$\langle x, y \rangle = 0.$$

**Definition BA 5:** The orthogonal projection of a vector  $F \in X$  onto  $M$  is that vector  $PF \in M$  which satisfies

$$F = Pf + r$$

where  $PF$  is chosen such that  $r$  is orthogonal to every vector  $m \in M$ , in short, where  $r$  is orthogonal to  $M$ .

The orthogonal projection is unique, for if

$$F = PF_1 + r_1$$

and

$$F = PF_2 + r_2$$

then

$$\langle PF_1 - PF_2, PF_1 - PF_2 \rangle = \langle r_2 - r_1, PF_1 - PF_2 \rangle = 0$$

because  $PF_1 - PF_2 \in M$  and hence orthogonal to  $r_1$  and  $r_2$ .

In principle,  $PF$  is readily found. By definition  $r$  has to be orthogonal to each  $\varphi_i$ , but if it is orthogonal to each  $\varphi_i$  then it also is orthogonal to every linear combination of them. Hence it is necessary and sufficient that

$$(BA.3) \quad \langle F - Pf, \varphi_i \rangle = \langle r, \varphi_i \rangle = 0 \quad \text{for } i = 1, \dots, N.$$

Since

$$PF = \sum_{j=1}^N \alpha_j \varphi_j$$

equation (BA.3) requires that the coefficients  $\{\alpha_j\}$  must be chosen such

$$\left\langle \sum_{j=1}^N \alpha_j \varphi_j, \varphi_i \right\rangle = \langle F, \varphi_i \rangle, \quad i = 1, \dots, N.$$

These  $N$  equations can be conveniently written in matrix form as

$$(BA.4) \quad \mathcal{N} \vec{\alpha} = \vec{b}$$

where

$$\mathcal{N}_{ij} = \langle \varphi_j, \varphi_i \rangle, \quad b_i = \langle F, \varphi_i \rangle.$$

$\mathcal{N}$  is non-singular so that this linear system has a unique solution. This result follows from the observation that if  $\mathcal{N}$  were singular then there is a non-zero vector  $\vec{\beta} = (\beta_1, \dots, \beta_N)$  such

$$\mathcal{N} \vec{\beta} = 0.$$

But then

$$\left\| \sum_{j=1}^N \beta_j \varphi_j \right\|^2 = \left\langle \sum_{j=1}^N \beta_j \varphi_j, \sum_{i=1}^N \beta_i \varphi_i \right\rangle = \sum_{i=1}^N \bar{\beta}_i \sum_{j=1}^N \langle \varphi_j, \varphi_i \rangle \beta_j = \mathcal{N} \vec{\beta} \cdot \vec{\beta} = 0$$

which contradicts that the vectors  $\{\varphi_i\}$  are linearly independent. Note from

$$\|PF\|^2 = \langle PF - F + F, PF \rangle = \langle F, PF \rangle \leq \|F\| \|PF\|$$

that

$$\|PF\| \leq \|F\|$$

regardless of the dimension of the subspace  $N$ . It is easy to see that the orthogonal projection  $PF$  is indeed a best approximation in  $M$ . Let  $G$  be any other element in  $M$ . Then

$$\begin{aligned} \|F - G\|^2 &= \|F - PF + (PF - G)\|^2 = \langle F - PF + (PF - G), F - PF + (PF - G) \rangle \\ &= \|F - PF\|^2 + \|PF - G\|^2 \end{aligned}$$

because  $PF - G \in M$  and thus is orthogonal to  $F - PF$ . Hence

$$\|F - PF\| \leq \|F - G\|$$

for any  $G \in M$  which shows that the orthogonal projection is a best approximation.

Conversely, let  $G$  be a best approximation to  $F$  in  $M$ . Then for any real  $t$ , arbitrary  $m \in M$  and any scalar  $\alpha$  the element  $G + t\alpha m$  belongs to  $M$  so that function

$$g(t) = \langle (G + t\alpha m) - F, (G + t\alpha m) - F \rangle$$

must have a minimum at  $t = 0$ . This implies that

$$g'(0) = \langle \alpha m, G - F \rangle + \langle G - F, \alpha m \rangle = 0.$$

If we choose  $\alpha \neq 0$  such that  $\langle \alpha m, F - G \rangle$  is real then  $\langle m, F - G \rangle = 0$ . Since  $m$  was arbitrary this implies  $\langle m, F - G \rangle = 0$  for any  $m \in M$ . Since  $F = G + F - G$  and  $F - G$  is orthogonal to  $M$  we see that  $G$  must be the orthogonal projection of  $F$  onto  $M$ . We summarize this discussion in the

**Theorem BA 1.** Let  $M = \text{span}\{\varphi_1, \dots, \varphi_N\}$  be a subspace of the vector space  $X$  with inner product  $\langle \cdot, \cdot \rangle$  and norm  $\| \cdot \| = \langle \cdot, \cdot \rangle^{1/2}$ . Then  $\hat{m} \in M$  is the best approximation to a given  $F \in X$  if and only if  $\hat{m}$  is the orthogonal projection of  $F$  onto  $M$ .

From a computational point of view the ideal situation is given when the  $\{\varphi_i\}$  are mutually orthogonal. Such vectors are necessarily linearly independent. In this case  $\mathcal{N}$  is diagonal and the coefficients of the projection  $PF$  are given explicitly as

$$(BA.5) \quad \alpha_i = \frac{\langle F, \varphi_i \rangle}{\langle \varphi_i, \varphi_i \rangle}.$$

Borrowing from the terminology of Fourier series we shall call  $\alpha_i$  the  $i$ th Fourier coefficient of  $F$  (with respect to the orthogonal set  $\{\varphi_i\}$ ).

Let us now apply the abstract theory to the approximation of functions. We begin with the simple case of functions of one variable.

The vector space  $X$  will be the space of all continuous real or complex valued functions defined on the closed interval  $D = [a, b]$ . We shall denote this space by  $C^0[a, b]$ . On this vector space we define the inner product

$$(BA.6) \quad \langle f, g \rangle = \int_a^b f(x) \overline{g(x)} w(x) dx$$

where  $w$  is a given “weight function” with the property that  $w$  is continuous and positive except at isolated points where it might be zero. (Often the weight function is  $w(x) \equiv 1$ .) This restriction on the weight function guarantees that

$$\langle f, f \rangle > 0$$

for any non-trivial function  $f$ . It is now straightforward to verify that we have indeed an inner product. The norm on  $X$  induced by this inner product is

$$\|f\| = \left[ \int_a^b |f(x)|^2 w(x) dx \right]^{1/2}$$

and represents a weighted and scaled root-mean-square value for the function  $f$ .

Typical subspaces  $M \subset X$  are made up of polynomials or trigonometric functions. For example,

$$M = \text{span}\{1, x, \dots, x^N\}$$

is the subspace of all polynomials of degree  $\leq N$ , while

$$M = \text{span} \left\{ \sin \frac{\pi k(x-a)}{b-a}, \cos \frac{\pi n(x-a)}{b-a} \right\}_{k=1, n=0}^{k=K, n=N}$$

is the  $K(N+1)$  dimensional subspace consisting of truncated Fourier series. Let us now illustrate the computation of an orthogonal projection in the following concrete case. Suppose that

$$D = [-1, 1]$$

$$w(x) = x^2$$

$$F(x) = x^3$$

and

$$M = \text{span}\{x, x^2\}.$$

Then

$$PF = \alpha_1 x + \alpha_2 x^2$$

where

$$\begin{pmatrix} \langle x, x \rangle & \langle x^2, x \rangle \\ \langle x, x^2 \rangle & \langle x^2, x^2 \rangle \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \langle x^3, x \rangle \\ \langle x^3, x^2 \rangle \end{pmatrix}.$$

We find

$$\langle x, x \rangle = \int_{-1}^1 x^2 x^2 dx = 2/5,$$

$$\langle x, x^2 \rangle = 0$$

(i.e. the vectors  $x$  and  $x^2$  happen to be orthogonal in this inner product)

$$\langle x^2, x^2 \rangle = 2/7$$

$$\langle x^3, x \rangle = 2/7$$

$$\langle x^3, x^2 \rangle = 0$$

which yields

$$Px^3 = \frac{5}{7} x.$$

Another common subspace of  $C^0[a, b]$  is

$$M = \text{span} \left\{ \cos \frac{k\pi(x-a)}{b-a} \right\}_{k=0}^K$$

which, as we shall see, is associated with the so-called half-range Fourier expansions. It is straightforward to verify that

$$\int_a^b \cos \frac{k\pi(x-a)}{b-a} \cos \frac{n\pi(x-a)}{b-a} dx = \begin{cases} 0 & n \neq k \\ b-a & n = k = 0 \\ (b-a)/2 & n = k > 0 \end{cases}$$

so that the functions spanning  $M$  are orthogonal with respect to the inner product (BA.6) with  $w(x) = 1$ . Hence the orthogonal projection of a continuous function  $F$  onto  $M$  is given by

$$PF(x) = \sum_{j=1}^N \alpha_j \cos \frac{j\pi(x-a)}{b-a}$$

where

$$\alpha_0 = \frac{\int_a^b F(x) dx}{(b-a)}$$

$$\alpha_j = \frac{\left\langle F(x), \cos \frac{j\pi(x-a)}{b-a} \right\rangle}{(b-a)/2}.$$

It will turn out that the vector space of continuous functions is not large enough for subsequent applications because it does not contain functions like step functions, square waves, and functions which are unbounded. We now turn the process around. We define

$$(BA.7) \quad \|f\| = \left( \int_a^b |f(x)|^2 w(x) dx \right)^{1/2}$$

for the weight functions described above and let  $X$  be the space of all functions for which

$$\|f\| < \infty.$$

This is indeed a vector space for if  $f$  and  $g \in X$  then it follows from

$$2|f(x)||g(x)| \leq (|f(x)|^2 + |g(x)|^2)$$

and

$$(|f(x)| + |g(x)|)^2 \leq 2(|f(x)|^2 + |g(x)|^2)$$

that  $f + g \in X$ . Hence  $X$  is closed under vector addition and, by inspection, under scalar multiplication and thus a vector space. If we now define for  $f, g \in X$

$$\langle f, g \rangle = \int_a^b f(x)\overline{g(x)}w(x)dx$$

then this integral is well defined and has the properties of an inner product.  $X$  becomes an inner product space and (BA.7) is the norm induced by the inner product. We denote this space by

$$L_2[[a, b], w] \quad \text{or} \quad L_2[D, w].$$

If  $w(x) \equiv 1$  then we use the simpler notation

$$L_2[a, b] \quad \text{or} \quad L_2[D].$$

Piecewise continuous bounded functions belong to  $L_2[D, w]$ , but so do some unbounded functions. For example, since

$$\int_0^1 (x^{-1/4})^2 dx = 2$$

we see that  $f(x) = x^{-1/4} \in L_2[0, 1]$ . Henceforth, whenever we say  $f \in L_2[D, w]$  we mean that the inner product is

$$\langle f, g \rangle = \int_D f(x)\overline{g(x)}w(x)dx,$$

that the norm is

$$\|f\| = \langle f, f \rangle^{1/2}$$

and that  $f$  and  $g$  are any functions for which these numbers are well defined. In applications, however, we shall rarely leave the realm of piecewise continuous functions.

Nothing changes if we deal with functions of several variables. Let  $x = (x_1, \dots, x_n)$  and  $D$  a (reasonable) closed and bounded subset of  $R_n$ . We shall imply that  $L_2[D, w]$  has the inner product

$$\langle f, g \rangle = \int_D f(x)g(x)w(x)dx$$

and that  $f \in L_2[D, w]$  if

$$\langle f, f \rangle < \infty.$$

Here

$$dx = dx_1, \dots, dx_n$$

although in applications integrals often are transformed into polar, cylindrical or spherical coordinates. Admissible weight functions are continuous functions with the property that for any continuous function  $f$  not identically zero we have

$$\langle f, f \rangle > 0.$$

Finally we take note that on occasion for a function of several variables we shall consider a subset of them as parameters and compute projections with respect to the remaining variables. For example, let  $F(x, t)$  be a function which for given  $t$  belongs to  $L_2[D, w]$  where  $D$  is an interval on the  $x$ -axis. Let  $M = \text{span}\{\varphi_1(x), \dots, \varphi_N(x)\} \subset L_2[D, w]$ . Then we shall write

$$PF(x, t) = \sum_{n=1}^N \alpha_n(t) \varphi_n(x)$$

where  $t$  is a parameter. If  $F$  depends smoothly on  $t$  then it follows from the computation of the  $\{\alpha_j(t)\}$  with (BA.4) that these coefficients likewise will depend smoothly on the parameter  $t$ .