Spline Methods
Draft

Tom Lyche and Knut Mørken

25th April 2003

Contents

CHAPTER 8

Quasi-interpolation methods

In Chapter 5 we considered a number of methods for computing spline approximations. The starting point for the approximation methods is a data set that is usually discrete and in the form of function values given at a set of abscissas. The methods in Chapter 5 roughly fall into two categories: global methods and local methods. A global method is one where any B-spline coefficient depends on all initial data points, whereas a local method is one where a B-spline coefficient only depends on data points taken from the neighbourhood of the support of the corresponding B-spline. Typical global methods are cubic spline interpolation and least squares approximation, while cubic Hermite interpolation and the Schoenberg variation diminishing spline approximation are popular local methods.

In this chapter we are going to describe a general recipe for developing local spline approximation methods. This will enable us to produce an infinite number of approximation schemes that can be tailored to any special needs that we may have or that our given data set dictates. In principle, the methods are local, but by allowing the area of influence for a given B-spline coefficient to grow, our general recipe may even encompass the global methods in Chapter 5.

The recipe we describe produces approximation methods known under the collective term *quasi-interpolation methods*. Their advantage is their flexibility and their simplicity. There is considerable freedom in the recipe to produce tailor-made approximation schemes for initial data sets with special structure. Quasi-interpolants also allow us to establish important properties of B-splines. In the next chapter we will employ them to study how well a given function can be approximated by splines, and to show that B-splines form a stable basis for splines.

8.1 A general recipe

A spline approximation method consists of two main steps: First the degree and knot vector are determined, and then the B-spline coefficients of the approximation are computed from given data according to some formula. For some methods like spline interpolation and least squares approximation, this formula corresponds to the solution of a linear system of equations. In other cases, like cubic Hermite interpolation and Schoenberg's Variation Diminishing spline approximation, the formula for the coefficients is given directly in terms of given values of the function to be interpolated.

8.1.1 The basic idea

The basic idea behind the construction of quasi-interpolants is very simple. We focus on how to compute the B-spline coefficients of the approximation and assume that the degree and knot vector are known. The procedure depends on two versions of the local support property of B-splines that we know well from earlier chapters: (i) The B-spline B_j is nonzero only within the interval $[t_j, t_{j+d+1}]$, and (ii) on the interval $[t_\mu, t_{\mu+1})$ there are only $d+1$ B-splines in $\mathbb{S}_{d,\boldsymbol{t}}$ that are nonzero so a spline g in $\mathbb{S}_{d,\boldsymbol{t}}$ can be written as $g(x) = \sum_{i=\mu-d}^{\mu} b_i B_i(x)$ when x is restricted to this interval.

Suppose we are to compute an approximation $g = \sum_i c_i B_i$ in $\mathbb{S}_{d,\boldsymbol{t}}$ to a given function f. To compute c_j we can select one knot interval $I = [t_\mu, t_{\mu+1}]$ which is a subinterval of $[t_j, t_{j+d+1}]$. We denote the restriction of f to this interval by f^I and determine an approximation $g^I = \sum_{i=\mu-d}^{\mu} b_i B_i$ to f^I. One of the coefficients of g^I will be b_j and we fix c_j by setting $c_j = b_j$. The whole procedure is then repeated until all the c_i have been determined.

It is important to note the flexibility of this procedure. In choosing the interval I we will in general have the $d+1$ choices $\mu = j, j, \ldots, j+d$ (fewer if there are multiple knots). As we shall see below we do not necessarily have to restrict I to be one knot interval; all that is required is that $I \cap [t_\mu, t_{\mu+d+1}]$ is nonempty. When approximating f^I by g^I we have a vast number of possibilities. We may use interpolation or least squares approximation, or any other approximation method. Suppose we settle for interpolation, then we have complete freedom in choosing the interpolation points within the interval I. In fact, there is so much freedom that we can have no hope of exploring all the possibilities.

It turns out that some of this freedom is only apparent — to produce useful quasi-interpolants we have to enforce certain conditions. With the general setup described above, a useful restriction is that if f^I should happen to be a polynomial of degree d then g^I should reproduce f^I, i.e., in this case we should have $g^I = f^I$. This has the important consequence that if f is a spline in $\mathbb{S}_{d,\boldsymbol{t}}$ then the approximation g will reproduce f exactly (apart from rounding errors in the numerical computations). To see why this is the case, suppose that $f = \sum_i \hat{b}_i B_i$ is a spline in $\mathbb{S}_{d,\boldsymbol{t}}$. Then f^I will be a polynomial that can be written as $f^I = \sum_{i=\mu-d}^{\mu} \hat{b}_i B_i$. Since we have assumed that polynomials will be reproduced we know that $g^I = f^I$ so $\sum_{i=\mu-d}^{\mu} b_i B_i = \sum_{i=\mu-d}^{\mu} \hat{b}_i B_i$, and by the linear independence of the B-splines involved we conclude that $b_i = \hat{b}_i$ for $i = \mu-d, \ldots, \mu$. But then we see that $c_j = b_j = \hat{b}_j$ so g will agree with f. An approximation scheme with the property that $Pf = f$ for all f in a space $\mathbb{S}$ is to *reproduce* the space.

8.1.2 A more detailed description

Hopefully, the basic idea behind the construction of quasi-interpolants became clear above. In this section we describe the construction in some more detail with the generalisations mentioned before. We first write down the general procedure for determining quasi-interpolants and then comment on the different steps afterwards.

Algorithm 8.1 (Construction of quasi-interpolants). *Let the spline space $\mathbb{S}_{d,\boldsymbol{t}}$ of dimension n and the real function f defined on the interval $[t_{d+1}, t_{n+1}]$ be given, and suppose that $\boldsymbol{t}$ is a $d+1$-regular knot vector. To approximate f from the space $\mathbb{S}_{d,\boldsymbol{t}}$ perform the following steps for $j = 1, 2, \ldots, n$:*

1. *Choose a subinterval $I = [t_\mu, t_\nu]$ of $[t_{d+1}, t_{n+1}]$ with the property that $I \cap (t_j, t_{j+d+1})$ is nonempty, and let f^I denote the restriction of f to this interval.*

2. *Choose a local approximation method P^I and determine an approximation g^I to f^I,*

$$g^I = P^I f^I = \sum_{i=\nu-d}^{\mu} b_i B_i, \tag{8.1}$$

   on the interval I.

3. *Set coefficient j of the global approximation Pf to b_j, i.e.,*

$$c_j = b_j.$$

The spline $Pf = \sum_{j=1}^{n} c_j B_j$ will then be an approximation to f.

The coefficient c_j obviously depends on f and this dependence on f is often indicated by using the notation $\lambda_j f$ for c_j. This will be our normal notation in the rest of the chapter.

An important point to note is that the restriction $\mathbb{S}_{d,\boldsymbol{t},I}$ of the spline space $\mathbb{S}_{d,\boldsymbol{t}}$ to the interval I can be written as a linear combination of the B-splines $\{B_i\}_{i=\nu-d}^{\mu}$. These are exactly the B-splines whose support intersect the interior of the interval I, and by construction, one of them must clearly be B_j. This ensures that the coefficient b_j that is needed in step 3 is computed in step 2.

Algorithm 8.1 generalises the simplified procedure in Section 8.1.1 in that I is no longer required to be a single knot interval in $[t_j, t_{j+d+1}]$. This gives us considerably more flexibility in the choice of local approximation methods. Note in particular that the classical global methods are included as special cases since we may choose $I = [t_{d+1}, t_{n+1}]$.

As we mentioned in Section 8.1.1, we do not get good approximation methods for free. If Pf is going to be a decent approximation to f we must make sure that the local methods used in step 2 reproduce polynomials or splines.

Lemma 8.2. *Suppose that all the local methods used in step 2 of Algorithm 8.1 reproduce all polynomials of some degree $d_1 \leq d$. Then the global approximation method P will also reproduce polynomials of degree d_1. If all the local methods reproduce all the splines in $\mathbb{S}_{d,\boldsymbol{t},I}$ then P will reproduce the whole spline space $\mathbb{S}_{d,\boldsymbol{t}}$.*

Proof. The proof of both claims follow just as in the special case in Section 8.1.1, but let us even so go through the proof of the second claim. We want to prove that if all the local methods P^I reproduce the local spline spaces $\mathbb{S}_{d,\boldsymbol{t},I}$ and f is a spline in $\mathbb{S}_{d,\boldsymbol{t}}$, then $Pf = f$. If f is in $\mathbb{S}_{d,\boldsymbol{t}}$ we clearly have $f = \sum_{i=1}^{n} \hat{b}_i B_i$ for appropriate coefficients $(\hat{b}_i)_{i=1}^{n}$, and the restriction of f to I can be represented as $f^I = \sum_{i=nu-d}^{\mu} \hat{b}_i B_i$. Since P^I reproduces $\mathbb{S}_{d,\boldsymbol{t},I}$ we will have $P^I f^I = f^I$ or

$$\sum_{i=\nu-d}^{\mu} b_i B_i = \sum_{i=\nu-d}^{\mu} \hat{b}_i B_i.$$

The linear independence of the B-splines involved over the interval I then allows us to conclude that $b_i = \hat{b}_i$ for all indices i involved in this sum. Since j is one the indices we therefore have $c_j = b_j = \hat{b}_j$. When this holds for all values of j we obviously have $Pf = f$. ∎

The reader should note that if I is a single knot interval, the local spline space $\mathbb{S}_{d,\boldsymbol{t},I}$ reduces to the space of polynomials of degree d. Therefore, when I is a single knot interval, local reproduction of polynomials of degree d leads to global reproduction of the whole spline space.

Why does reproduction of splines or polynomials ensure that P will be a good approximation method? We will study this in some detail in Chapter 9, but as is often the case the basic idea is simple: The functions we want to approximate are usually nice and smooth, like the exponential functions or the trigonometric functions. An important property of polynomials is that they approximate such smooth functions well, although if the interval becomes wide we may need to use polynomials of high degree. A quantitative manifestation of this phenomenon is that if we perform a Taylor expansion of a smooth function, then the error term will be small, at least if the degree is high enough. If our approximation method reproduces polynomials it will pick up the essential behaviour of the Taylor polynomial, while the approximation error will pick up the essence of the error in the Taylor expansion. The approximation method will therefore perform well whenever the error in the Taylor expansion is small. If we reproduce spline functions we can essentially reproduce Taylor expansions on each knot interval as long as the function we approximate has at least the same smoothness as the splines in the spline space we are using. So instead of increasing the polynomial degree because we are approximating over a wide interval, we can keep the spacing in the knot vector small and thereby keep the polynomial degree of the spline low. Another way to view this is that by using splines we can split our function into suitable pieces that each can be approximated well by polynomials of relatively low degree, even though this is not possible for the complete function. By constructing quasi-interpolants as outlined above we obtain approximation methods that actually utilise this approximation power of polynomials on each subinterval. In this way we can produce good approximations even to functions that are only piecewise smooth.

8.2 Some quasi-interpolants

It is high time to try out our new tool for constructing approximation methods. Let us see how some simple methods can be obtained from Algorithm 8.1.

8.2.1 Piecewise linear interpolation

Perhaps the simplest, local approximation method is piecewise linear interpolation. We assume that our n-dimensional spline space $\mathbb{S}_{1,\boldsymbol{t}}$ is given and that $\boldsymbol{t}$ is a 2-regular knot vector. For simplicity we also assume that all the interior knots are simple. The function f is given on the interval $[t_2, t_{n+1}]$. To determine c_j we choose the local interval to be $I = [t_j, t_{j+1}]$. In this case, we have no interior knots in I so $\mathbb{S}_{1,\boldsymbol{t},I}$ is the two dimensional space of linear polynomials. A basis for this space is given by the two linear B-splines B_{j-1} and B_j, restricted to the interval I. A natural candidate for our local approximation method is interpolation at t_j andt_{j+1}. On the interval I, the B-spline B_{j-1} is a straight line with value 1 at t_j and value 0 at t_{j+1}, while B_j is a straight line with value 0 at t_j and value 1 at t_{j+1}. The local interpolant can therefore be written

$$P_1^I f(x) = f(t_j)B_{j-1}(x) + f(t_{j+1})B_j(x).$$

From Algorithm 8.1 we know that the coefficient multiplying B_j is the one that should multiply B_j also in our global approximation, in other words $c_j = \lambda_j f = f(t_{j+1})$. The

global approximation is therefore

$$P_1 f(x) = \sum_{i=1}^{n} f(t_{j+1}) B_j(x).$$

Since a straight line is completely characterised by its value at two points, the local approximation will always give zero error and therefore reproduce all linear polynomials. Then we know from Lemma 8.2 that P_1 will reproduce all splines $\mathbb{S}_{1,\boldsymbol{t}}$.

This may seem like unnecessary formalism in this simple case where the conclusions are almost obvious, but it illustrates how the construction works in a very transparent situation.

8.2.2 A 3-point quadratic quasi-interpolant

In our repertoire of approximation methods, we only have one local, quadratic method, Schoenberg's variation diminishing spline. With the quasi-interpolant construction it is easy to construct alternative, local methods. Our starting point is a quadratic spline space $\mathbb{S}_{2,\boldsymbol{t}}$ based on a 3-regular knot vector with distinct interior knots, and a function f to be approximated by a scheme which we denote P_2. The support of the B-spline B_j is $[t_j, t_{j+3}]$, and we choose our local interval as $I = [t_{j+1}, t_{j+2}]$. Since I is one knot interval, we need a local approximation method that reproduces quadratic polynomials. One such method is interpolation at three distinct points. We therefore choose three distinct points $x_{j,0}$, $x_{j,1}$ and $x_{j,2}$ in I. Some degree of symmetry is always a good guide so we choose

$$x_{j,0} = t_{j+1}, \quad x_{j,1} = \frac{t_{j+1} + t_{j+2}}{2}, \quad x_{j,2} = t_{j+2}.$$

To determine $P_2^I f$ we have to solve the linear system of three equations in the three unknowns b_{j-1}, b_j and b_{j+1} given by

$$P_2^I f(x_{j,k}) = \sum_{i=j-1}^{j+1} b_i B_i(x_{j,k}) = f(x_{j,k}), \quad \text{for } k = 0,\, 1,\, 2.$$

With the aid of a tool like Mathematica we can solve these equations symbolically. The result is that

$$b_j = \frac{1}{2}\big(-f(t_{j+1}) + 4f(t_{j+3/2}) - f(t_{j+2})\big),$$

where $t_{j+3/2} = (t_{j+1} + t_{j+2})/2$. The expressions for b_{j-1} and b_{j+1} are much more complicated and involve the knots t_j and t_{j+3} as well. The simplicity of the expression for b_j stems from the fact that $x_{j,1}$ was chosen as the midpoint between t_{j+1} and t_{j+2}.

The expression for b_j is valid whenever $t_{j+1} < t_{j+2}$ which is not the case for $j = 1$ and $j = n$ since $t_1 = t_2 = t_3$ and $t_{n+1} = t_{n+2} = t_{n+3}$. But from Lemma 2.12 we know that any spline g in $\mathbb{S}_{3,\boldsymbol{t}}$ will interpolate its first and last B-spline coefficient at these points so we simply set $c_1 = f(t_1)$ and $c_n = f(t_{n+1})$.

Having constructed the local interpolants, we have all the ingredients necessary to

construct the quasi-interpolant $P_2 f = \sum_{j=1}^{n} \lambda_j f B_j$, namely

$$\lambda_j f = \begin{cases} f(t_1), & \text{when } j = 1; \\ \frac{1}{2}(-f(x_{j,0}) + 4f(x_{j,1}) - f(x_{j,2}), & \text{when } 1 < j < n; \\ f(t_{n+1}), & \text{when } j = n. \end{cases}$$

Since the local approximation reproduced the local spline space (the space of quadratic polynomials in this case), the complete quasi-interpolant will reproduce the whole spline space $\mathbb{S}_{2,\boldsymbol{t}}$.

8.2.3 A 5-point cubic quasi-interpolant

The most commonly used splines are cubic, so let us construct a cubic quasi-interpolant. We assume that the knot vector is 4-regular and that the interior knots are all distinct. As usual we focus on the coefficient $c_j = \lambda_j f$. It turns out that the choice $I = [t_{j+1}, t_{j+3}]$ is convenient. The local spline space $\mathbb{S}_{3,\boldsymbol{t},I}$ has dimension 5 and is spanned by the (restriction of the) B-splines $\{B_i\}_{i=j-2}^{j+2}$. We want the quasi-interpolant to reproduce the whole spline space and therefore need P^I to reproduce $\mathbb{S}_{3,\boldsymbol{t},I}$. We want to use interpolation as our local approximation method, and we know from Chapter 5 that spline interpolation reproduces the spline space as long as it has a unique solution. The solution is unique if the coefficient matrix of the resulting linear system is nonsingular, and from Theorem 5.18 we know that a B-spline coefficient matrix is nonsingular if and only if its diagonal is positive. Since the dimension of $\mathbb{S}_{3,\boldsymbol{t},I}$ is 5 we need 5 interpolation points. We use the three knots t_{j+1}, t_{j+2} and t_{j+3} and one point from each of the knot intervals in I,

$$x_{j,0} = t_{j+1}, \quad x_{j,1} \in (t_{j+1}, t_{j+2}), \quad x_{j,2} = t_{j+2}, \quad x_{j,3} \in (t_{j+2}, t_{j+3}), \quad x_{j,4} = t_{j+3}.$$

Our local interpolation problem is

$$\sum_{i=j-2}^{j+2} b_i B_i(x_{j,k}) = f(x_{j,k}), \quad \text{for } k = 0, 1, \ldots, 4.$$

In matrix-vector form this becomes

$$\begin{pmatrix} B_{j-2}(x_{j,0}) & B_{j-1}(x_{j,0}) & 0 & 0 & 0 \\ B_{j-2}(x_{j,1}) & B_{j-1}(x_{j,1}) & B_j(x_{j,1}) & B_j(x_{j,1}) & 0 \\ B_{j-2}(x_{j,2}) & B_{j-1}(x_{j,2}) & B_j(x_{j,2}) & B_j(x_{j,2}) & B_j(x_{j,2}) \\ 0 & B_{j-1}(x_{j,3}) & B_j(x_{j,3}) & B_j(x_{j,3}) & B_j(x_{j,3}) \\ 0 & 0 & 0 & B_j(x_{j,4}) & B_j(x_{j,4}) \end{pmatrix} \begin{pmatrix} b_{j-2} \\ b_{j-1} \\ b_j \\ b_{j+1} \\ b_{j+2} \end{pmatrix} = \begin{pmatrix} f(x_{j,0}) \\ f(x_{j,1}) \\ f(x_{j,2}) \\ f(x_{j,3}) \\ f(x_{j,4}) \end{pmatrix}$$

when we insert the matrix entries that are zero. Because of the way we have chosen the interpolation points we see that all the entries on the diagonal of the coefficient matrix will be positive so the matrix is nonsingular. The local problem therefore has a unique solution and will reproduce $\mathbb{S}_{3,\boldsymbol{t},I}$. The expression for $\lambda_j f$ is in general rather complicated, but in the special case where the width of the two knot intervals is equal and $x_{j,2}$ and $x_{j,4}$ are chosen as the midpoints of the two intervals we end up with

$$\lambda_j f = \frac{1}{6}\big(f(t_{j+1}) - 8f(t_{j+3/2}) + 20f(t_{j+2}) - 8f(t_{j+5/2}) + f(t_{j+3})\big)$$

where $t_{j+3/2} = (t_{j+1} + t_{j+2})/2$ and $t_{j+5/2} = (t_{j+2} + t_{j+3})/2$. Unfortunately, this formula is not valid when $j = 1$, 2, $n-1$ or n since then one or both of the knot intervals in I collapse to one point. However, our procedure is sufficiently general to derive alternative formulas for computing the first two coefficients. The first value of j for which the general procedure works is $j = 3$. In this case $I = [t_4, t_6]$ and our interpolation problem involves the B-splines $\{B_i\}_{i=1}^5$. This means that when we solve the local interpolation problem we obtain B-spline coefficients multiplying all of these B-splines, including B_1 and B_2. There is nothing stopping us from using the same interval I for computation of several coefficients, so in addition to obtaining $\lambda_3 f$ from this local interpolant, we also use it as our source for the first two coefficients. In the special case when the interior knots are uniformly distributed and $x_{3,1} = t_{9/2}$ and $x_{3,3} = t_{11/2}$, we find

$$\begin{aligned}\lambda_1 f &= f(t_4),\\ \lambda_2 f &= \frac{1}{18}\big(-5f(t_4) + 40f(t_{9/2}) - 36f(t_5) + 18f(t_{11/2}) - f(t_6)\big).\end{aligned}$$

In general, the second coefficient will be much more complicated, but the first one will not change.

This same procedure can obviously be used to determine values for the last two coefficients, and under the same conditions of uniformly distributed knots and interpolation points we find

$$\begin{aligned}\lambda_{n-1} f &= \frac{1}{18}\big(-f(t_{n-1}) + 18f(t_{n-1/2}) - 36f(t_n) + 40f(t_{n+1/2}) - 5f(t_{n+1})\big),\\ \lambda_n f &= f(t_{n+1}).\end{aligned}$$

8.2.4 Some remarks on the constructions

In all our constructions, we have derived specific formulas for the B-spline coefficients of the quasi-interpolants in terms of the function f to be approximated, which makes it natural to use the notation $c_j = \lambda_j f$. To do this, we had to solve the local linear system of equations symbolically. When the systems are small this can be done quite easily with a computer algebra system like Maple or Mathematica, but the solutions quickly become complicated and useless unless the knots and interpolation points are nicely structured, preferably with uniform spacing. The advantage of solving the equations symbolically is of course that we obtain explicit formulas for the coefficients once and for all and can avoid solving equations when we approximate a particular function.

For general knots, the local systems of equations usually have to be solved numerically, but quasi-interpolants can nevertheless prove very useful. One situation is real-time processing of data. Suppose we are in a situation where data are measured and need to be fitted with a spline in real time. With a global approximation method we would have to recompute the whole spline each time we receive new data. This would be acceptable at the beginning, but as the data set grows, we would not be able to compute the new approximation quickly enough. We could split the approximation into smaller pieces at regular intervals, but quasi-interpolants seem to be a perfect tool for this kind of application. In a real-time application the data will often be measured at fixed time intervals, and as we have seen it is then easy to construct quasi-interpolants with explicit formulas for the coefficients. Even if this is not practicable because the explicit expressions are not

available or become too complicated, we just have to solve a simple, linear set of equations to determine each new coefficient. The important fact is that the size of the system is constant so that we can handle almost arbitrarily large data sets, the only limitation being available storage space.

Another important feature of quasi-interpolants is their flexibility. In our constructions we have assumed that the function we approximate can be evaluated at any point that we need. This may sometimes be the case, but often the function is only partially known by a few discrete, measured values at specific abscissas. The procedure for constructing quasi-interpolants has so much inherent freedom that it can be adapted in a number of ways to virtually any specific situation, whether the whole data set is available a priori or the approximation has to be produced in real-time as the data is generated.

8.3 Quasi-interpolants are linear operators

Now that we have seen some examples of quasi-interpolants, let us examine them from a more general point of view. The basic ingredient of quasi-interpolants is the computation of each B-spline coefficient, and we have have used the notation $c_j = \lambda_j f = \lambda_j(f)$ to indicate that each coefficient depends on f. It is useful to think of λ_j as a 'function' that takes an ordinary function as input and gives a real number as output; such 'functions' are usually called functionals. If we go back and look at our examples, we notice that in each case the dependency of our coefficient functionals on f is quite simple: The function values occur explicitly in the coefficient expressions and are not multiplied or operated on in any way other than being added together and multiplied by real numbers. This is familiar from linear algebra.

Definition 8.3. *In the construction of quasi-interpolants, each B-spline coefficient is computed by evaluating a* linear *functional. A linear functional λ is a mapping from a suitable space of functions $\mathbb{S}$ into the real numbers $\mathbb{R}$ with the property that if f and g are two arbitrary functions in $\mathbb{S}$ and α and β are two real numbers then*

$$\lambda(\alpha f + \beta g) = \alpha\lambda f + \beta\lambda g.$$

Linearity is a necessary property of a functional that is being used to compute B-spline coefficients in the construction of quasi-interpolants. If one of the coefficient functionals are not linear, then the resulting approximation method is not a quasi-interpolant. Linearity of the coefficient functionals leads to linearity of the approximation scheme.

Lemma 8.4. *Any quasi-interpolant P is a linear operator, i.e., for any two admissible functions f and g and any real numbers α and β,*

$$P(\alpha f + \beta g) = \alpha Pf + \beta Pg.$$

Proof. Suppose that the linear coefficient functionals are $(\lambda_j)_{j=1}^n$. Then we have

$$P(\alpha f + \beta g) = \sum_{i=1}^{n} \lambda_j(\alpha f + \beta g)B_i = \alpha \sum_{i=1}^{n} \lambda_j f B_i + \beta \sum_{i=1}^{n} \lambda_j g B_i = \alpha Pf + \beta Pg$$

which demonstrates the linearity of P. ∎

This lemma is simple, but very important since there are so many powerful mathematical tools available to analyse linear operators. In Chapter 9 we are going to see how well a given function can be approximated by splines. We will do this by applying basic tools in the analysis of linear operators to some specific quasi-interpolants.

8.4 Different kinds of linear functionals and their uses

In our examples of quasi-interpolants in Section 8.2 the coefficient functionals were all linear combinations of function values, but there are other functionals that can be useful. In this section we will consider some of these and how they turn up in approximation problems.

8.4.1 Point functionals

Let us start by recording the form of the functionals that we have already encountered. The coefficient functionals in Section 8.2 were all in the form

$$\lambda f = \sum_{i=0}^{\ell} w_i f(x_i) \tag{8.2}$$

for suitable numbers $(w_i)_{i=0}^{\ell}$ and $(x_i)_{i=0}^{\ell}$. Functionals of this kind can be used if a procedure is available to compute values of the function f or if measured values of f at specific points are known. Most of our quasi-interpolants will be of this kind.

Point functionals of this type occur naturally in at least two situations. The first is when the local approximation method is interpolation, as in our examples above. The second is when the local approximation method is discrete least squares approximation. As a simple example, suppose our spline space is $\mathbb{S}_{2,\boldsymbol{t}}$ and that in determining c_j we consider the single knot interval $I = [t_{j+1}, t_{j+2}]$. Suppose also that we have 10 function values at the points $(x_{j,k})_{k=0}^{9}$ in this interval. Since the dimension of $\mathbb{S}_{2,\boldsymbol{t},I}$ is 3, we cannot interpolate all 10 points. The solution is to perform a local least squares approximation and determine the local approximation by minimising the sum of the squares of the errors,

$$\min_{g \in \mathbb{S}_{2,\boldsymbol{t},I}} \sum_{k=0}^{9} \big(g(x_{j,k}) - f(x_{j,k})\big)^2.$$

The result is that c_j will be a linear combination of the 10 function values,

$$c_j = \lambda_j f = \sum_{k=0}^{9} w_{j,k} f(x_{j,k}).$$

8.4.2 Derivative functionals

In addition to function values, we can also compute derivatives of a function at a point. Since differentiation is a linear operator it is easy to check that a functional like $\lambda f = f''(x_i)$ is linear. The most general form of a derivative functional based at a point that we will consider is

$$\lambda f = \sum_{k=0}^{r} w_k f^{(k)}(x)$$

where x is a suitable point in the domain of f. We will construct a quasi-interpolant based on this kind of coefficient functionals in Section 8.6.1. By combining derivative functionals based at different points we obtain

$$\lambda f = \sum_{i=0}^{\ell} \sum_{k=0}^{r_i} w_{i,k} f^{(k)}(x_i)$$

where each r_i is a nonnegative integer. A typical functional of this kind is the divided difference of a function when some of the arguments are repeated. Such functionals are fundamental in interpolation with polynomials. Recall that if the same argument occurs $r+1$ times in a divided difference, this signifies that all derivatives of order 0, 1, ..., r are to be interpolated at the point. Note that the point functionals above are derivative functionals with $r_i = 0$ for all i.

8.4.3 Integral functionals

The final kind of linear functionals that we will consider are based on integration. A typical functional of this kind is

$$\lambda f = \int_a^b f(x)\phi(x)\,dx \tag{8.3}$$

where ϕ is some fixed function. Because of basic properties of integration, it is easy to check that this is a linear functional. Just as with point functionals, we can combine several functionals like the one in (8.3) together,

$$\lambda f = w_0 \int_a^b f(x)\phi_0(x)\,dx + w_1 \int_a^b f(x)\phi_1(x)\,dx + \cdots + w_\ell \int_a^b f(x)\phi_\ell(x)\,dx,$$

where $(w_i)_{i=0}^{\ell}$ are real numbers and $\{\phi_i\}_{i=0}^{\ell}$ are suitable functions. Note that the right-hand side of this equation can be written in the form (8.3) if we define ϕ by

$$\phi(x) = w_0\phi_0(x) + w_1\phi_1(x) + \cdots + w_\ell\phi_\ell(x).$$

Point functionals can be considered a special case of integral functionals. For if ϕ_ϵ is a function that is positive on the interval $I_\epsilon = (x_i - \epsilon, x_i + \epsilon)$ and $\int_{I_\epsilon} \phi_\epsilon = 1$, then we know from the mean value theorem that

$$\int_{I_\epsilon} f(x)\phi_\epsilon(x)\,dx = f(\xi)$$

for some ξ in I_ϵ, as long as f is a nicely behaved (for example continuous) function. If we let ϵ tend to 0 we clearly have

$$\lim_{\epsilon \to 0} \int_{I_\epsilon} f(x)\phi_\epsilon(x)\,dx = f(x_i), \tag{8.4}$$

so by letting ϕ in (8.3) be a nonnegative function with small support around x and unit integral we can come as close to point interpolation as we wish.

If we include the condition that $\int_a^b \phi\,dx = 1$, then the natural interpretation of (8.3) is that λf gives a weighted average of the function f, with $\phi(x)$ giving the weight of the

function value $f(x)$. A special case of this is when ϕ is the constant $\phi(x) = 1/(b-a)$; then λf is the traditional average of f. From this point of view the limit (8.4) is quite obvious: if we take the average of f over ever smaller intervals around x_i, the limit must be $f(x_i)$.

The functional $\int_a^b f(x)\,dx$ is often referred to as the *first moment* of f. As the name suggests there are more moments. The $i+1$st moment of f is given by

$$\int_a^b f(x)x^i\,dx.$$

Moments of a function occur in many applications of mathematics like physics and the theory of probability.

8.4.4 Preservation of moments and interpolation of linear functionals

Interpolation of function values is a popular approximation method, and we have used it repeatedly in this book. However, is it a good way to approximate a given function f? Is it not a bit haphazard to pick out a few, rather arbitrary, points on the graph of f and insist that our approximation should reproduce these points exactly and then ignore all other information about f? As an example of what can happen, suppose that we are given a set of function values $\big(x_i, f(x_i)\big)_{i=1}^m$ and that we use piecewise linear interpolation to approximate the underlying function. If f has been sampled densely and we interpolate all the values, we would expect the approximation to be good, but consider what happens if we interpolate only two of the values. In this case we cannot expect the resulting straight line to be a good approximation. If we are only allowed to reproduce two pieces of information about f we would generally do much better by reproducing its first two moments, i.e., the two integrals $\int f(x)\,dx$ and $\int f(x)x\,dx$, since this would ensure that the approximation would reproduce some of the *average* behaviour of f.

Reproduction of moments is quite easy to accomplish. If our approximation is g, we just have to ensure that the conditions

$$\int_a^b g(x)x^i\,dx = \int_a^b f(x)x^i\,dx, \quad i = 0,\, 1,\, \ldots,\, n-1$$

are enforced if we want to reproduce n moments. In fact, this can be viewed as a generalisation of interpolation if we view interpolation to be preservation of the values of a set of linear functionals $(\rho_i)_{i=1}^n$,

$$\rho_i g = \rho_i f, \quad \text{for } i = 1,\, 2,\, \ldots,\, n. \tag{8.5}$$

When $\rho_i f = \int_a^b f(x)x^{i-1}\,dx$ for $i = 1, \ldots, n$ we preserve moments, while if $\rho_i f = f(x_i)$ for $i = 1, \ldots, n$ we preserve function values. Suppose for example that g is required to lie in the linear space spanned by the basis $\{\psi_j\}_{j=1}^n$. Then we can determine coefficients $(c_j)_{j=1}^n$ so that $g(x) = \sum_{j=1}^n c_j\psi_j(x)$ satisfies the interpolation conditions (8.5) by inserting this expression for g into (8.5). By exploiting the linearity of the functionals, we end up with the n linear equations

$$c_1\rho_i(\psi_1) + c_2\rho_i(\psi_2) + \cdots + c_n\rho_i(\psi_n) = \rho_i(f), \quad i = 1,\, \ldots,\, n$$

in the n unknown coefficients $(c_i)_{i=1}^n$. In matrix-vector form this becomes

$$\begin{pmatrix} \rho_1(\psi_1) & \rho_1(\psi_2) & \cdots & \rho_1(\psi_n) \\ \rho_2(\psi_1) & \rho_2(\psi_2) & \cdots & \rho_1(\psi_n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_n(\psi_1) & \rho_n(\psi_2) & \cdots & \rho_n(\psi_n) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} \rho_1(f) \\ \rho_2(f) \\ \vdots \\ \rho_n(f) \end{pmatrix}. \tag{8.6}$$

A fundamental property of interpolation by point functionals is that the only polynomial of degree d that interpolates the value 0 at $d+1$ points is the zero polynomial. This corresponds to the fact that when $\rho_i f = f(x_i)$ and $\psi_i(x) = x^i$ for $i = 0, \ldots, d$, the matrix in (8.6) is nonsingular. Similarly, it turns out that the only polynomial of degree d whose $d+1$ first moments vanish is the zero polynomial, which corresponds to the fact that the matrix in (8.6) is nonsingular when $\rho_i f = \int_a^b f(x)x^i\,dx$ and $\psi_i(x) = x^i$ for $i = 0, \ldots, d$.

If the equations (8.6) can be solved, each coefficient will be a linear combination of the entries on the right-hand side,

$$c_j = \lambda_j f = w_{j,1}\rho_1(f) + w_{j,2}\rho_2(f) + \cdots + w_{j,n}\rho_n(f).$$

We recognise this as (8.2) when the ρ_i correspond to point functionals, whereas we have

$$\begin{aligned} c_j = \lambda_j f &= w_{j,1}\int_a^b f(x)\,dx + w_{j,2}\int_a^b f(x)x\,dx + \cdots + w_{j,n}\int_a^b f(x)x^{n-1}\,dx \\ &= \int_a^b f(x)\big(w_{j,1} + w_{j,2}x + \cdots + w_{j,n}x^{n-1}\big)\,dx \end{aligned}$$

when the ρ_i correspond to preservation of moments.

8.4.5 Least squares approximation

In the discussion of point functionals, we mentioned that least squares approximation leads to coefficients that are linear combinations of point functionals when the error is measured by summing up the squares of the errors at a given set of data points. This is naturally termed *discrete* least squares approximation. In *continuous* least squares approximation we minimise the integral of the square of the error. If the function to be approximated is f and the approximation g is required to lie in a linear space $\mathbb{S}$, we solve the minimisation problem

$$\min_{g\in\mathbb{S}} \int_a^b \big(f(x) - g(x)\big)^2\,dx.$$

If $\mathbb{S}$ is spanned by $(\psi_i)_{i=1}^n$, we can write g as $g = \sum_{i=1}^n c_i\psi$ and the minimisation problem becomes

$$\min_{(c_1,\ldots,c_n)\in\mathbb{R}^n} \int_a^b \Big(f(x) - \sum_{i=1}^n c_i\psi(x)\Big)^2\,dx.$$

To determine the minimum we differentiate with respect to each coefficient and set the derivatives to zero which leads to the so-called *normal equations*

$$\sum_{i=1}^n c_i \int_a^b \psi_i(x)\psi_j(x)\,dx = \int_a^b \psi_j(x)f(x)\,dx, \quad \text{for } j = 1, \ldots, n.$$

If we use the notation above and introduce the linear functionals $\rho_i f = \int_a^b \psi_i(x) f(x)$ represented by the basis functions, we recognise this linear system as an instance of (8.6). In other words, least squares approximation is nothing but interpolation of the linear functionals represented by the basis functions. In particular, preservation of moments corresponds to least squares approximation by polynomials.

8.4.6 Computation of integral functionals

In our discussions involving integral functionals we have tacitly assumed that the values of integrals like $\int_a^b f(x)\psi(x)\,dx$ are readily available. This is certainly true if both f and ψ are polynomials, and it turns out that it is also true if both f and ψ are splines. However, if f is some general function, then the integral cannot usually be determined exactly, even when ψ_i is a polynomial. In such situations we have to resort to numerical integration methods. Numerical integration amounts to computing an approximation to an integral by evaluating the function to be integrated at certain points, multiplying the function values by suitable weights, and then adding up to obtain the approximate value of the integral,

$$\int_a^b f(x)\,dx \approx w_0 f(x_0) + w_1 f(x_1) + \cdots + w_\ell f(x_\ell).$$

In other words, when it comes to practical implementation of integral functionals we have to resort to point functionals. In spite of this, integral functionals and continuous least squares approximation are such important concepts that it is well worth while to have an exact mathematical description. And it is important to remember that we do have exact formulas for the integrals of polynomials and splines.

8.5 Alternative ways to construct coefficient functionals

In Section 8.2 we constructed three quasi-interpolants by following the general procedure in Section 8.1. In this section we will deduce two alternative ways to construct quasi-interpolants.

8.5.1 Computation via evaluation of linear functionals

Let us use the 3-point, quadratic quasi-interpolant in subsection 8.2.2 as an example. In this case we used $I = [t_{j+1}, t_{j+2}]$ as the local interval for determining $c_j = \lambda_j f$. This meant that the local spline space $\mathbb{S}_{2,\boldsymbol{t},I}$ become the space of quadratic polynomials on I which has dimension three. This space is spanned by the three B-splines $\{B_i\}_{i=j-1}^{j+1}$ and interpolation at the three points

$$t_{j+1}, \quad t_{j+3/2} = \frac{t_{j+1} + t_{j+2}}{2}, \quad t_{j+2}$$

allowed us to determine a local interpolant $g^I = \sum_{i=j-1}^{j+1} b_i B_i$ whose middle coefficient b_j we used as $\lambda_j f$.

An alternative way to do this is as follows. Since g^I is constructed by interpolation at the three points t_{j+1}, $t_{j+3/2}$ and t_{j+2}, we know that $\lambda_j f$ can be written in the form

$$\lambda_j f = w_1 f(t_{j+1}) + w_2 f(t_{j+3/2}) + w_3 f(t_{j+2}). \tag{8.7}$$

We want to reproduce the local spline space which in this case is just the space of quadratic polynomials. This means that (8.7) should be valid for all quadratic polynomials. Reproduction of quadratic polynomials can be accomplished by demanding that (8.7) should be exact when f is replaced by the three elements of a basis for $\mathbb{S}_{2,\boldsymbol{t},I}$. The natural basis to use in our situation is the B-spline basis $\{B_i\}_{i=j-1}^{j+1}$. Inserting this, we obtain the system

$$\begin{aligned}
\lambda_j B_{j-1} &= w_1 B_{j-1}(t_{j+1}) + w_2 B_{j-1}(t_{j+3/2}) + w_3 B_{j-1}(t_{j+2}),\\
\lambda_j B_j &= w_1 B_j(t_{j+1}) + w_2 B_j(t_{j+3/2}) + w_3 B_j(t_{j+2}),\\
\lambda_j B_{j+1} &= w_1 B_{j+1}(t_{j+1}) + w_2 B_{j+1}(t_{j+3/2}) + w_3 B_{j+1}(t_{j+2}).
\end{aligned}$$

in the three unknowns w_1, w_2 and w_3. The left-hand sides of these equations are easy to determine. Since $\lambda_j f$ denotes the jth B-spline coefficient, it is clear that $\lambda_j B_i = \delta_{i,j}$, i.e., it is 1 when $i = j$ and 0 otherwise.

To determine the right-hand sides we have to compute the values of the B-splines. For this it is useful to note that the w_j's in equation (8.7) cannot involve any of the knots other than t_{j+1} and t_{j+2} since a general polynomial knows nothing about these knots. This means that we can choose the other knots so as to make life simple for ourselves. The easiest option is to choose the first three knots equal to t_{j+1} and the last three equal to t_{j+2}. But then we are in the Bezier setting, and we know that the B-splines in this case will have the same values if we choose $t_{j+1} = 0$ and $t_{j+2} = 1$. The knots are then $(0, 0, 0, 1, 1, 1)$ which means that $t_{j+3/2} = 1/2$. If we denote the B-splines on these knots by $\{\tilde{B}_i\}_{i=1}^3$, we can replace B_i in (8.5.1) by $\tilde{B}_{i-j+2}$ for $i = 1, 2, 3$. We can now simplify (8.5.1) to

$$\begin{aligned}
0 &= w_1 \tilde{B}_1(0) + w_2 \tilde{B}_1(1/2) + w_3 \tilde{B}_1(1),\\
1 &= w_1 \tilde{B}_2(0) + w_2 \tilde{B}_2(1/2) + w_3 \tilde{B}_2(1),\\
0 &= w_1 \tilde{B}_3(0) + w_2 \tilde{B}_3(1/2) + w_3 \tilde{B}_3(1).
\end{aligned}$$

If we insert the values of the B-splines we end up with the system

$$\begin{aligned}
w_1 + w_2/4 &= 0,\\
w_2/2 &= 1,\\
w_2/4 + w_3 &= 0,
\end{aligned}$$

which has the solution $w_1 = -1/2$, $w_2 = 2$ and $w_3 = -1/2$. In conclusion we have

$$\lambda_j f = \frac{-f(t_{j+1}) + 4f(t_{j+3/2}) - f(t_{j+2})}{2},$$

as we found in Section 8.2.2.

This approach to determining the linear functional works quite generally and is often the easiest way to compute the weights (w_i).

8.5.2 Computation via explicit representation of the local approximation

There is a third way to determine the expression for $\lambda_j f$. For this we write down an explicit expression for the approximation g^I. Using the 3-point quadratic quasi-interpolant as our

example again, we introduce the abbreviations $a = t_{j+1}$, $b = t_{j+3/2}$ and $c = t_{j+2}$. We can write the local interpolant g^I as

$$g^I(x) = \frac{(x-b)(x-c)}{(a-b)(a-c)}f(a) + \frac{(x-a)(x-c)}{(b-a)(b-c)}f(b) + \frac{(x-a)(x-b)}{(c-a)(c-b)}f(c),$$

as it is easily verified that g^I then satisfies the three interpolation conditions $g^I(a) = f(a)$, $g^I(b) = f(b)$ and $g^I(c) = f(c)$. What remains is to write this in terms of the B-spline basis $\{B_i\}_{i=j-1}^{j+1}$ and pick out coefficient number j. Recall that we have the notation $\gamma_j(f)$ for the jth B-spline coefficient of a spline f. Coefficient number j on the left-hand side is $\lambda_j f$. On the right, we find the B-spline coefficients of each of the three polynomials and add up. The numerator of the first polynomial is $(x-b)(x-c) = x^2 - (b+c)x + bc$. To find the jth B-spline of this polynomial, we make use of Corollary 3.5 which tells that, when $d = 2$, we have $\gamma_j(x^2) = t_{j+1}t_{j+2} = ac$ and $\gamma_j(x) = (t_{j+1} + t_{j+2})/2 = (a+c)/2 = b$, respectively. The jth B-spline coefficient of the first polynomial is therefore

$$\gamma_j\Big(\frac{ac - (b+c)b + bc}{(a-b)(a-c)}\Big) = \frac{ac - b^2}{(a-b)(a-c)} \tag{8.8}$$

which simplifies to $-1/2$ since $b = (a+c)/2$. Similarly, we find that the jth B-spline coefficient of the second and third polynomials are 2 and $-1/2$, respectively. The complete jth B-spline coefficient of the right-hand side of (8.8) is therefore $-f(a)/2 + 2f(b) - f(c)/2$. In total, we have therefore obtained

$$\lambda_j f = \gamma_j(g^I) = -\frac{f(t_{j+1})}{2} + 2f(t_{j+3/2}) - \frac{f(t_{j+2})}{2},$$

as required.

This general procedure also works generally, and we will see another example of it in Section 8.6.1.

8.6 Two quasi-interpolants based on point functionals

In this section we consider two particular quasi-interpolants that can be constructed for any polynomial degree. They may be useful for practical approximation problems, but we are going to use them to prove special properties of spline functions in Chapters 9 and 10. Both quasi-interpolants are based on point functionals: In the first case all the points are identical which leads to derivative functionals, in the second case all the points are distinct.

8.6.1 A quasi-interpolant based on the Taylor polynomial

A very simple local, polynomial approximation is the Taylor polynomial. This leads to a quasi-interpolant based on derivative functionals. Even though we use splines of degree d, our local approximation can be of lower degree; in Theorem 8.5 this degree is given by r.

Theorem 8.5 (de Boor-Fix). *Let r be an integer with $0 \le r \le d$ and let x_j be a number in $[t_j, t_{j+d+1}]$ for $j = 1, \ldots, n$. Consider the quasi-interpolant*

$$Q_{d,r}f = \sum_{j=1}^{n} \lambda_j(f) B_{j,d}, \quad \text{where} \quad \lambda_j(f) = \frac{1}{d!}\sum_{k=0}^{r} (-1)^k D^{d-k}\rho_{j,d}(x_j) D^k f(x_j), \tag{8.9}$$

and $\rho_{j,d}(y) = (y - t_{j+1}) \cdots (y - t_{j+d})$. *Then* $Q_{d,r}$ *reproduces all polynomials of degree* r *and* $Q_{d,d}$ *reproduces all splines in* $\mathbb{S}_{d,\boldsymbol{t}}$.

Proof. To construct $Q_{d,r}$ we let I be the knot interval that contains x_j and let the local approximation $g^I = P_r^I f$ be the Taylor polynomial of degree r at the point x_j,

$$g^I(x) = P_r^I f(x) = \sum_{k=0}^{r} \frac{(x - x_j)^k}{k!} D^k f(x_j).$$

To construct the linear functional $\lambda_j f$, we have to find the B-spline coefficients of this polynomial. We use the same approach as in Section 8.5.2. For this Marsden's identity,

$$(y - x)^d = \sum_{j=1}^{n} \rho_{j,d}(y) B_{j,d}(x),$$

will be useful. Setting $y = x_j$, we see that the jth B-spline coefficient of $(x_j - x)^d$ is $\rho_{j,d}(x_j)$. Differentiating Marsden's identity $d - k$ times with respect to y, setting $y = x_i$ and rearranging, we obtain the jth B-spline coefficient of $(x - x_j)^k/k!$ as

$$\gamma_j\big((x - x_j)^k/k!\big) = (-1)^k D^{d-k} \rho_{j,d}(x_j)/d! \qquad \text{for } k = 0, \ldots, r.$$

Summing up, we find that

$$\lambda_j(f) = \frac{1}{d!} \sum_{k=0}^{r} (-1)^k D^{d-k} \rho_{j,d}(x_j) D^k f(x_j).$$

Since the Taylor polynomial of degree r reproduces polynomials of degree r, we know that the quasi-interpolant will do the same. If $r = d$, we reproduce polynomials of degree d which agree with the local spline space $\mathbb{S}_{d,\boldsymbol{t},I}$ since I is a single knot interval. The quasi-interpolant therefore reproduces the whole spline space $\mathbb{S}_{d,\boldsymbol{t}}$ in this case. ∎

Example 8.6. We find

$$D^d \rho_{j,d}(y)/d! = 1, \quad D^{d-1} \rho_{j,d}(y)/d! = y - t_j^*, \quad \text{where} \quad t_j^* = \frac{t_{j+1} + \cdots + t_{j+d}}{d}. \tag{8.10}$$

For $r = 1$ and $x_j = t_j^*$ we therefore obtain

$$Q_{d,r} f = \sum_{j=1}^{n} f(t_j^*) B_{j,d}$$

which is the Variation Diminishing spline approximation. For $d = r = 2$ we obtain

$$Q_{2,2} f = \sum_{j=1}^{n} \big[f(x_j) - (x_j - t_{j+3/2}) D f(x_j) + \frac{1}{2} (x_j - t_{j+1})(x_j - t_{j+2}) D^2 f(x_j) \big] B_{j,2}. \tag{8.11}$$

while for $d = r = 3$ and $x_j = t_{j+2}$ we obtain

$$Q_{3,3} f = \sum_{j=1}^{n} \big[f(t_{j+2}) + \frac{1}{3}(t_{j+3} - 2t_{j+2} + t_{j+1}) D f(t_{j+2}) - \frac{1}{6}(t_{j+3} - t_{j+2})(t_{j+2} - t_{j+1}) D^2 f(t_{j+2}) \big] B_{j,3}. \tag{8.12}$$

We leave the detailed derivation as a problem for the reader.

Since $Q_{d,d}f = f$ for all $f \in \mathbb{S}_{d,\boldsymbol{t}}$ it follows that the coefficients of a spline $f = \sum_{j=1}^{n} c_j B_{j,d}$ can be written in the form

$$c_j = \frac{1}{d!}\sum_{k=0}^{d}(-1)^k D^{d-k}\rho_{j,d}(x_j)D^k f(x_j), \qquad \text{for } j = 1, \ldots, n, \tag{8.13}$$

where x_j is any number in $[t_j, t_{j+d+1}]$.

8.6.2 Quasi-interpolants based on evaluation

Another natural class of linear functionals is the one where each λ_j used to define Q is constructed by evaluating the data at $r+1$ distinct points

$$t_j \le x_{j,0} < x_{j,1} < \cdots < x_{j,r} \le t_{j+d+1} \tag{8.14}$$

located in the support $[t_j, t_{j+d+1}]$ of the B-spline $B_{j,d}$ for $j = 1, \ldots, n$. We consider the quasi-interpolant

$$P_{d,r}f = \sum_{j=1}^{n} \lambda_{j,r}(f)B_{j,d}, \tag{8.15}$$

where

$$\lambda_{j,r}(f) = \sum_{k=0}^{r} w_{j,k} f(x_{j,k}). \tag{8.16}$$

From the preceding theory we know how to choose the constants $w_{j,k}$ so that $P_{d,r}f = f$ for all $f \in \pi_r$.

Theorem 8.7. *Let $\mathbb{S}_{d,\boldsymbol{t}}$ be a spline space with a $d+1$-regular knot vector $\boldsymbol{t} = (t_i)_{i=1}^{n+d+1}$. Let $(x_{j,k})_{k=0}^{r}$ be $\ell+1$ distinct points in $[t_j, t_{j+d+1}]$ for $j = 1, \ldots, n$, and let $w_{j,k}$ be the jth B-spline coefficient of the polynomial*

$$p_{j,k}(x) = \prod_{\substack{r=0 \\ r\neq k}}^{r} \frac{x - x_{j,r}}{x_{j,k} - x_{j,r}}.$$

Then $P_{d,r}f = f$ for all $f \in \pi_r$, and if $r = d$ and all the numbers $(x_{j,k})_{k=0}^{r}$ lie in one subinterval

$$t_j \le t_{\ell_j} \le x_{j,0} < x_{j,1} < \cdots < x_{j,r} \le t_{\ell_j+1} \le t_{j+d+1} \tag{8.17}$$

then $P_{d,d}f = f$ for all $f \in \mathbb{S}_{d,\boldsymbol{t}}$.

Proof. It is not hard to see that

$$p_{j,k}(x_{j,i}) = \delta_{k,i}, \quad k, i = 0, \ldots, r$$

so that the polynomial

$$P^I_{d,r}f(x) = \sum_{k=0}^{r} p_{j,k}(x) f(x_{j,k})$$

satisfies the interpolation conditions $P^I_{d,r}f(x_{j,r}) = f(x_{j,r})$ for all j and r. The result therefore follows from the general recipe. ∎

In order to give examples of quasi-interpolants based on evaluation we need to know the B-spline coefficients of the polynomials $p_{j,k}$. We will return to this in more detail in Chapter 9, see (9.14) in the case $r = d$. A similar formula can be given for $r < d$.

Example 8.8. For $r = 1$ we have

$$p_{j,0}(x) = \frac{x_{j,1} - x}{x_{j,1} - x_{j,0}}, \quad p_{j,1}(x) = \frac{x - x_{j,0}}{x_{j,1} - x_{j,0}}$$

and (8.15) takes the form

$$P_{d,1}f = \sum_{j=1}^{n} \left[\frac{x_{j,1} - t_j^*}{x_{j,1} - x_{j,0}} f(x_{j,0}) + \frac{t_j^* - x_{j,0}}{x_{j,1} - x_{j,0}} f(x_{j,1}) \right] B_{j,d}. \tag{8.18}$$

This quasi-interpolant reproduces straight lines for any choice of $t_j \le x_{j,0} < x_{j,1} \le t_{j+d+1}$. If we choose $x_{j,0} = t_j^*$ the method simplifies to

$$\tilde{P}_{d,1}f = \sum_{j=1}^{n} f(t_j^*) B_{j,d}. \tag{8.19}$$

This is again the *Variation diminishing method of Schoenberg.*

Exercises for Chapter 8

8.1 In this exercise we assume that the points $(x_{i,k})$ and the spline space $\mathbb{S}_{d,\boldsymbol{t}}$ are as in Theorem 8.7.

a) Show that for $r = d = 2$

$$\begin{aligned} P_{2,2}f = \sum_{j=1}^{n} \Bigg[& \frac{(t_{j+1} - x_{j,1})(t_{j+2} - x_{j,2}) + (t_{j+2} - x_{j,1})(t_{j+1} - x_{j,2})}{2(x_{j,0} - x_{j,1})(x_{j,0} - x_{j,2})} f(x_{j,0}) \\ & + \frac{(t_{j+1} - x_{j,0})(t_{j+2} - x_{j,2}) + (t_{j+2} - x_{j,0})(t_{j+1} - x_{j,2})}{2(x_{j,1} - x_{j,0})(x_{j,1} - x_{j,2})} f(x_{j,1}) \\ & + \frac{(t_{j+1} - x_{j,0})(t_{j+2} - x_{j,1}) + (t_{j+2} - x_{j,0})(t_{j+1} - x_{j,1})}{2(x_{j,2} - x_{j,0})(x_{j,2} - x_{j,1})} f(x_{j,2}) \Bigg] B_{j,2} \end{aligned} \tag{8.20}$$

b) Show that (8.20) reduces to the operator (9.4) for a suitable choice of $(x_{j,k})_{k=0}^{2}$.

8.2 Derive the following operators $Q_{d,l}$ and show that they are exact for π_r for the indicated r. Again we the points $(x_{j,k})$ and the spline space $\mathbb{S}_{d,\boldsymbol{t}}$ are is in Theorem 8.7. Which of the operators reproduce the whole spline space?

a) $Q_{d,0}f = \sum_{j=1}^{n} f(x_j) B_{j,d}, \quad (r = 0).$

b) $Q_{d,1}f = \sum_{j=1}^{n} \big[f(x_j) + (t_j - x_j) Df(x_j) \big] B_{j,d}, \quad (r = 1).$

c) $\tilde{Q}_{d,1}f = \sum_{j=1}^{n} f(t_j^*) B_{j,d}, \quad (r = 1).$

d)

$$\begin{aligned} Q_{2,2}f = \sum_{j=1}^{n} \big[& f(x_j) - (x_j - t_{j+3/2}) Df(x_j) \\ & + \frac{1}{2}(x_j - t_{j+1})(x_j - t_{j+2}) D^2 f(x_j) \big] B_{j,2}, \text{(r=2)}. \end{aligned}$$

e) $\tilde{Q}_{2,2}f = \sum_{j=1}^{n} \left[f(t_{j+3/2}) - \frac{1}{2}(t_{j+2} - t_{j+1})^2 D^2 f(t_{j+3/2}) \right] B_{j,2}, \quad (r = 2).$

f)

$$Q_{3,3}f = \sum_{j=1}^{n} \big[f(t_{j+2}) + \frac{1}{3}(t_{j+3} - 2t_{j+2} + t_{j+1}) Df(t_{j+2}) - \frac{1}{6}(t_{j+3} - t_{j+2})(t_{j+2} - t_{j+1}) D^2 f(t_{j+2}) \big] B_{j,3}, \quad (r = 3).$$

CHAPTER 9

Approximation theory and stability

Polynomials of degree d have $d+1$ degrees of freedom, namely the $d+1$ coefficients relative to some polynomial basis. It turns out that each of these degrees of freedom can be utilised to gain approximation power so that the possible rate of approximation by polynomials of degree d is h^{d+1}, see Section 9.1. The meaning of this is that when a smooth function is approximated by a polynomial of degree d on an interval of length h, the error is bounded by Ch^{d+1}, where C is a constant that is independent of h. The exponent $d+1$ therefore controls how fast the error tends to zero with h.

When several polynomials are linked smoothly together to form a spline, each polynomial piece has $d+1$ coefficients, but some of these are tied up in satisfying the smoothness conditions. It therefore comes as a nice surprise that the approximation power of splines of degree d is the same as for polynomials, namely h^{d+1}, where h is now the largest distance between two adjacent knots. In passing from polynomials to splines we have therefore gained flexibility without sacrificing approximation power. We prove this in Section 9.2, by making use of some of the simple quasi-interpolants that we constructed in Chapter 8; it turns out that these produce spline approximations with the required accuracy.

The quasi-interpolants also allow us to establish two important properties of B-splines. The first is that B-splines form a stable basis for splines, see Section 9.3. This means that small perturbations of the B-spline coefficients can only lead to small perturbations in the spline, which is of fundamental importance for numerical computations. We have already seen that an important consequence of the stability of the B-spline basis is that the control polygon of a spline converges to the spline as the knot spacing tends to zero; this was proved in Section 4.1.

9.1 The distance to polynomials

We start by determining how well a given a real valued function f defined on an interval $[a, b]$ can be approximated by a polynomial of degree d. To measure the error in the approximation we will use the uniform norm which for a bounded function g defined on an interval $[a, b]$ is defined by

$$||g||_{\infty,[a,b]} = \sup_{a \le x \le b} |g(x)|.$$

Whenever we have an approximation p to f we can then measure the error by $||f - p||_{\infty,a,b}$. There are many possible approximations to f by polynomials of degree d, and the approximation that makes the error as small as possible is of course of special interest. This error is referred to as the *distance* from f to the space π_d of polynomials of degree $\leq d$ and is defined formally as

$$\text{dist}_{\infty,[a,b]}(f, \pi_d) = \inf_{p \in \pi_d} ||f - p||_{\infty,[a,b]}.$$

In order to bound this approximation error, we have to place some restrictions on the functions that we approximate, and we will only consider functions with piecewise continuous derivatives. Such functions lie in a space that we denote $C^k_{\mathbf{\Delta}}[a, b]$ for some integer $k \geq 0$. A function f lies in this space if it has $k - 1$ continuous derivatives on the interval $[a, b]$, and the kth derivative $D^k f$ is continuous everywhere except for a finite number of points in the interior (a, b), given by $\mathbf{\Delta} = (z_j)$. At the points of discontinuity $\mathbf{\Delta}$ the limits from the left and right given by $D^k f(z_j+)$ and $D^k f(z_j-)$, should exist so all the jumps are finite. If there are no continuous derivatives we write $C_{\mathbf{\Delta}}[a, b] = C^0_{\mathbf{\Delta}}[a, b]$. Note that we will often refer to these spaces without stating explicitly what the singularities $\mathbf{\Delta}$ are.

An upper bound for the distance of f to polynomials of degree d is fairly simple to give by choosing a particular approximation, namely Taylor expansion.

Theorem 9.1. *Given a polynomial degree d and a function f in $C^{d+1}_{\mathbf{\Delta}}[a, b]$, then*

$$\text{dist}_{\infty,[a,b]}(f, \pi_d) \leq K_d h^{d+1} ||D^{d+1} f||_{\infty,[a,b]},$$

where $h = b - a$ and

$$K_d = \frac{1}{2^{d+1}(d + 1)!}$$

depends only on d.

Proof. Consider the truncated Taylor series of f at the midpoint $m = (a + b)/2$ of $[a, b]$.

$$T_d f(x) = \sum_{k=0}^{d} \frac{(x - m)^k}{k!} D^k f(m), \quad \text{for } x \in [a, b].$$

Since $T_d f$ is a polynomial of degree d we clearly have

$$\text{dist}_{\infty,[a,b]}(f, \pi_d) \leq ||f - T_d f||_{\infty,[a,b]}. \tag{9.1}$$

To study the error we use the integral form of the remainder in the Taylor expansion,

$$f(x) - T_d f(x) = \frac{1}{d!} \int_m^x (x - y)^d D^{d+1} f(y) dy,$$

which is valid for any $x \in [a, b]$. If we restrict x to the interval $[m, b]$ we obtain

$$|f(x) - T_d f(x)| \leq ||D^{d+1} f||_{\infty,[a,b]} \frac{1}{d!} \int_m^x (x - y)^d dy.$$

The integral is given by

$$\frac{1}{d!} \int_m^x (x - y)^d dy = \frac{1}{(d + 1)!} (x - m)^{d+1} \leq \frac{1}{(d + 1)!} \left(\frac{h}{2}\right)^{d+1},$$

so for $x \geq m$ we have

$$|f(x) - T_d f(x)| \leq \frac{1}{2^{d+1}(d+1)!} h^{d+1} ||D^{d+1} f||_{\infty,[a,b]}.$$

By symmetry this estimate must also hold for $x \leq m$ and combining it with (9.1) completes the proof of the theorem. ∎

We remark that the best possible constant K_d can actually be computed. In fact, for each $f \in C^{d+1}[a,b]$ there is a point $\xi \in [a,b]$ such that

$$\text{dist}_{\infty,[a,b]}(f, \pi_d) = \frac{2}{4^{d+1}(d+1)!} h^{d+1} |D^{d+1} f(\xi)|$$

Applying this formula to the function $f(x) = x^{d+1}$ we see that the exponent $d+1$ in h^{d+1} is best possible.

9.2 The distance to splines

Just as we defined the distance from a function f to the space of polynomials of degree d we can define the distance from f to a spline space. Our aim is to show that on one knot interval, the distance from f to a spline space of degree d is essentially the same as the distance from f to the space of polynomials of degree d on a slightly larger interval, see Theorem 9.2 and Corollary 9.11. Our strategy is to consider the cases $d = 0$, 1 and 2 separately and then generalise to degree d. The main ingredient in the proof is a family of simple approximation methods called quasi-interpolants. As well as leading to good estimates of the distance between f and a spline space, many of the quasi-interpolants are good, practical approximation methods.

We consider a spline space $\mathbb{S}_{d,\boldsymbol{t}}$ where d is a nonnegative integer and $\boldsymbol{t} = (t_i)_{i=1}^{n+d+1}$ is a $d+1$ regular knot vector. We set

$$a = t_1, \quad b = t_{n+d+1}, \quad h_j = t_{j+1} - t_j, \quad h = \max_{1 \leq j \leq n} h_j.$$

Given a function f we consider the distance from f to $\mathbb{S}_{d,\boldsymbol{t}}$ defined by

$$\text{dist}_{\infty,[a,b]}(f, \mathbb{S}_{d,\boldsymbol{t}}) = \inf_{g \in \mathbb{S}_{d,\boldsymbol{t}}} ||f - g||_{\infty,[a,b]}.$$

We want to show the following.

Theorem 9.2. *Let the polynomial degree d and the function f in $C^{d+1}_{\boldsymbol{\Delta}}[a,b]$ be given. Then for any spline space $\mathbb{S}_{d,\boldsymbol{t}}$*

$$\text{dist}_{\infty,[a,b]}(f, \mathbb{S}_{d,\boldsymbol{t}}) \leq K_d h^{d+1} ||D^{d+1} f||_{\infty,[a,b]}, \tag{9.2}$$

where the constant K_d depends on d, but not on f, h or $\boldsymbol{t}$.

We will prove this theorem by constructing a spline $P_d f$ such that

$$|f(x) - P_d f(x)| \leq K_d h^{d+1} ||D^{d+1} f||_{\infty,[a,b]}, \quad x \in [a,b] \tag{9.3}$$

for a constant K_d depending only on d. The approximation $P_d f$ will be on the form

$$P_d f = \sum_{i=1}^{n} \lambda_i(f) B_{i,d}$$

where λ_i is a rule for computing the ith B-spline coefficient. We will restrict ourselves to rules λ_i like

$$\lambda_i(f) = \sum_{k=0}^{d} w_{i,k} f(x_{i,k})$$

where the points $(x_{i,k})_{k=0}^{d}$ all lie in one knot interval and $(w_{i,k})_{k=0}^{d}$ are suitable coefficients. These kinds of approximation methods are called *quasi-interpolants*.

9.2.1 The constant and linear cases

We first prove Theorem 9.2 in the low degree cases $d = 0$ and $d = 1$. For $d = 0$ the knots form a partition $a = t_1 < \cdots < t_{n+1} = b$ of $[a, b]$ and the B-spline $B_{i,0}$ is the characteristic function of the interval $[t_i, t_{i+1})$ for $i = 1, \ldots, n-1$, while $B_{n,0}$ is the characteristic function of the closed interval $[t_n, t_{n+1}]$. We consider the step function

$$g = P_0 f = \sum_{i=1}^{n} f(t_{i+1/2}) B_{i,0},$$

where $t_{i+1/2} = (t_i + t_{i+1})/2$. Fix $x \in [a, b]$ and let l be an integer such that $t_l \leq x < t_{l+1}$. We then have

$$f(x) - P_0 f(x) = f(x) - f(t_{l+1/2}) = \int_{t_{l+1/2}}^{x} Df(y) dy$$

so

$$|f(x) - P_0 f(x)| \leq |x - t_{l+1/2}| \; ||Df||_{\infty,[t_l,t_{l+1}]} \leq \frac{h}{2} ||Df||_{\infty,[a,b]}.$$

In this way we obtain (9.2) with $K_0 = 1/2$.

In the linear case $d = 1$ we define $P_1 f$ to be the piecewise linear interpolant to f on $\boldsymbol{t}$

$$g = P_1 f = \sum_{i=1}^{n} f(t_{i+1}) B_{i,1}.$$

Proposition 5.2 gives an estimate of the error in linear interpolation and by applying this result on each interval we obtain

$$||f - P_1 f||_{\infty,[a,b]} \leq \frac{h^2}{8} ||D^2 f||_{\infty,[a,b]}$$

which is (9.2) with $K_1 = 1/8$.

9.2.2 The quadratic case

Consider next the quadratic case $d = 2$. We shall approximate f by the quasi-interpolant $P_2 f$ that we constructed in Section 8.2.2. Its properties is summarised in the following lemma.

Lemma 9.3. *Suppose $\boldsymbol{t} = (t_i)_{i=1}^{n+3}$ is a knot vector with $t_{i+3} > t_i$ for $i = 1, \ldots, n$. The operator*

$$P_2 f = \sum_{i=1}^{n} \lambda_i(f) B_{i,2,\boldsymbol{t}}, \quad \text{with} \quad \lambda_i(f) = -\frac{1}{2} f(t_{i+1}) + 2f(t_{i+3/2}) - \frac{1}{2} f(t_{i+2}) \tag{9.4}$$

satisfies $P_2 p = p$ for all $p \in \pi_2$.

To show that (9.3) holds for $d = 2$ we now give a sequence of small lemmas.

Lemma 9.4. *Let $P_2(f)$ be as in* (9.4)*. Then*

$$|\lambda_i(f)| \le 3||f||_{\infty,[t_{i+1},t_{i+2}]}, \quad i = 1, \ldots, n. \tag{9.5}$$

Proof. Fix an integer i. Then

$$|\lambda_i(f)| = |-\frac{1}{2} f(t_{i+1}) + 2f(t_{i+3/2}) - \frac{1}{2} f(t_{i+2})| \le (\frac{1}{2} + 2 + \frac{1}{2})||f||_{\infty,[t_{i+1},t_{i+2}]}$$

from which the result follows. ∎

Lemma 9.5. *For $\ell = 3, \ldots, n$ we can bound $P_2 f$ on a subinterval $[t_\ell, t_{\ell+1}]$ by*

$$||P_2 f||_{\infty,[t_\ell,t_{\ell+1}]} \le 3||f||_{\infty,[t_{\ell-1},t_{\ell+2}]}. \tag{9.6}$$

Proof. Fix $x \in [t_\ell, t_{\ell+1}]$. Since the B-splines are nonnegative and form a partition of unity we have

$$\begin{aligned} |P_2 f(x)| &= \Big| \sum_{i=\ell-2}^{\ell} \lambda_i(f) B_{i,2,\boldsymbol{t}}(x) \Big| \le \max_{\ell-2 \le i \le \ell} |\lambda_i(f)| \\ &\le 3 \max_{\ell-2 \le i \le \ell} ||f||_{\infty,[t_{i+1},t_{i+2}]} = 3||f||_{\infty,[t_{\ell-1},t_{\ell+2}]}, \end{aligned}$$

where we used Lemma 9.4. This completes the proof. ∎

The following lemma shows that locally, the spline $P_2 f$ approximates f essentially as well as the best quadratic polynomial.

Lemma 9.6. *For $\ell = 3, \ldots, n$ the error $f - P_2 f$ on the interval $[t_\ell, t_{\ell+1}]$ is bounded by*

$$||f - P_2 f||_{\infty,[t_\ell,t_{\ell+1}]} \le 4 \operatorname{dist}_{\infty,[t_{\ell-1},t_{\ell+2}]}(f, \pi_2). \tag{9.7}$$

Proof. Let $p \in \pi_2$ be any quadratic polynomial. Since $P_2 p = p$ and P_2 is a linear operator, application of (9.6) to $f - p$ yields

$$\begin{aligned} \big|f(x) - (P_2 f)(x)\big| &= \big|f(x) - p(x) - \big((P_2 f)(x) - p(x)\big)\big| \\ &\le \big|f(x) - p(x)\big| + \big|P_2(f - p)(x)\big| \\ &\le (1 + 3)||f - p||_{\infty,[t_{\ell-1},t_{\ell+2}]}. \end{aligned} \tag{9.8}$$

Since p is arbitrary we obtain (9.7). ∎

We can now prove (9.2) for $d = 2$. For any interval $[a, b]$ Theorem 9.1 with $d = 2$ gives

$$\operatorname{dist}_{\infty,[a,b]}(f, \pi_2) \le K_2 h^3 ||D^3 f||_{\infty,[a,b]},$$

where $h = b - a$ and $K_2 = 1/(2^3 3!)$. Combining this estimate on $[a, b] = [t_{\ell-1}, t_{\ell+2}]$ with (9.7) we obtain (9.3) and hence (9.2).

9.2.3 The general case

The general case is analogous to the quadratic case, but the details are more complicated. Recall that for $d = 2$ we picked three points $x_{i,k} = t_{i+1} + k(t_{i+2} - t_{i+1})/2$ for $k = 0, 1, 2$ in each subinterval $[t_{i+1}, t_{i+2}]$ and then chose constants $w_{i,k}$ for $k = 0$, 1, 2 such that the operator

$$P_2 f = \sum_{i=1}^{n} \lambda_i(f) B_{i,2,\boldsymbol{t}}, \quad \text{with} \quad \lambda_i(f) = w_{i,0} f(x_{i,0}) + w_{i,1} f(x_{i,1}) + w_{i,2} f(x_{i,2}),$$

reproduced quadratic polynomials. We will follow the same strategy for general degree. The resulting quasi-interpolant is a special case of the one given in Theorem 8.7.

Suppose that $d \geq 2$ and fix an integer i such that $t_{i+d} > t_{i+1}$. We pick the largest subinterval $[a_i, b_i] = [t_l, t_{l+1}]$ of $[t_{i+1}, t_{i+d}]$ and define the uniformly spaced points

$$x_{i,k} = a_i + \frac{k}{d}(b_i - a_i), \quad \text{for } k = 0, 1, \ldots, d \tag{9.9}$$

in this interval. Given $f \in C_{\boldsymbol{\Delta}}[a, b]$ we define $P_d f \in \mathbb{S}_{d,\boldsymbol{t}}$ by

$$P_d f(x) = \sum_{i=1}^{n} \lambda_i(f) B_{i,d}(x), \quad \text{where} \quad \lambda_i(f) = \sum_{k=0}^{d} w_{i,k} f(x_{i,k}). \tag{9.10}$$

The following lemma shows how the coefficients $(w_{i,k})_{k=0}^{d}$ should be chosen so that $P_d p = p$ for all $p \in \pi_d$.

Lemma 9.7. *Suppose that in* (9.10) *the functionals* λ_i *are given by* $\lambda_i(f) = f(t_{i+1})$ *if* $t_{i+d} = t_{i+1}$, *while if* $t_{i+d} > t_{i+1}$ *we set*

$$w_{i,k} = \gamma_i(p_{i,k}), \quad k = 0, 1, \ldots, d, \tag{9.11}$$

where $\gamma_i(p_{i,k})$ *is the* i*th B-spline coefficient of the polynomial*

$$p_{i,k}(x) = \prod_{\substack{j=0 \\ j \neq k}}^{d} \frac{x - x_{i,j}}{x_{i,k} - x_{i,j}}. \tag{9.12}$$

Then the operator P_d *in* (9.10) *satisfies* $P_d p = p$ *for all* $p \in \pi_d$.

Proof. Suppose first that $t_{i+d} > t_{i+1}$. Any $p \in \pi_d$ can be written in the form

$$p(x) = \sum_{k=0}^{d} p(x_{i,k}) p_{i,k}(x). \tag{9.13}$$

For if we denote the function on the right by $q(x)$ then $q(x_{i,k}) = p(x_{i,k})$ for $k = 0$, 1, …, d, and since $q \in \pi_d$ it follows by the uniqueness of the interpolating polynomial that $p = q$. Now, by linearity of γ_i we have

$$\begin{aligned} \lambda_i(p) &= \sum_{k=0}^{d} w_{i,k} p(x_{i,k}) = \sum_{k=0}^{d} \gamma_i(p_{i,k}) p(x_{i,k}) \\ &= \gamma_i\big(\sum_{k=0}^{d} p_{i,k} p(x_{i,k})\big) = \gamma_i(p). \end{aligned}$$

If $t_{i+1} = t_{i+d}$ we know that a spline of degree d with knots $\boldsymbol{t}$ agrees with its $i+1$st coefficient at t_{i+1}. In particular, for any polynomial p we have $\lambda_i(p) = f(t_{i+1}) = \gamma_i(p)$. Alltogether this means that

$$P_d(p) = \sum_{i=1}^{n} \lambda_i(p) B_{i,d}(x) = \sum_{i=1}^{n} \gamma_i(p) B_{i,d}(x) = p$$

which confirms the lemma. ∎

The B-spline coefficients of $p_{i,k}$ can be found from the following lemma.

Lemma 9.8. *Given a spline space $\mathbb{S}_{d,\boldsymbol{t}}$ and numbers $v_1, \ldots, v_d$. The ith B-spline coefficient of the polynomial $p(x) = (x - v_1) \ldots (x - v_d)$ can be written*

$$\gamma_i(p) = \frac{1}{d!} \sum_{(j_1,\ldots,j_d)\in\Pi_d} (t_{i+j_1} - v_1) \cdots (t_{i+j_d} - v_d), \tag{9.14}$$

where Π_d is the set of all permutations of the integers $1, 2, \ldots, d$.

Proof. By Theorem 4.16 we have

$$\gamma_i(p) = \mathcal{B}[p](t_{i+1}, \ldots, t_{i+d}),$$

where $\mathcal{B}[p]$ is the blossom of p. It therefore suffices to verify that the expression (9.14) for $\gamma_i(p)$ satisfies the three properties of the blossom, but this is immediate. ∎

As an example, for $d = 2$ the set of all permutations of $1, 2$ are $\Pi_2 = \{(1, 2), (2, 1)\}$ and therefore

$$\gamma_i\big((x - v_1)(x - v_2)\big) = \frac{1}{2}\bigg((t_{i+1} - v_1)(t_{i+2} - v_2) + (t_{i+2} - v_1)(t_{i+1} - v_2)\bigg).$$

We can now give a bound for $\lambda_i(f)$.

Theorem 9.9. *Let $P_d(f) = \sum_{i=1}^{n} \lambda_i(f) B_{i,d}$ be the operator in Lemma 9.7. Then*

$$|\lambda_i(f)| \le K_d ||f||_{\infty,[t_{i+1},t_{i+d}]}, \quad i = 1, \ldots, n, \tag{9.15}$$

where

$$K_d = \frac{2^d}{d!}[d(d-1)]^d \tag{9.16}$$

depends only on d.

Proof. Fix an integer i. From Lemma 9.8 we have

$$w_{i,k} = \sum_{(j_1,\ldots,j_d)\in\Pi_d} \prod_{r=1}^{d} \left(\frac{t_{i+j_r} - v_r}{x_{i,k} - v_r}\right)/d!, \tag{9.17}$$

where $(v_r)_{r=1}^{d} = (x_{i,0}, \ldots, x_{i,k-1}, x_{i,k+1}, \ldots, x_{i,d})$. and Π_d denotes the set of all permutations of the integers 1, 2, ..., d. Since the numbers t_{i+j_r} and v_r belongs to the interval $[t_{i+1}, t_{i+d}]$ for all r we have the inequality

$$\prod_{r=1}^{d} (t_{i+j_r} - v_r) \le (t_{i+d} - t_{i+1})^d.$$

We also note that $x_{i,k} - v_r = (k-q)(b_i - a_i)/d$ for some q in the range $1 \le q \le d$ but with $q \ne k$. Taking the product over all r we therefore obtain

$$\prod_{r=1}^{d} |x_{i,k} - v_r| = \prod_{\substack{q=0 \\ q \ne k}}^{d} \frac{|k-q|}{d}(b_i - a_i) \ge k!(d-k)! \left(\frac{b_i - a_i}{d}\right)^d \ge k!(d-k)! \left(\frac{t_{i+d} - t_{i+1}}{d(d-1)}\right)^d$$

for all values of k and r since $[a_i, b_i]$ is the largest subinterval of $[t_{i+1}, t_{i+d}]$. Since the sum in (9.17) contains $d!$ terms, we find

$$\sum_{k=0}^{d} |w_{i,k}| \le \frac{[d(d-1)]^d}{d!} \sum_{k=0}^{d} \binom{d}{k} = \frac{2^d}{d!}[d(d-1)]^d = K_d$$

and hence

$$|\lambda_i(f)| \le ||f||_{\infty,[t_{i+1},t_{i+d}]} \sum_{k=0}^{d} |w_{i,k}| \le K_d ||f||_{\infty,[t_{i+1},t_{i+d}]} \tag{9.18}$$

which is the required inequality. ∎

From the bound for $\lambda_i(f)$ we easily obtain a bound for the norm of $P_d f$.

Theorem 9.10. *For $d+1 \le l \le n$ and $f \in C_{\mathbf{\Delta}}[a,b]$ we have the bound*

$$||P_d f||_{\infty,[t_l,t_{l+1}]} \le K_d ||f||_{\infty,[t_{l-d+1},t_{l+d}]}, \tag{9.19}$$

where K_d is the constant in Theorem 9.9.

Proof. Fix $x \in [t_l, t_{l+1}]$. Since the B-splines are nonnegative and form a partition of unity we have by Theorem 9.9

$$\begin{aligned} |P_d f(x)| = |\sum_{i=l-d}^{l} \lambda_i(f) B_{i,d,\mathbf{t}}(x)| &\le \max_{l-d \le i \le l} |\lambda_i(f)| \\ &\le K_d \max_{l-d \le i \le l} ||f||_{\infty,[t_{i+1},t_{i+d}]} = K_d ||f||_{\infty,[t_{l-d+1},t_{l+d}]} \end{aligned}$$

This completes the proof. ∎

The following corollary shows that $P_d f$ locally approximates f essentially as well as the best polynomial approximation of f of degree d.

Corollary 9.11. *For $l = d+1, \ldots, n$ the error $f - P_d f$ on the interval $[t_l, t_{l+1}]$ is bounded by*

$$||f - P_d f||_{\infty,[t_l,t_{l+1}]} \le (1 + K_d) \operatorname{dist}_{\infty,[t_{l-d+1},t_{l+d}]}(f, \pi_d), \tag{9.20}$$

where K_d is the constant in Theorem 9.9

Proof. We argue exactly as in the quadratic case. Let $p \in \pi_d$ be any polynomial in π_d. Since $P_d p = p$ and P_d is a linear operator we therefore have

$$\begin{aligned} \big|f(x) - (P_d f)(x)\big| &= \big|f(x) - p(x) - \big((P_d f)(x) - p(x)\big)\big| \\ &\le \big|f(x) - p(x)\big| + \big|P_d(f-p)(x)\big| \\ &\le (1 + K_d)||f - p||_{\infty,[t_{l-d+1},t_{l+d}]}. \end{aligned}$$

Since p is arbitrary we obtain (9.20). ∎

We can now prove (9.2) for general d. By Theorem 9.1 we have for any interval $[a, b]$

$$\text{dist}_{\infty,[a,b]}(f, \pi_d) \leq K_d h^{d+1} ||D^{d+1} f||_{\infty,[a,b]},$$

where $h = b-a$ and K_d only depends on d. Combining this estimate on $[a, b] = [t_{l-d+1}, t_{l+d}]$ with (9.20) we obtain (9.3) and hence (9.2).

9.3 Stability of the B-spline basis

In order to compute with polynomials or splines we need to choose a basis to represent the functions. If a basis is to be suitable for computer manipulations then it should be reasonably insensitive to round-off errors. In particular, functions with 'small' function values should have 'small' coefficients and vice versa. A basis with this property is said to be *well conditioned* or *stable*. In this section we will study the relationship between a spline and its coefficients quantitatively by introducing the *condition number* of a basis.

We have already seen that the size of a spline is bounded by its B-spline coefficients. There is also a reverse inequality, i.e., a bound on the B-spline coefficients in terms of the size of f. There are several reasons why such inequalities are important. In Section 4.1 we made use of this fact to estimate how fast the control polygon converges to the spline as more and more knots are inserted. A more direct consequence is that small relative perturbations in the coefficients can only lead to small changes in the function values. Both properties reflect the fact that the B-spline basis is well conditioned.

9.3.1 A general definition of stability

The stability of a basis can be defined quite generally. Instead of considering polynomials, we can consider a general linear vector space where we can measure the size of the elements through a norm; this is called a *normed linear space*.

Definition 9.12. *Let $\mathbb{U}$ be a normed linear space. A basis (ϕ_j) for $\mathbb{U}$ is said to be stable with respect to a vector norm $||\cdot||$ if there are small positive constants C_1 and C_2 such that*

$$C_1^{-1} \big\|(c_j)\big\| \leq \Big\| \sum_j c_j \phi_j \Big\| \leq C_2 \big\|(c_j)\big\|, \tag{9.21}$$

for all sets of coefficients $\boldsymbol{c} = (c_j)$. Let C_1^* and C_2^* denote the smallest possible values of C_1 and C_2 such that (9.21) holds. The condition number of the basis is then defined to be $\kappa = \kappa((\phi_i)_i) = C_1^* C_2^*$.

At the risk of confusion, we have used the same symbol both for the norm in $\mathbb{U}$ and the vector norm of the coefficients. In our case $\mathbb{U}$ will of course be some spline space $\mathbb{S}_{d,\boldsymbol{t}}$ and the basis (ϕ_j) will be the B-spline basis. The norms we will consider are the p-norms which are defined by

$$||f||_p = ||f||_{p,[a,b]} = \left(\int_a^b |f(x)|^p dx \right)^{1/p}, \quad \text{and} \quad ||\boldsymbol{c}||_p = \left(\sum_j |c_j|^p \right)^{1/p}$$

where f is a function on the interval $[a, b]$ and $\boldsymbol{c} = (c_j)$ is a real vector, and p is a real number in the range $1 \leq p < \infty$ for any real number. For $p = \infty$ the norms are defined by

$$||f||_\infty = ||f||_{\infty,[a,b]} = \max_{a \leq x \leq b} |f(x)|, \quad \text{and} \quad ||\boldsymbol{c}||_\infty = \big\|(c_j)\big\|_\infty = \max_j |c_j|,$$

In practice, the most important norms are the 1-, 2- and ∞-norms.

In Definition 9.12 we require the constants C_1 and C_2 to be 'small', but how small is 'small'? There is no unique answer to this question, but it is typically required that C_1 and C_2 should be independent of the dimension n of $\mathbb{U}$, or at least grow very slowly with n. Note that we always have $\kappa \geq 1$, and $\kappa = 1$ if and only if we have equality in both inequalities in (9.21).

A stable basis is desirable for many reasons, and the constant $\kappa = C_1 C_2$ crops up in many different contexts. The condition number κ does in fact act as a sort of derivative of the basis and gives a measure of how much an error in the coefficients is magnified in a function value.

Proposition 9.13. *Suppose (ϕ_j) is a stable basis for $\mathbb{U}$. If $f = \sum_j c_j \phi_j$ and $g = \sum_j b_j \phi_j$ are two elements in $\mathbb{U}$ with $f \neq 0$, then*

$$\frac{||f - g||}{||f||} \leq \kappa \frac{||\boldsymbol{c} - \boldsymbol{b}||}{||\boldsymbol{c}||}, \tag{9.22}$$

where κ is the condition number of the basis as in Definition 9.12.

Proof. From (9.21), we have the two inequalities $||f - g|| \leq C_2 ||(c_j - b_j)||$ and $1/||f|| \leq C_1/||(c_j)||$. Multiplying these together gives the result. ∎

If we think of g as an approximation to f, then (9.22) says that the relative error in $f - g$ is bounded by at most κ times the relative error in the coefficients. If κ is small, then a small relative error in the coefficients gives a small relative error in the function values. This is important in floating point calculations on a computer. A function is usually represented by its coefficients relative to some basis. Normally, the coefficients are real numbers that must be represented inexactly as floating point numbers in a computer. This round-off error means that the computed spline, here g, will differ from the exact f. Proposition 9.13 shows that this is not so serious if the perturbed coefficients of g are close to those of f and the basis is stable.

Proposition 9.13 also provides some information as to what are acceptable values of C_1^* and C_2^*. If for example $\kappa = C_1^* C_2^* = 100$ we risk losing 2 decimal places in evaluation of a function; exactly how much accuracy one can afford to lose will of course vary.

One may wonder whether there are any unstable polynomial bases. It turns out that the power basis 1, x, x^2, ..., on the interval $[0, 1]$ is unstable even for quite low degrees. Already for degree 10, one risks losing as much as 4 or 5 decimal digits in the process of computing the value of a polynomial on the interval $[0, 1]$ relative to this basis, and other operations such as numerical root finding is even more sensitive.

9.3.2 The condition number of the B-spline basis. Infinity norm

Since splines and B-splines are defined via the knot vector, it is quite conceivable that the condition number of the B-spline basis could become arbitrarily large for certain knot configurations, for example in the limit when two knots merge into one. We will now prove that the condition number of the B-spline basis can be bounded independently of the knot vector so it cannot grow beyond all bounds when the knots vary.

The best constant C_2^* in Definition 9.12 can be found quite easily for the B-spline basis.

Lemma 9.14. *In all spline spaces $\mathbb{S}_{d,\boldsymbol{t}}$ the bound*

$$\Big\|\sum_{i=1}^{m} b_i B_{i,d}\Big\|_{\infty,[t_1,t_{m+1+d}]} \le ||\boldsymbol{b}||_\infty$$

holds. Equality holds if $b_i = 1$ for all i and the knot vector $\boldsymbol{t} = (t_i)_{i=0}^{n+d}$ is $d+1$-extended; in this case $C_2^* = 1$.

Proof. This follows since the B-splines are nonnegative and sum to one. ∎

To find a bound for the constant C_1 we shall use the operator P_d given by (9.3). We recall that P_d reproduces polynomials of degree d, i.e., $P_d p = p$ for all $p \in \pi_d$. We now show that more is true; we have in fact that P_d reproduces all splines in $\mathbb{S}_{d,\boldsymbol{t}}$.

Theorem 9.15. *The operator*

$$P_d f = \sum_{i=1}^{n} \lambda_i(f) B_{i,d}$$

given by (9.3) *reproduces all splines in $\mathbb{S}_{d,\boldsymbol{t}}$, $P_d f = f$ for all $f \in \mathbb{S}_{d,\boldsymbol{t}}$.*

Proof. We first show that

$$\lambda_j(B_{k,d}) = \delta_{j,k}, \qquad \text{for } j,k = 1,\ldots,n. \tag{9.23}$$

Fix i and let

$$I_i = [a_i, b_j] = [t_{l_i}, t_{l_i+1}]$$

be the interval used to define $\lambda_i(f)$. We consider the polynomials

$$\phi_k = B_{k,d}|_{I_i} \qquad \text{for } l_i - d \le k \le l_i$$

obtained by restricting the B-splines $\{B_{k,d}\}_{k=l_i-d}^{l_i}$ to the interval I_i. Since P_d reproduces π_d we have

$$\phi_k(x) = (P_d\phi_k)(x) = \sum_{j=l_i-d}^{l_i} \lambda_j(\phi_k)\phi_j(x)$$

for x in the interval I_i. By the linear independence of the the polynomials (ϕ_k) we therefore obtain

$$\lambda_j(B_{k,d}) = \lambda_j(\phi_k) = \delta_{j,k}, \qquad \text{for } j,k = l_i - d,\ldots,l_i.$$

In particular we have $\lambda_i B_{i,d} = 1$ since $l_i - d \le i \le l_i$. For $k < l_i - d$ or $k > l_i$ the support of $B_{k,d}$ has empty intersection with I_i so $\lambda_i(B_{k,d}) = 0$ for these values of k. Thus (9.23) holds for all k.

To complete the proof suppose $f = \sum_{k=1}^{n} c_k B_{k,d}$ is a spline in $\mathbb{S}_{d,\boldsymbol{t}}$. From (9.23) we then have

$$Qf = \sum_{j=1}^{n}\Big(\sum_{k=1}^{n} c_k\lambda_j(B_{k,d})\Big)B_{j,d} = \sum_{j=1}^{n} c_j B_{j,d} = f. \quad ∎$$

To obtain an upper bound for C_1^* we note that the leftmost inequality in (9.21) is equivalent to

$$|b_i| \leq C_1||f||, \quad i = 1, \ldots, m.$$

Lemma 9.16. *There is a constant K_d, depending only on the polynomial degree d, such that for all splines $f = \sum_{i=1}^m b_i B_{i,d}$ in some given spline space $\mathbb{S}_{d,\boldsymbol{t}}$ the inequality*

$$|b_i| \leq K_d||f||_{[t_{i+1},t_{i+d}]} \tag{9.24}$$

holds for all i.

Proof. Consider the operator P_d given in Lemma 9.7. Since $P_d f = f$ we have $b_i = \lambda_i(f)$. The result now follows from (9.15) ∎

Note that if $[a,b] \subseteq [c,d]$, then $||f||_{\infty,[a,b]} \leq ||f||_{\infty,[c,d]}$. From (9.24) we therefore conclude that $|b_i| \leq K_d\|f\|_{\infty,[t_1,t_{m+1+d}]}$ for all i or briefly $||\boldsymbol{b}|| \leq K_d\|f\|$. The constant K_d can therefore be used as C_1 in Definition 9.12 in the case where the norm is the ∞-norm. Combining the two lemmas we obtain the following theorem.

Theorem 9.17. *There is a constant K_1, depending only on the polynomial degree d, such that for all spline spaces $\mathbb{S}_{d,\boldsymbol{t}}$ and all splines $f = \sum_{i=1}^m b_i B_{i,d} \in \mathbb{S}_{d,\boldsymbol{t}}$ with B-spline coefficients $\boldsymbol{b} = (b_i)_{i=1}^m$ the inequalities*

$$K_1^{-1}||\boldsymbol{b}||_\infty \leq ||f||_{\infty,[t_1,t_{m+d}]} \leq ||\boldsymbol{b}||_\infty \tag{9.25}$$

hold.

The condition number of the B-spline basis on the knot vector $\boldsymbol{t}$ with respect to the ∞-norm is usually denoted $\kappa_{d,\infty,\boldsymbol{t}}$. By taking the supremum over all knot vectors we obtain the knot independent condition number $\kappa_{d,\infty}$,

$$\kappa_{d,\infty} = \sup_{\boldsymbol{t}} \kappa_{d,\infty,\boldsymbol{t}}.$$

Theorem 9.17 shows that $\kappa_{d,\infty}$ is bounded above by K_1.

The estimate K_d for C_1^* given by (9.16) is a number which grows quite rapidly with d and does not indicate that the B-spline basis is stable. However, it is possible to find better estimates for the condition number, and it is known that the B-spline basis is very stable, at least for moderate values of d. To determine the condition number is relatively simple for $d \leq 2$; we have $\kappa_{0,\infty} = \kappa_{1,\infty} = 1$ and $\kappa_{2,\infty} = 3$. For $d \geq 3$ it has recently been shown that $\kappa_{d,\infty} = O(2^d)$. The first few values are known numerically to be $\kappa_{3,\infty} \approx 5.5680$ and $\kappa_{4,\infty} \approx 12.088$.

9.3.3 The condition number of the B-spline basis. p-norm

With $1 \leq p \leq \infty$ and q such that $1/p+1/q = 1$ we recall the Hölder inequality for functions

$$\int_a^b |f(x)g(x)|dx \leq ||f||_p||g||_q,$$

and the Hölder inequality for sums

$$\sum_{i=1}^m |b_i c_i| \leq ||(b_i)_{i=1}^m||_p||(c_i)_{i=1}^m||_q.$$

We also note that for any polynomial $g \in \pi_d$ and any interval $[a,b]$ we have

$$|g(x)| \leq \frac{C}{b-a}\int_a^b |g(x)|\,dx, \quad x \in [a,b], \tag{9.26}$$

where the constant C only depends on the degree d. This follows on $[a,b] = [0,1]$ since all norms on a finite dimensional vector space are equivalent, and then on an arbitrary interval $[a,b]$ by a change of variable.

In order to generalise the stability result (9.25) to arbitrary p-norms we need to scale the B-splines differently. We define the p-norm B-splines to be identically zero if $t_{i+d+1} = t_i$ and

$$B^p_{i,d,\boldsymbol{t}} = \left(\frac{d+1}{t_{i+d+1}-t_i}\right)^{1/p} B_{i,d,\boldsymbol{t}}, \tag{9.27}$$

otherwise.

Theorem 9.18. *There is a constant K, depending only on the polynomial degree d, such that for all $1 \leq p \leq \infty$, all spline spaces $\mathbb{S}_{d,\boldsymbol{t}}$ and all splines $f = \sum_{i=1}^m b_i B^p_{i,d} \in \mathbb{S}_{d,\boldsymbol{t}}$ with p-norm B-spline coefficients $\boldsymbol{b} = (b_i)_{i=1}^m$ the inequalities*

$$K^{-1}||\boldsymbol{b}||_p \leq ||f||_{p,[t_1,t_{m+d}]} \leq ||\boldsymbol{b}||_p \tag{9.28}$$

hold.

Proof. We first prove the upper inequality. Let $\gamma_i = (d+1)/(t_{i+d+1}-t_i)$ for $i = 1,\ldots,m$ and set $[a,b] = [t_1, t_{m+d+1}]$. Using the Hölder inequality for sums we have

$$\sum_i |b_i B^p_{i,d}| = \sum_i |b_i \gamma_i^{1/p} B^{1/p}_{i,d}| B^{1/q}_{i,d} \leq \Big(\sum_i |b_i|^p \gamma_i B_{i,d}\Big)^{1/p} \Big(\sum_i B_{i,d}\Big)^{1/q}.$$

Raising this to the pth power and using the partition of unity property we obtain the inequality

$$\big|\sum_i b_i B^p_{i,d}(x)\big|^p \leq \sum_i |b_i|^p \gamma_i B_{i,d}(x), \quad x \in \mathbb{R}.$$

Therefore, recalling that $\int B_{i,d}(x)dx = 1/\gamma_i$ we find

$$||f||^p_{p,[a,b]} = \int_a^b \big|\sum_i b_i B^p_{i,d}(x)\big|^p\,dx \leq \sum_i |b_i|^p \gamma_i \int_a^b B_{i,d}(x)\,dx = \sum_i |b_i|^p.$$

Taking pth roots proves the upper inequality.

Consider now the lower inequality. Recall from (9.24) that we can bound the B-spline coefficients in terms of the infinity norm of the function. In terms of the coefficients b_i of the p-norm B-splines we obtain from (9.24) for all i

$$\left(\frac{d+1}{t_{i+d+1}-t_i}\right)^{1/p} |b_i| \leq K_1 \max_{t_{i+1}\leq x\leq t_{i+d}} |f(x)|,$$

where the constant K_1 only depends on d. Taking max over a larger subinterval, using (9.26), and then Hölder for integrals we find

$$\begin{aligned}
|b_i| &\le K_1(d+1)^{-1/p}\big(t_{i+d+1}-t_i\big)^{1/p}|\max_{t_i\le x\le t_{i+d+1}}|f(x)| \\
&\le CK_1(d+1)^{-1/p}\big(t_{i+d+1}-t_i\big)^{-1+1/p}\int_{t_i}^{t_{i+d+1}}|f(y)|\,dy \\
&\le CK_1(d+1)^{-1/p}\bigg(\int_{t_i}^{t_{i+d+1}}|f(y)|^p\,dy\bigg)^{1/p}
\end{aligned}$$

Raising both sides to the pth power and summing over i we obtain

$$\sum_i |b_i|^p \le C^pK_1^p(d+1)^{-1}\sum_i\int_{t_i}^{t_{i+d+1}}|f(y)|^p\,dy \le C^pK_1^p||f||^p_{p,[a,b]}.$$

Taking pth roots we obtain the lower inequality in (9.28) with $K = CK_1$. ∎

Exercises for Chapter 9

9.1 In this exercise we will study the order of approximation by the Schoenberg Variation Diminishing Spline Approximation of degree $d \ge 2$. This approximation is given by

$$V_df = \sum_{i=1}^n f(t_i^*)B_{i,d}, \quad \text{with} \quad t_i^* = \frac{t_{i+1}+\cdots t_{i+d}}{d}.$$

Here $B_{i,d}$ is the ith B-spline of degree d on a $d+1$-regular knot vector $\boldsymbol{t} = (t_i)_{i=1}^{n+d+1}$. We assume that $t_{i+d} > t_i$ for $i = 2,\dots,n$. Moreover we define the quantities

$$a = t_1, \quad b = t_{n+d+1}, \quad h = \max_{1\le i\le n} t_{i+1}-t_i.$$

We want to show that V_df is an $O(h^2)$ approximation to a sufficiently smooth f.

We first consider the more general spline approximation

$$\tilde{V}_df = \sum_{i=1}^n \lambda_i(f)B_{i,d}, \quad \text{with} \quad \lambda_i(f) = w_{i,0}f(x_{i,0}) + w_{i,1}f(x_{i,1}).$$

Here $x_{i,0}$ and $x_{i,1}$ are two distinct points in $[t_i, t_{i+d}]$ and $w_{i,0}, w_{i,1}$ are constants, $i = 1,\dots,n$.

Before attempting to solve this exercise the reader might find it helpful to review Section 9.2.2

a) Suppose for $i = 1,\dots,n$ that $w_{i,0}$ and $w_{i,1}$ are such that

$$\begin{aligned}
w_{i,0} + w_{i,1} &= 1 \\
x_{i,0}w_{i,0} + x_{i,1}w_{i,1} &= t_i^*
\end{aligned}$$

Show that then $\tilde{V}_dp = p$ for all $p \in \pi_1$. (Hint: Consider the polynomials $p(x) = 1$ and $p(x) = x$.)

b) Show that if we set $x_{i,0} = t_i^*$ for all i then $\tilde{V}_d f = V_d f$ for all f, regardless of how we choose the value of $x_{i,1}$.

In the rest of this exercise we set $\lambda_i(f) = f(t_i^*)$ for $i = 1, \ldots, n$, i.e. we consider $V_d f$. We define the usual uniform norm on an interval $[c, d]$ by

$$||f||_{[c,d]} = \sup_{c \le x \le d} |f(x)|, \quad f \in C_{\mathbf{\Delta}}[c, d].$$

c) Show that for $d + 1 \le l \le n$

$$||V_d f||_{[t_l, t_{l+1}]} \le ||f||_{[t_{l-d}^*, t_l^*]}, \quad f \in C_{\mathbf{\Delta}}[a, b].$$

d) Show that for $f \in C_{\mathbf{\Delta}}[t_{l-d}^*, t_l^*]$ and $d + 1 \le l \le n$

$$||f - V_d f||_{[t_l, t_{l+1}]} \le 2 \operatorname{dist}_{[t_{l-d}^*, t_l^*]}(f, \pi_1).$$

e) Explain why the following holds for $d + 1 \le l \le n$

$$\operatorname{dist}_{[t_{l-d}^*, t_l^*]}(f, \pi_1) \le \frac{(t_l^* - t_{l-d}^*)^2}{8} ||D^2 f||_{[t_{l-d}^*, t_l^*]}.$$

f) Show that the following $O(h^2)$ estimate holds

$$||f - V_d f||_{[a,b]} \le \frac{d^2}{4} h^2 ||D^2 f||_{[a,b]}.$$

(Hint: Verify that $t_l^* - t_{l-d}^* \le hd$.)

9.2 In this exercise we want to perform a numerical simulation experiment to determine the order of approximation by the quadratic spline approximations

$$V_2 f = \sum_{i=1}^{n} f(t_i^*) B_{i,2}, \quad \text{with} \quad t_i^* = \frac{t_{i+1} + t_{i+2}}{2},$$
$$P_2 f = \sum_{i=1}^{n} \big(-\frac{1}{2} f(t_{i+1}) + 2 f(t_i^*) - \frac{1}{2} f(t_{i+2}) \big) B_{i,2}.$$

We want to test the hypotheses $f - V_2 f = O(h^2)$ and $f - P_2 f = O(h^3)$ where $h = \max_i t_{i+1} - t_i$. We test these on the function $f(x) = \sin x$ on $[0, \pi]$ for various values of h. Consider for $m \ge 0$ and $n_m = 2 + 2^m$ the 3-regular knot vector $\boldsymbol{t}^m = (t_i^m)_{i=1}^{n_m+3}$ on the interval $[0, \pi]$ with uniform spacing $h_m = \pi 2^{-m}$. We define

$$V_2^m f = \sum_{i=1}^{n} f(t_{i+3/2}^m) B_{i,2}^m, \quad \text{with} \quad t_i^m = \frac{t_{i+1}^m + t_{i+2}^m}{2},$$
$$P_2^m f = \sum_{i=1}^{n} \big(-\frac{1}{2} f(t_{i+1}^m) + 2 f(t_{i+3/2}^m) - \frac{1}{2} f(t_{i+2}^m) \big) B_{i,2}^m,$$

and $B_{i,2}^m$ is the ith quadratic B-spline on $\boldsymbol{t}^m$. As approximations to the norms $||f - V_2^m f||_{[0,\pi]}$ and $||f - P_2^m f||_{[0,\pi]}$ we use

$$E_V^m = \max_{0 \le j \le 100} |f(j\pi/100) - V_2^m f(j\pi/100)|,$$
$$E_P^m = \max_{0 \le j \le 100} |f(j\pi/100) - P_2^m f(j\pi/100)|.$$

Write a computer program to compute numerically the values of E_V^m and E_P^m for $m = 0, 1, 2, 3, 4, 5$, and the ratios E_V^m/E_V^{m-1} and E_P^m/E_P^{m-1} for $1 \le m \le 5$. What can you deduce about the approximation order of the two methods?

Make plots of $V_2^m f$, $P_2^m f$, $f - V_2^m f$, and $f - P_2^m f$ for some values of m.

9.3 Suppose we have $m \ge 3$ data points $\big(x_i, f(x_i)\big)_{i=1}^m$ sampled from a function f, where the abscissas $\boldsymbol{x} = (x_i)_{i=1}^m$ satisfy $x_1 < \cdots < x_m$. In this exercise we want to derive a local quasi-interpolation scheme which only uses the data values at the x_i's and which has $O(h^3)$ order of accuracy if the y-values are sampled from a smooth function f. The method requires m to be odd.

From $\boldsymbol{x}$ we form a 3-regular knot vector by using every second data point as a knot

$$\boldsymbol{t} = (t_j)_{j=1}^{n+3} = (x_1, x_1, x_1, x_3, x_5, \ldots, x_{m-2}, x_m, x_m, x_m), \tag{9.29}$$

where $n = (m+3)/2$. In the quadratic spline space $\mathbb{S}_{2,\boldsymbol{t}}$ we can then construct the spline

$$Q_2 f = \sum_{j=1}^{n} \lambda_j(f) B_{j,2}, \tag{9.30}$$

where the B-spline coefficients $\lambda_j(f)_{j=1}^n$ are defined by the rule

$$\lambda_j(f) = \frac{1}{2}\bigg(-\theta_j^{-1} f(x_{2j-3}) + \theta_j^{-1}(1+\theta_j)^2 f(x_{2j-2}) - \theta_j f(x_{2j-1}) \bigg), \tag{9.31}$$

for $j = 1, \ldots, n$. Here $\theta_1 = \theta_n = 1$ and

$$\theta_j = \frac{x_{2j-2} - x_{2j-3}}{x_{2j-1} - x_{2j-2}}$$

for $j = 2, \ldots, n-1$.

a) Show that Q_2 simplifies to P_2 given by (9.4) when the data abscissas are uniformly spaced.

b) Show that $Q_2 p = p$ for all $p \in \pi_2$ and that because of the multiple abscissas at the ends we have $\lambda_1(f) = f(x_1)$, $\lambda_n(f) = f(x_m)$, so only the original data are used to define $Q_2 f$. (Hint: Use the formula in Exercise 1.

c) Show that for $j = 1, \ldots, n$ and $f \in C_{\boldsymbol{\Delta}}[x_1, x_m]$

$$|\lambda_j(f)| \le (2\theta + 1)||f||_{\infty,[t_{j+1},t_{j+2}]},$$

where

$$\theta = \max_{1 \le j \le n} \{\theta_j^{-1}, \theta_j\}.$$

d) Show that for $l = 3, \ldots, n$, $f \in C_{\mathbf{\Delta}}[x_1, x_m]$, and $x \in [t_l, t_{l+1}]$

$$|Q_2(f)(x)| \leq (2\theta + 1)||f||_{\infty,[t_{l-1},t_{l+2}]}.$$

e) Show that for $l = 3, \ldots, n$ and $f \in C_{\mathbf{\Delta}}[x_1, x_m]$

$$||f - Q_2 f||_{\infty,[t_l,t_{l+1}]} \leq (2\theta + 2)\,\mathrm{dist}_{[t_{l-1},t_{l+2}]}(f, \pi_2).$$

f) Show that for $f \in C^3_{\mathbf{\Delta}}[x_1, x_m]$ we have the $O(h^3)$ estimate

$$||f - Q_2 f||_{\infty,[x_1,x_m]} \leq K(\theta)|\Delta x|^3||D^3 f||_{\infty,[x_1,x_m]},$$

where

$$|\Delta x| = \max_j |x_{j+1} - x_j|$$

and the constant $K(\theta)$ only depends on θ.

CHAPTER 10

Shape Preserving Properties of B-splines

In earlier chapters we have seen a number of examples of the close relationship between a spline function and its B-spline coefficients. This is especially evident in the properties of the Schoenberg operator, but the same phenomenon is apparent in the diagonal property of the blossom, the stability of the B-spline basis, the convergence of the control polygon to the spline it represents and so on. In the present chapter we are going to add to this list by relating the number of zeros of a spline to the number of sign changes in the sequence of its B-spline coefficients. From this property we shall obtain an accurate characterisation of when interpolation by splines is uniquely solvable. In the final section we show that the knot insertion matrix and the B-spline collocation matrix are totally positive, i.e., all their square submatrices have nonnegative determinants.

10.1 Bounding the number of zeros of a spline

In Section 4.5 of Chapter 4 we showed that the number of sign changes in a spline is bounded by the number of sign changes in its B-spline coefficients, a generalisation of Descartes' rule of signs for polynomials, Theorem 4.23. Theorem 4.25 is not a completely satisfactory generalisation of Theorem 4.23 since it does not allow multiple zeros. In this section we will prove a similar result that does allow multiple zeros, but we cannot allow the most general spline functions. we have to restrict ourselves to *connected splines*.

Definition 10.1. *A spline* $f = \sum_{j=1}^{n} c_j B_{j,d}$ *in* $\mathbb{S}_{d,\boldsymbol{\tau}}$ *is said to be* connected *if for each* x *in* (τ_1, τ_{n+d+1}) *there is some* j *such that* $\tau_j < x < \tau_{j+d+1}$ *and* $c_j \neq 0$. *A point* x *where this condition fails is called a* splitting point *for* f.

To develop some intuition about connected splines, let us see when a spline is not connected. A splitting point of f can be of two kinds:

(i) The splitting point x is not a knot. If $\tau_\mu < x < \tau_{\mu+1}$, then $\tau_j < x < \tau_{j+d+1}$ for $j = \mu - d, \ldots, \mu$ (assuming the knot vector is long enough) so we must have $c_{\mu-d} = \cdots = c_\mu = 0$. In other words f must be identically zero on $(\tau_\mu, \tau_{\mu+1})$. In this case f splits into two spline functions f_1 and f_2 with knot vectors $\boldsymbol{\tau}^1 = (\tau_j)_{j=1}^{\mu}$ and

$\boldsymbol{\tau}^2 = (\tau_j)_{j=\mu+1}^{n+d+1}$. We clearly have

$$f_1 = \sum_{j=1}^{\mu-d-1} c_j B_{j,d}, \qquad f_2 = \sum_{j=\mu+1}^{n} c_j B_{j,d}.$$

(ii) The splitting point x is a knot of multiplicity m, say

$$\tau_\mu < x = \tau_{\mu+1} = \cdots = \tau_{\mu+m} < \tau_{\mu+m+1}.$$

In this case we have $\tau_j < x < \tau_{j+1+d}$ for $j = \mu + m - d, \ldots, \mu$. We must therefore have $c_{\mu+m-d} = \cdots = c_\mu = 0$. (Note that if $m = d + 1$, then no coefficients need to be zero). This means that all the B-splines that "cross" x do not contribute to f. It therefore splits into two parts f_1 and f_2, but now the two pieces are not separated by an interval, but only by the single point x. The knot vector of f_1 is $\boldsymbol{\tau}^1 = (\tau_j)_{j=1}^{\mu+m}$ and the knot vector of f_2 is $\boldsymbol{\tau}^2 = (\tau_j)_{j=\mu+1}^{n+d+1}$, while

$$f_1 = \sum_{j=1}^{\mu+m-d-1} c_j B_{j,d}, \qquad f_2 = \sum_{j=\mu+1}^{n} c_j B_{j,d}.$$

Before getting on with our zero counts we need the following lemma.

Lemma 10.2. *Suppose that z is a knot that occurs m times in $\boldsymbol{\tau}$,*

$$\tau_i < z = \tau_{i+1} = \cdots = \tau_{i+m} < \tau_{i+m+1}$$

for some i. Let $f = \sum_j c_j B_{j,d}$ be a spline in $\mathbb{S}_{d,\boldsymbol{\tau}}$. Then

$$c_j = \frac{1}{d!} \sum_{k=0}^{d-m} (-1)^k D^{d-k} \rho_{j,d}(z) D^k f(z) \tag{10.1}$$

for all j such that $\tau_j < z < \tau_{j+d+1}$, where $\rho_{j,d}(y) = (y - \tau_{j+1}) \cdots (y - \tau_{j+d})$.

Proof. Recall from Theorem 8.5 that the B-spline coefficients of f can be written as

$$c_j = \lambda_j f = \frac{1}{d!} \sum_{k=0}^{d} (-1)^k D^{d-k} \rho_{j,d}(y) D^k f(y),$$

where y is a number such that $B_{j,d}(y) > 0$. In particular, we may choose $y = z$ for $j = i + m - d, \ldots, i$ so

$$c_j = \lambda_j f = \frac{1}{d!} \sum_{k=0}^{d} (-1)^k D^{d-k} \rho_{j,d}(z) D^k f(z), \tag{10.2}$$

for these values of j. But in this case $\rho_{j,d}(y)$ contains the factor $(y - \tau_{i+1}) \cdots (y - \tau_{i+m}) = (y - z)^m$ so $D^{d-k} \rho_{j,d}(z) = 0$ for $k > d - m$ and $j = i + m - d, \ldots, i$, i.e., for all values of j such that $\tau_j < z < \tau_{j+d+1}$. The formula (10.1) therefore follows from (10.2). ∎

In the situation of Lemma 10.2, we know from Lemma 2.6 that $D^k f$ is continuous at z for $k = 0, \ldots, d-m$, but $D^{d+1-m} f$ may be discontinuous. Equation (10.1) therefore shows that the B-spline coefficients of f can be computed solely from continuous derivatives of f at a point.

Lemma 10.3. *Let f be a spline that is connected. For each x in (τ_1, τ_{n+d+1}) there is then a nonnegative integer r such that $D^r f$ is continuous at x and $D^r f(x) \neq 0$.*

Proof. The claim is clearly true if x is not a knot, for otherwise f would be identically zero on an interval and therefore not connected. Suppose next that x is a knot of multiplicity m. Then the first discontinuous derivative at x is $D^{d-m+1} f$, so if the claim is not true, we must have $D^j f(x) = 0$ for $j = 0, \ldots, d-m$. But then we see from Lemma 10.2 that $c_l = \lambda_l f = 0$ for all l such that $\tau_l < x < \tau_{l+d+1}$. But this is impossible since f is connected. ∎

The lemma shows that we can count zeros of connected splines precisely as for smooth functions. If f is a connected spline then a zero must be of the form $f(z) = Df(z) = \cdots = D^{r-1} f(z) = 0$ with $D^r f(z) \neq 0$ for some integer r. Moreover $D^r f$ is continuous at z. The total number of zeros of f on (a, b), counting multiplicities, is denoted $Z(f) = Z_{(a,b)}(f)$. Recall from Definition 4.21 that $S^-(\boldsymbol{c})$ denotes the number of sign changes in the vector $\boldsymbol{c}$ (zeros are completely ignored).

Example 10.4. Below are some examples of zero counts of functions. For comparison we have also included counts of sign changes. All zero counts are over the whole real line.

$$
\begin{array}{llll}
Z(x) = 1, & S^-(x) = 1, & Z\big(x(1-x)^2\big) = 3, & S^-\big(x(1-x)^2\big) = 1, \\
Z(x^2) = 2, & S^-(x^2) = 0, & Z\big(x^3(1-x)^2\big) = 5, & S^-\big(x^3(1-x)^2\big) = 1, \\
Z(x^7) = 7, & S^-(x^7) = 1, & Z(-1-x^2+\cos x) = 2, & S^-(-1-x^2+\cos x) = 0.
\end{array}
$$

We are now ready to prove a generalization of Theorem 4.23 that allows zeros to be counted with multiplicities.

Theorem 10.5. *Let $f = \sum_{j=1}^n c_j B_{j,d}$ be a spline in $\mathbb{S}_{d,\boldsymbol{\tau}}$ that is connected. Then*

$$Z_{(\tau_1, \tau_{n+d+1})}(f) \leq S^-(\boldsymbol{c}) \leq n-1.$$

Proof. Let $z_1 < z_2 < \cdots < z_\ell$ be the zeros of f in the interval (τ_1, τ_{n+d+1}), each of multiplicity r_i; Lemma 10.2 shows that z_i occurs at most $d - r_i$ times in $\boldsymbol{\tau}$. For if z_i occured $m > d - r_i$ times in $\boldsymbol{\tau}$ then $d - m < r_i$, and hence $c_j = 0$ by (10.1) for all j such that $\tau_j < z < \tau_{j+d+1}$, which means that z is a splitting point for f. But this is impossible since f is connected.

Now we form a new knot vector $\hat{\boldsymbol{\tau}}$ where z_i occurs exactly $d - r_i$ times and the numbers $z_i - h$ and $z_i + h$ occur $d+1$ times. Here h is a number that is small enough to ensure that there are no other zeros of f or knots from $\boldsymbol{\tau}$ other than z_i in $[z_i - h, z_i + h]$ for $1 \leq i \leq \ell$. Let $\hat{\boldsymbol{c}}$ be the B-spline coefficients of f relative to $\hat{\boldsymbol{\tau}}$. By Lemma 4.24 we then have $S^-(\hat{\boldsymbol{c}}) \leq S^-(\boldsymbol{c})$ so it suffices to prove that $Z_{(\tau_1, \tau_{n+d+1})}(f) \leq S^-(\hat{\boldsymbol{c}})$. But since

$$Z_{(\tau_1, \tau_{n+d+1})}(f) = \sum_{i=1}^{\ell} Z_{(z_i - h, z_i + h)}(f),$$

it suffices to establish the theorem in the following situation: The knot vector is given by

$$\boldsymbol{\tau} = (\overbrace{z-h,\ldots,z-h}^{d+1},\overbrace{z,\ldots,z}^{d-r},\overbrace{z+h,\ldots,z+h}^{d+1})$$

and z is a zero of f of multiplicity r. We want to show that

$$c_j = \frac{(d-r)!}{d!}(-1)^{d+1-j}h^r D^r f(z), \quad j = d+1-r,\ldots,d+1, \tag{10.3}$$

so that the $r+1$ coefficients $(c_j)_{j=d+1-r}^{d+1}$ alternate in sign. For then $S^-(\boldsymbol{c}) \geq r = Z_{(z-h,z+h)}(f)$. Fix j in the range $d+1-r \leq j \leq d+1$. By equation (10.1) we have

$$c_j = \frac{1}{d!}\sum_{k=0}^{r}(-1)^k D^{d-k}\rho_{j,d}(z)D^k f(z) = \frac{(-1)^r}{d!}D^{d-r}\rho_{j,d}(z)D^r f(z),$$

since $D^j f(z) = 0$ for $j = 0\ldots, r-1$. With our special choice of knot vector we have

$$\rho_{j,d}(y) = (y-z+h)^{d+1-j}(y-z)^{d-r}(y-z-h)^{r-d-1+j}.$$

Taking $d-r$ derivatives we therefore obtain

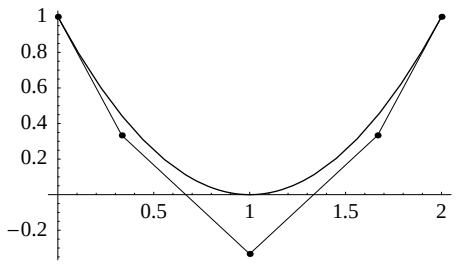
$$D^{d-r}\rho_{j,d}(z) = (d-r)!h^{d+1-j}(-h)^{r-d-1+j} = (d-r)!(-1)^{r-d-1+j}h^r$$

and (10.3) follows. ∎

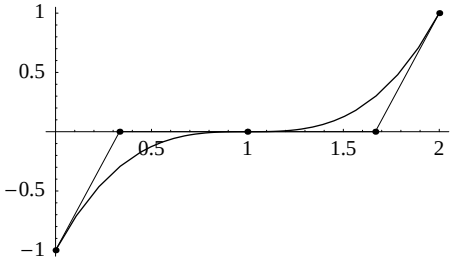
Figures 10.1 (a)–(d) show some examples of splines with multiple zeros of the sort discussed in the proof of Theorem 10.5. All the knot vectors are $d+1$-regular on the interval $[0,2]$, with additional knots at $x=1$. In Figure 10.1 (a) there is one knot at $x=1$ and the spline is the polynomial $(x-1)^2$ which has a double zero at $x=1$. The control polygon models the spline in the normal way and has two sign changes. In Figure 10.1 (b) the knot vector is the same, but the spline is now the polynomial $(x-1)^3$. In this case the multiplicity of the zero is so high that the spline has a splitting point at $x=1$. The construction in the proof of Theorem 10.5 prescribes a knot vector with no knots at $x=1$ in this case. Figure 10.1 (c) shows the polynomial $(x-1)^3$ as a degree 5 spline on a 6-regular knot vector with a double knot at $x=1$. As promised by the theorem and its proof the coefficients change sign exactly three times. The spline in Figure 10.1 (d) is more extreme. It is the polynomial $(x-1)^8$ represented as a spline of degree 9 with one knot at $x=1$. The control polygon has the required 8 changes of sign.

10.2 Uniqueness of spline interpolation

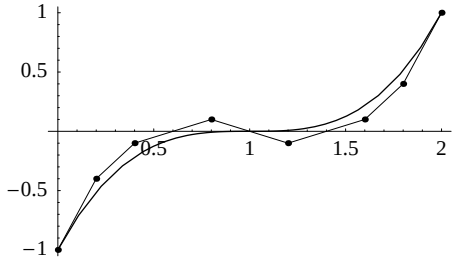
Having established Theorem 10.5, we return to the problem of showing that the B-spline collocation matrix that occurs in spline interpolation, is nonsingular. We first consider Lagrange interpolation, and then turn to Hermite interpolation where we also allow interpolation derivatives.



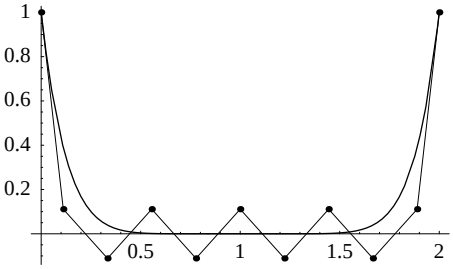
(a) Cubic, 2 zeros, simple knot.



(b) Cubic, multiplicity 3, simple knot.



(c) Degree 5, multiplicity 3, double knot.



(d) Degree 9, multiplicity 8, simple knot.

Figure 10.1. Splines of varying degree with a varying number of zeros at and knots at $x = 1$.

10.2.1 Lagrange Interpolation

In Chapter 8 we studied spline interpolation. With a spline space $\mathbb{S}_{d,\boldsymbol{\tau}}$ of dimension n and data $(y_i)_{i=1}^n$ given at n distinct points $x_1 < x_2 < \cdots < x_n$, the aim is to determine a spline $g = \sum_{i=1}^n c_i B_{i,d}$ in $\mathbb{S}_{d,\boldsymbol{\tau}}$ such that

$$g(x_i) = y_i, \qquad \text{for } i = 1, \ldots, n. \tag{10.4}$$

This leads to the linear system of equations

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{y},$$

where

$$\boldsymbol{A} = \begin{pmatrix} B_{1,d}(x_1) & B_{2,d}(x_1) & \ldots & B_{n,d}(x_1) \\ B_{1,d}(x_2) & B_{2,d}(x_2) & \ldots & B_{n,d}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_{1,d}(x_n) & B_{2,d}(x_n) & \ldots & B_{n,d}(x_n) \end{pmatrix}, \qquad \boldsymbol{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

The matrix $\boldsymbol{A}$ is often referred to as the *B-spline collocation matrix*. Since $B_{i,d}(x)$ is nonzero only if $\tau_i < x < \tau_{i+d+1}$ (we may allow $\tau_i = x$ if $\tau_i = \tau_{i+d} < \tau_{i+d+1}$), the matrix $\boldsymbol{A}$ will in general be sparse. The following theorem tells us exactly when $\boldsymbol{A}$ is nonsingular.

Theorem 10.6. *Let $\mathbb{S}_{d,\boldsymbol{\tau}}$ be a given spline space, and let $x_1 < x_2 < \cdots < x_n$ be n distinct numbers. The collocation matrix $\boldsymbol{A}$ with entries $\big(B_{j,d}(x_i)\big)_{i,j=1}^n$ is nonsingular if and only if its diagonal is positive, i.e.,*

$$B_{i,d}(x_i) > 0 \qquad \text{for } i = 1, \ldots, n. \tag{10.5}$$

Proof. We start by showing that $\boldsymbol{A}$ is singular if a diagonal entry is zero. Suppose that $x_q \le \tau_q$ (strict inequality if $\tau_q = \tau_{q+d} < \tau_{q+d+1}$) for some q so that $B_{q,d}(x_q) = 0$. By the support properties of B-splines we must have $a_{i,j} = 0$ for $i = 1, \ldots, q$ and $j = q, \ldots, n$. But this means that only the $n - q$ last entries of each of the last $n - q + 1$ columns of $\boldsymbol{A}$ can be nonzero; these columns must therefore be linearly dependent and $\boldsymbol{A}$ must be singular. A similar argument shows that $\boldsymbol{A}$ is also singular if $x_q \ge \tau_{q+d+1}$.

To show the converse, suppose that (10.5) holds but $\boldsymbol{A}$ is singular. Then there is a nonzero vector $\boldsymbol{c}$ such that $\boldsymbol{A}\boldsymbol{c} = 0$. Let $f = \sum_{i=1}^n c_i B_{i,d}$ denote the spline with B-spline coefficients $\boldsymbol{c}$. We clearly have $f(x_q) = 0$ for $q = 1, \ldots, n$. Let G denote the set

$$G = \cup_i \big\{(\tau_i, \tau_{i+d+1}) \mid c_i \neq 0\big\}.$$

Since each x in G must be in (τ_i, τ_{i+d+1}) for some i with $c_i \neq 0$, we note that G contains no splitting points of f. Note that if $x_i = \tau_i = \tau_{i+d} < \tau_{i+d+1}$ occurs at a knot of multiplicity $d+1$, then $0 = f(x_i) = c_i$. To complete the proof, suppose first that G is an open interval. Since x_i is in G if $c_i \neq 0$, the number of zeros of f in G is greater than or equal to the number ℓ of nonzero coefficients in $\boldsymbol{c}$. Since we also have $S^-(\boldsymbol{c}) < \ell \le Z_G(f)$, we have a contradiction to Theorem 10.5. In general G consists of several subintervals which means that f is not connected, but can be written as a sum of connected components, each defined on one of the subintervals. The above argument then leads to a contradiction on each subinterval, and hence we conclude that $\boldsymbol{A}$ is nonsingular. ∎

Theorem 10.6 makes it simple to ensure that the collocation matrix is nonsingular. We just place the knots and interpolation points in such a way that $\tau_i < x_i < \tau_{i+d+1}$ for $i = 1, \ldots, n$ (note again that if $\tau_i = \tau_{i+d} < \tau_{i+d+1}$, then $x_i = \tau_i$ is allowed).

10.2.2 Hermite Interpolation

In earlier chapters, particularly in Chapter 8, we made use of polynomial interpolation with Hermite data—data based on derivatives as well as function values. This is also of interest for splines, and as for polynomials this is conveniently indicated by allowing the interpolation point to coalesce. If for example $x_1 = x_2 = x_3 = x$, we take x_1 to signify interpolation of function value at x, the second occurrence of x signifies interpolation of first derivative, and the third tells us to interpolate second derivative at x. If we introduce the notation

$$\lambda_{\boldsymbol{x}}(i) = \max_j\{j \mid x_{i-j} = x_i\}$$

and assume that the interpolation points are given in nondecreasing order as $x_1 \le x_2 \le \cdots \le x_n$, then the interpolation conditions are

$$D^{\lambda_{\boldsymbol{x}}(i)}g(x_i) = D^{\lambda_{\boldsymbol{x}}(i)}f(x_i) \tag{10.6}$$

where f is a given function and g is the spline to be determined. Since we are dealing with splines of degree d we cannot interpolate derivatives of higher order than d; we therefore assume that $x_i < x_{i+d+1}$ for $i = 1, \ldots, n-d-1$. At a point of discontinuity (10.6) is to be interpreted according to our usual convention of taking limits from the right. The (i,j)-entry of the collocation matrix $\boldsymbol{A}$ is now given by

$$a_{i,j} = D^{\lambda_{\boldsymbol{x}}(i)}B_{j,d}(x_i),$$

and as before the interpolation problem is generally solvable if and only if the collocation matrix is nonsingular. Also as before, it turns out that the collocation matrix is nonsingular if and only if $\tau_i \le x_i < \tau_{i+d+1}$, where equality is allowed in the first inequality only if $D^{\lambda_{\boldsymbol{x}}(i)}B_{i,d}(x_i) \neq 0$. This result will follow as a special case of our next theorem where we consider an even more general situation.

At times it is of interest to know exactly when a submatrix of the collocation matrix is nonsingular. The submatrices we consider are obtained by removing the same number of rows and columns from $\boldsymbol{A}$. Any columns may be removed, or equivalently, we consider a subset $\{B_{j_1,d}, \ldots, B_{j_\ell,d}\}$ of the B-splines. When removing rows we have to be a bit more careful. The convention is that if a row with derivatives of order r at z is included, then we also include all the lower order derivatives at z. This is most easily formulated by letting the sequence of interpolation points only contain ℓ points as in the following theorem.

Theorem 10.7. *Let $\mathbb{S}_{d,\boldsymbol{\tau}}$ be a spline space and let $\{B_{j_1,d}, \ldots, B_{j_\ell,d}\}$ be a subsequence of its B-splines. Let $x_1 \le \cdots \le x_\ell$ be a sequence of interpolation points with $x_i \le x_{i+d+1}$ for $i = 1, \ldots, \ell-d-1$. Then the $\ell \times \ell$ matrix $\boldsymbol{A}(\boldsymbol{j})$ with entries given by*

$$a_{i,q} = D^{\lambda_{\boldsymbol{x}}(i)}B_{j_q,d}(x_i)$$

for $i = 1, \ldots, \ell$ and $q = 1, \ldots, \ell$ is nonsingular if and only if

$$\tau_{j_i} \le x_i < \tau_{j_i+d+1}, \qquad \text{for } i = 1, \ldots, \ell, \tag{10.7}$$

where equality is allowed in the first inequality if $D^{\lambda_{\boldsymbol{x}}(i)}B_{j_i,d}(x_i) \neq 0$.

Proof. The proof follows along the same lines as the proof of Theorem 10.6. The most challenging part is the proof that condition (10.7) is necessary so we focus on this. Suppose that (10.7) holds, but $\boldsymbol{A}(\boldsymbol{j})$ is singular. Then we can find a nonzero vector $\boldsymbol{c}$ such that $\boldsymbol{A}(\boldsymbol{j})\boldsymbol{c} = \boldsymbol{0}$. Let $f = \sum_{i=1}^{\ell} c_i B_{j_i,d}$ denote the spline with $\boldsymbol{c}$ as its B-spline coefficients, and let G denote the set

$$G = \cup_{i=1}^{\ell} \{(\tau_{j_i}, \tau_{j_i+d+1}) \mid c_i \neq 0\}.$$

To carry through the argument of Theorem 10.6 we need to verify that in the exceptional case where $x_i = \tau_{j_i}$ then $c_i = 0$.

Set $r = \lambda_{\boldsymbol{x}}(i)$ and suppose that the knot τ_{j_i} occurs m times in $\boldsymbol{\tau}$ and that $\tau_{j_i} = x_i$ so $D^r B_{j_i,d}(x_i) \neq 0$. In other words

$$\tau_\mu < x_i = \tau_{\mu+1} = \cdots = \tau_{\mu+m} < \tau_{\mu+m+1}$$

for some integer μ, and in addition $j_i = \mu + k$ for some integer k with $1 \leq k \leq m$. Note that f satisfies

$$f(x_i) = Df(x_i) = \cdots = D^r f(x_i) = 0.$$

(Remember that if a derivative is discontinuous at x_i we take limits from the right.) Recall from Lemma 2.6 that all B-splines have continuous derivatives up to order $d - m$ at x_i. Since $D^r B_{j_i}$ clearly is discontinuous at x_i, it must be true that $r > d-m$. We therefore have $f(x_i) = Df(x_i) = \cdots = D^{d-m} f(x_i) = 0$ and hence $c_{\mu+m-d} = \cdots = c_\mu = 0$ by Lemma 10.2. The remaining interpolation conditions at x_i are $D^{d-m+1} f(x_i) = D^{d-m+2} f(x_i) = \cdots = D^r f(x_i) = 0$. Let us consider each of these in turn. By the continuity properties of B-splines we have $D^{d-m+1} B_{\mu+1}(x_i) \neq 0$ and $D^{d-m+1} B_{\mu+\nu} = 0$ for $\nu > 1$. This means that

$$0 = D^{d-m+1} f(x_i) = c_{\mu+1} D^{d-m+1} B_{\mu+1}(x_i)$$

and $c_{\mu+1} = 0$. Similarly, we also have

$$0 = D^{d-m+2} f(x_i) = c_{\mu+2} D^{d-m+2} B_{\mu+2}(x_i),$$

and hence $c_{\mu+2} = 0$ since $D^{d-m+2} B_{\mu+2}(x_i) \neq 0$. Continuing this process we find

$$0 = D^r f(x_i) = c_{\mu+r+m-d} D^r B_{\mu+r+m-d}(x_i),$$

so $c_{\mu+r+m-d} = 0$ since $D^r B_{\mu+r+m-d}(x_i) \neq 0$. This argument also shows that j_i cannot be chosen independently of r; we must have $j_i = \mu + r + m - d$.

For the rest of the proof it is sufficient to consider the case where G is an open interval, just as in the proof of Theorem 10.6. Having established that $c_i = 0$ if $x_i = \tau_{j_i}$, we know that if $c_i \neq 0$ then $x_i \in G$. The number of zeros of f in G (counting multiplicities) is therefore greater than or equal to the number of nonzero coefficients. But this is impossible according to Theorem 10.5. ∎

10.3 Total positivity

In this section we are going to deduce another interesting property of the knot insertion matrix and the B-spline collocation matrix, namely that they are totally positive. We follow the same strategy as before and establish this first for the knot insertion matrix and then obtain the total positivity of the collocation matrix by recognising it as a submatrix of a knot insertion matrix.

Definition 10.8. *A matrix $\boldsymbol{A}$ in $\mathbb{R}^{m,n}$ is said to be totally positive if all its square submatrices have nonnegative determinant. More formally, let $\boldsymbol{i} = (i_1, i_2, \ldots, i_\ell)$ and $\boldsymbol{j} = (j_1, j_2, \ldots, j_\ell)$ be two integer sequences such that*

$$1 \le i_1 < i_2 < \cdots < i_\ell \le m, \tag{10.8}$$

$$1 \le i_1 < i_2 < \cdots < i_\ell \le n, \tag{10.9}$$

and let $\boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j})$ denote the submatrix of $\boldsymbol{A}$ with entries $(a_{i_p, j_q})_{p,q=1}^{\ell}$. Then $\boldsymbol{A}$ is totally positive if $\det \boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j}) \ge 0$ for all sequences $\boldsymbol{i}$ and $\boldsymbol{j}$ on the form (10.8) and (10.9), for all ℓ with $1 \le \ell \le \min\{m, n\}$.

We first show that knot insertion matrices are totally positive.

Theorem 10.9. *Let $\boldsymbol{\tau}$ and $\boldsymbol{t}$ be two knot vectors with $\boldsymbol{\tau} \subseteq \boldsymbol{t}$. Then the knot insertion matrix from $\mathbb{S}_{d,\boldsymbol{\tau}}$ to $\mathbb{S}_{d,\boldsymbol{t}}$ is totally positive.*

Proof. Suppose that there are k more knots in $\boldsymbol{t}$ than in $\boldsymbol{\tau}$; our proof is by induction on k. We first note that if $k = 0$, then $\boldsymbol{A} = I$, the identity matrix, while if $k = 1$, then $\boldsymbol{A}$ is a bi-diagonal matrix with one more rows than columns. Let us denote the entries of $\boldsymbol{A}$ by $\big(\alpha_j(i)\big)_{i,j=1}^{n+1,n}$ (if $k = 0$ the range of i is 1, ..., n). In either case all the entries are nonnegative and $\alpha_j(i) = 0$ for $j < i - 1$ and $j > i$. Consider now the determinant of $\boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j})$. If $j_\ell \ge i_\ell$ then $j_\ell > i_q$ for $q = 1, \ldots, \ell - 1$ so $\alpha_{j_\ell}(i_q) = 0$ for $q < \ell$. This means that only the last entry of the last column of $\boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j})$ is nonzero. The other possibility is that $j_\ell \le i_\ell - 1$ so that $j_q < i_\ell - 1$ for $q < \ell$. Then $\alpha_{j_q}(i_\ell) = 0$ for $q < \ell$ so only the last entry of the last row of $\boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j})$ is nonzero. Expanding the determinant either by the last column or last row we therefore have $\det \boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j}) = \alpha_{j_\ell}(i_\ell) \det \boldsymbol{A}(\boldsymbol{i}', \boldsymbol{j}')$ where $\boldsymbol{i}' = (i_1, \ldots, i_{\ell-1})$ and $\boldsymbol{j}' = (j_1, \ldots, j_{\ell-1})$. Continuing this process we find that

$$\det \boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j}) = \alpha_{j_1}(i_1)\alpha_{j_2}(i_2) \cdots \alpha_{j_\ell}(i_\ell)$$

which clearly is nonnegative.

For $k \ge 2$, we make use of the factorization

$$\boldsymbol{A} = \boldsymbol{A}_k \cdots \boldsymbol{A}_1 = \boldsymbol{A}_k \boldsymbol{B}, \tag{10.10}$$

where each $\boldsymbol{A}_r$ corresponds to insertion of one knot and $\boldsymbol{B} = \boldsymbol{A}_{k-1} \cdots \boldsymbol{A}_1$ is the knot insertion matrix for inserting $k - 1$ of the knots. By the induction hypothesis we know that both $\boldsymbol{A}_k$ and $\boldsymbol{B}$ are totally positive; we must show that $\boldsymbol{A}$ is totally positive. Let $(\boldsymbol{a}_i)$ and $(\boldsymbol{b}_i)$ denote the rows of $\boldsymbol{A}$ and $\boldsymbol{B}$, and let $\big(\alpha_j(i)\big)_{i,j=1}^{m,m-1}$ denote the entries of $\boldsymbol{A}_k$. From (10.10) we have

$$\boldsymbol{a}_i = \alpha_{i-1}(i)\boldsymbol{b}_{i-1} + \alpha_i(i)\boldsymbol{b}_i \qquad \text{for } i = 1, \ldots, m,$$

where $\alpha_0(1) = \alpha_m(m) = 0$. Let $\boldsymbol{a}_i(\boldsymbol{j})$ and $\boldsymbol{b}_i(\boldsymbol{j})$ denote the vectors obtained by keeping only entries $(j_q)_{q=1}^{\ell}$ of $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ respectively. Row q of $\boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j})$ of $\boldsymbol{A}$ is then given by

$$\boldsymbol{a}_{i_q}(\boldsymbol{j}) = \alpha_{i_q - 1}(i_q)\boldsymbol{b}_{i_q - 1}(\boldsymbol{j}) + \alpha_{i_q}(i_q)\boldsymbol{b}_{i_q}(\boldsymbol{j}).$$

Using the linearity of the determinant in row q we therefore have

$$\det \begin{pmatrix} \boldsymbol{a}_{i_1}(\boldsymbol{j}) \\ \vdots \\ \boldsymbol{a}_{i_q}(\boldsymbol{j}) \\ \vdots \\ \boldsymbol{a}_{i_\ell}(\boldsymbol{j}) \end{pmatrix} = \det \begin{pmatrix} \boldsymbol{a}_{i_1}(\boldsymbol{j}) \\ \vdots \\ \alpha_{i_q-1}(i_q)\boldsymbol{b}_{i_q-1}(\boldsymbol{j}) + \alpha_{i_q}(i_q)\boldsymbol{b}_{i_q}(\boldsymbol{j}) \\ \vdots \\ \boldsymbol{a}_{i_\ell}(\boldsymbol{j}) \end{pmatrix}$$

$$= \alpha_{i_q-1}(i_q) \det \begin{pmatrix} \boldsymbol{a}_{i_1}(\boldsymbol{j}) \\ \vdots \\ \boldsymbol{b}_{i_q-1}(\boldsymbol{j}) \\ \vdots \\ \boldsymbol{a}_{i_\ell}(\boldsymbol{j}) \end{pmatrix} + \alpha_{i_q}(i_q) \det \begin{pmatrix} \boldsymbol{a}_{i_1}(\boldsymbol{j}) \\ \vdots \\ \boldsymbol{b}_{i_q}(\boldsymbol{j}) \\ \vdots \\ \boldsymbol{a}_{i_\ell}(\boldsymbol{j}) \end{pmatrix}.$$

By expanding the other rows similarly we find that $\det \boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j})$ can be written as a sum of determinants of submatrices of $\boldsymbol{B}$, multiplied by products of $\alpha_j(i)$'s. By the induction hypothesis all these quantities are nonnegative, so the determinant of $\boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j})$ must also be nonnegative. Hence $\boldsymbol{A}$ is totally positive. ∎

Knowing that the knot insertion matrix is totally positive, we can prove a similar property of the B-spline collocation matrix, even in the case where multiple collocation points are allowed.

Theorem 10.10. *Let $\mathbb{S}_{d,\boldsymbol{\tau}}$ be a spline space and let $\{B_{j_1,d}, \ldots, B_{j_\ell,d}\}$ be a subsequence of its B-splines. Let $x_1 \leq \cdots \leq x_\ell$ be a sequence of interpolation points with $x_i \leq x_{i+d+1}$ for $i = 1, \ldots, \ell - d - 1$, and denote by $\boldsymbol{A}(\boldsymbol{j})$ the $\ell \times \ell$ matrix with entries given by*

$$a_{i,q} = D^{\lambda_{\boldsymbol{x}}(i)} B_{j_q,d}(x_i)$$

for $i = 1, \ldots, \ell$ and $q = 1, \ldots, \ell$. Then

$$\det \boldsymbol{A}(\boldsymbol{j}) \geq 0.$$

Proof. We first prove the claim in the case $x_1 < x_2 < \cdots < x_\ell$. By inserting knots of multiplicity $d + 1$ at each of $(x_i)_{i=1}^{\ell}$ we obtain a knot vector $\boldsymbol{t}$ that contains $\boldsymbol{\tau}$ as a subsequence. If $t_{i-1} < t_i = t_{i+d} < t_{i+d+1}$ we know from Lemma 2.6 that $B_{j,d,\boldsymbol{\tau}}(t_i) = \alpha_{j,d}(i)$. This means that the matrix $\boldsymbol{A}(\boldsymbol{j})$ appears as a submatrix of the knot insertion matrix from $\boldsymbol{\tau}$ to $\boldsymbol{t}$. It therefore follows from Theorem 10.9 that $\det \boldsymbol{A}(\boldsymbol{j}) \geq 0$ in this case.

To prove the theorem in the general case we consider a set of distinct collocation points $y_1 < \cdots < y_\ell$ and let $\boldsymbol{A}(\boldsymbol{j}, \boldsymbol{y})$ denote the corresponding collocation matrix. Set $\lambda^i = \lambda_{\boldsymbol{x}}(i)$ and let ρ_i denote the linear functional given by

$$\rho_i f = \lambda^i! \, [y_{i-\lambda^i}, \ldots, y_i] f \tag{10.11}$$

for $i = 1, \ldots, \ell$. Here $[\cdot, \ldots, \cdot] f$ is the divided difference of f. By standard properties of divided differences we have

$$\rho_i B_{j,d} = \sum_{s=i-\lambda^i}^{i} \gamma_{i,s} B_{j,d}(y_s)$$

and $\gamma_{i,i} > 0$. Denoting by $\boldsymbol{D}$ the matrix with (i,j)-entry $\rho_i B_{j,d}$, we find by properties of determinants and (10.11) that

$$\det \boldsymbol{D} = \gamma_{1,1} \cdots \gamma_{\ell,\ell} \det \boldsymbol{A}(\boldsymbol{j}, \boldsymbol{y}).$$

If we now let $\boldsymbol{y}$ tend to $\boldsymbol{x}$ we know from properties of the divided difference functional that $\rho_i B_j$ tends to $D^{\lambda^i} B_j$ in the limit. Hence $\boldsymbol{D}$ tends to $\boldsymbol{A}(\boldsymbol{j})$ so $\det \boldsymbol{A}(\boldsymbol{j}) \geq 0$. ∎

APPENDIX A

Some Linear Algebra

A.1 Matrices

The collection of m, n matrices

$$\boldsymbol{A} = \begin{pmatrix} a_{1,1}, & \dots & , a_{1,n} \\ \cdots & & \cdots \\ a_{m,1}, & \dots & , a_{m,n} \end{pmatrix}$$

with real elements $a_{i,j}$ is denoted by $\mathbb{R}^{m,n}$. If $n = 1$ then $\boldsymbol{A}$ is called a column vector. Similarly, if $m = 1$ then $\boldsymbol{A}$ is a row vector. We let $\mathbb{R}^m$ denote the collection of all column or row vectors with m real components.

A.1.1 Nonsingular matrices, and inverses.

Definition A.1. *A collection of vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n \in \mathbb{R}^m$ is* linearly independent *if $x_1\boldsymbol{a}_1 + \cdots + x_n\boldsymbol{a}_n = \boldsymbol{0}$ for some real numbers $x_1, \dots, x_n$, implies that $x_1 = \cdots = x_n = 0$.*

Suppose $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ are the columns of a matrix $\boldsymbol{A} \in \mathbb{R}^{m,n}$. For a vector $\boldsymbol{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ we have $\boldsymbol{Ax} = \sum_{j=1}^n x_j \boldsymbol{a}_j$. It follows that the collection $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ is linearly independent if and only if $\boldsymbol{Ax} = \boldsymbol{0}$ implies $\boldsymbol{x} = \boldsymbol{0}$.

Definition A.2. *A square matrix $\boldsymbol{A}$ such that $\boldsymbol{Ax} = \boldsymbol{0}$ implies $\boldsymbol{x} = \boldsymbol{0}$ is said to be nonsingular.*

Definition A.3. *A square matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is said to be invertible if for some $\boldsymbol{B} \in \mathbb{R}^{n,n}$*

$$\boldsymbol{BA} = \boldsymbol{AB} = \boldsymbol{I},$$

where $\boldsymbol{I} \in \mathbb{R}^{n,n}$ is the identity matrix.

An invertible matrix $\boldsymbol{A}$ has a unique inverse $\boldsymbol{B} = \boldsymbol{A}^{-1}$. If $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$ are square matrices, and $\boldsymbol{A} = \boldsymbol{BC}$, then $\boldsymbol{A}$ is invertible if and only if both $\boldsymbol{B}$ and $\boldsymbol{C}$ are also invertible. Moreover, the inverse of $\boldsymbol{A}$ is the product of the inverses of $\boldsymbol{B}$ and $\boldsymbol{C}$ in reverse order, $\boldsymbol{A}^{-1} = \boldsymbol{C}^{-1}\boldsymbol{B}^{-1}$.

A.1.2 Determinants.

The determinant of a square matrix $\boldsymbol{A}$ will be denoted $\det(\boldsymbol{A})$ or

$$\begin{vmatrix} a_{1,1}, & \dots & , a_{1,n} \\ \vdots & & \vdots \\ a_{n,1}, & \dots & , a_{n,n} \end{vmatrix}.$$

Recall that the determinant of a 2×2 matrix is

$$\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}.$$

A.1.3 Criteria for nonsingularity and singularity.

We state without proof the following criteria for nonsingularity.

Theorem A.4. *The following is equivalent for a square matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$.*

1. *$\boldsymbol{A}$ is nonsingular.*
2. *$\boldsymbol{A}$ is invertible.*
3. *$\boldsymbol{Ax} = \boldsymbol{b}$ has a unique solution $\boldsymbol{x} = \boldsymbol{A}^{-1}b$ for any $\boldsymbol{b} \in \mathbb{R}^n$.*
4. *$\boldsymbol{A}$ has linearly independent columns.*
5. *$\boldsymbol{A}^T$ is nonsingular.*
6. *$\boldsymbol{A}$ has linearly independent rows.*
7. $\det(\boldsymbol{A}) \neq \boldsymbol{0}$.

We also have a number of criteria for a matrix to be singular.

Theorem A.5. *The following is equivalent for a square matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$.*

1. *There is a nonzero $\boldsymbol{x} \in \mathbb{R}^n$ so that $\boldsymbol{Ax} = \boldsymbol{0}$.*
2. *$\boldsymbol{A}$ has no inverse.*
3. *$\boldsymbol{Ax} = \boldsymbol{b}$ has either no solution or an infinite number of solutions.*
4. *$\boldsymbol{A}$ has linearly dependent columns.*
5. *There is a nonzero $\boldsymbol{x}$ so that $\boldsymbol{x}^T\boldsymbol{A} = \boldsymbol{0}$.*
6. *$\boldsymbol{A}$ has linearly dependent rows.*
7. $\det(\boldsymbol{A}) = \boldsymbol{0}$.

Corollary A.6. *A matrix with more columns than rows has linearly dependent columns.*

Proof. Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$ with $n > m$. By adding $n - m$ rows of zeros to $\boldsymbol{A}$ we obtain a square matrix $\boldsymbol{B} \in \mathbb{R}^{n,n}$. This matrix has linearly dependent rows. By Theorem A.4 the matrix $\boldsymbol{B}$ has linearly dependent columns. But then the columns of $\boldsymbol{A}$ are also linearly dependent. ∎

A.2 Vector Norms

Formally, a *vector norm* $||\ ||=||\boldsymbol{x}||$, is a function $\|\ \ \|:\mathbb{R}^n\to[0,\infty)$ that satisfies for $\boldsymbol{x},\boldsymbol{y},\in\mathbb{R}^n$, and $\alpha\in\mathbb{R}$ the following properties

$$\begin{array}{ll}1. & ||\boldsymbol{x}||=0 \quad \text{implies} \quad \boldsymbol{x}=\boldsymbol{0}.\\ 2. & ||\alpha\boldsymbol{x}||=|\alpha|||\boldsymbol{x}||.\\ 3. & ||\boldsymbol{x}+\boldsymbol{y}||\le||\boldsymbol{x}||+||\boldsymbol{y}||.\end{array} \tag{A.1}$$

Property 3 is known as the *Triangle Inequality*. For us the most useful class of norms are the p or ℓ^p norms. They are defined for $p\ge1$ and $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)^T\in\mathbb{R}^n$ by

$$\begin{array}{ll}||\boldsymbol{x}||_p & =(|x_1|^p+|x_2|^p+\cdots+|x_n|^p)^{1/p}.\\ ||\boldsymbol{x}||_\infty & =\max_i|x_i|.\end{array} \tag{A.2}$$

Since

$$||\boldsymbol{x}||_\infty\le||\boldsymbol{x}||_p\le n^{1/p}||\boldsymbol{x}||_\infty,\qquad p\ge1 \tag{A.3}$$

and $\lim_{p\to\infty}n^{1/p}=1$ for any $n\in\mathbb{N}$ we see that $\lim_{p\to\infty}||\boldsymbol{x}||_p=||\boldsymbol{x}||_\infty$.

The 1,2, and ∞ norms are the most important. We have

$$||\boldsymbol{x}||_2^2=x_1^2+\cdots+x_n^2=\boldsymbol{x}^T\boldsymbol{x}. \tag{A.4}$$

Lemma A.7 (The Hölder inequality). *We have for* $1\le p\le\infty$ *and* $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}$

$$\sum_{i=1}^n|x_iy_i|\le||\boldsymbol{x}||_p||\boldsymbol{y}||_q,\quad \textit{where}\quad \frac{1}{p}+\frac{1}{q}=1. \tag{A.5}$$

Proof. We base the proof on properties of the exponential function. Recall that the exponential function is convex, i.e. with $f(x)=e^x$ we have the inequality

$$f(\lambda x+(1-\lambda)y)\le\lambda f(x)+(1-\lambda)f(y) \tag{A.6}$$

for every $\lambda\in[0,1]$ and $x,y\in\mathbb{R}$.

If $\boldsymbol{x}=\boldsymbol{0}$ or $\boldsymbol{y}=\boldsymbol{0}$, there is nothing to prove. Suppose $\boldsymbol{x},\boldsymbol{y}\neq\boldsymbol{0}$. Define $\boldsymbol{u}=\boldsymbol{x}/||\boldsymbol{x}||_p$ and $\boldsymbol{v}=\boldsymbol{y}/||\boldsymbol{y}||_q$. Then $||\boldsymbol{u}||_p=||\boldsymbol{v}||_q=1$. If we can prove that $\sum_i|u_iv_i|\le1$, we are done because then $\sum_i|x_iy_i|=||\boldsymbol{x}||_p||\boldsymbol{y}||_q\sum_i|u_iv_i|\le||\boldsymbol{x}||_p||\boldsymbol{y}||_q$. Since $|u_iv_i|=|u_i||v_i|$, we can assume that $u_i\ge0$ and $v_i\ge0$. Moreover, we can assume that $u_i>0$ and $v_i>0$ because a zero term contributes no more to the left hand side than to the right hand side of (A.5). Let s_i, t_i be such that $u_i=e^{s_i/p}$, $v_i=e^{t_i/q}$. Taking $f(x)=e^x,\lambda=1/p$, $1-\lambda=1/q$, $x=s_i$ and $y=t_i$ in (A.6) we find

$$e^{s_i/p+t_i/q}\le\frac{1}{p}e^{s_i}+\frac{1}{q}e^{t_i}.$$

But then

$$\sum_i|u_iv_i|=\sum_ie^{s_i/p+t_i/q}\le\frac{1}{p}\sum_ie^{s_i}+\frac{1}{q}\sum_ie^{t_i}=\frac{1}{p}\sum_iu_i^p+\frac{1}{q}\sum_iv_i^q=\frac{1}{p}+\frac{1}{q}=1.$$

This completes the proof of (A.5). ∎

When $p = 2$ then $q = 2$ and the Hölder inequality is associated with the names Buniakowski-Cauchy-Schwarz.

Lemma A.8 (The Minkowski inequality). *We have for* $1 \leq p \leq \infty$ *and* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}$

$$||\boldsymbol{x} + \boldsymbol{y}||_p \leq ||\boldsymbol{x}||_p + ||\boldsymbol{y}||_p. \tag{A.7}$$

Proof. Let $\boldsymbol{u} = (u_1, \ldots, u_n)$ with $u_i = |x_i + y_i|^{p-1}$. Since $q(p-1) = p$ and $p/q = p-1$, we find

$$||\boldsymbol{u}||_q = (\sum_i |x_i + y_i|^{q(p-1)})^{1/q} = (\sum_i |x_i + y_i|^p)^{1/q} = ||\boldsymbol{x} + \boldsymbol{y}||_p^{p/q} = ||\boldsymbol{x} + \boldsymbol{y}||_p^{p-1}.$$

Using this and the Hölder inequality we obtain

$$\begin{aligned} ||\boldsymbol{x} + \boldsymbol{y}||_p^p &= \sum_i |x_i + y_i|^p \leq \sum_i |u_i||x_i| + \sum_i |u_i||y_i| \leq (||\boldsymbol{x}||_p + ||\boldsymbol{y}||_p)||\boldsymbol{u}||_q \\ &\leq (||\boldsymbol{x}||_p + ||\boldsymbol{y}||_p)||\boldsymbol{x} + \boldsymbol{y}||_p^{p-1}. \end{aligned}$$

Dividing by $||\boldsymbol{x} + \boldsymbol{y}||_p^{p-1}$ proves Minkowski. ∎

Using the Minkowski inequality it follows that the p norms satisfies the axioms for a vector norm.

In (A.3) we established the inequality

$$||\boldsymbol{x}||_\infty \leq ||\boldsymbol{x}||_p \leq n^{1/p}||\boldsymbol{x}||_\infty, \quad p \geq 1.$$

More generally, we say that two vector norms $||\ ||$ and $||\ ||'$ are *equivalent* if there exists positive constants μ and M such that

$$\mu||\boldsymbol{x}|| \leq ||\boldsymbol{x}||' \leq M||\boldsymbol{x}|| \tag{A.8}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$.

Theorem A.9. *All vector norms on* $\mathbb{R}^n$ *are equivalent.*

Proof. It is enough to show that a vector norm $||\ ||$ is equivalent to the l_∞ norm, $||\ ||_\infty$. Let $\boldsymbol{x} \in \mathbb{R}^n$ and let $\boldsymbol{e}_i, i = 1, \ldots, n$ be the unit vectors in $\mathbb{R}^n$. Writing $\boldsymbol{x} = x_1\boldsymbol{e}_1 + \cdots + x_n\boldsymbol{e}_n$ we have

$$||\boldsymbol{x}|| \leq \sum_i |x_i|||\boldsymbol{e}_i|| \leq ||\boldsymbol{x}||_\infty M, \quad M = \sum_i ||\boldsymbol{e}_i||.$$

To find $\mu > 0$ such that $||\boldsymbol{x}|| \geq \mu||\boldsymbol{x}||_\infty$ for all $\boldsymbol{x} \in \mathbb{R}^n$ is less elementary. Consider the function f given by $f(x) = ||\boldsymbol{x}||$ defined on the l_∞ "unit ball"

$$S = \{\boldsymbol{x} \in \mathbb{R}^n : ||\boldsymbol{x}||_\infty = 1\}.$$

S is a closed and bounded set. From the inverse triangle inequality

$$|\,||\boldsymbol{x}|| - ||\boldsymbol{y}||\,| \leq ||\boldsymbol{x} - \boldsymbol{y}||, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

it follows that f is continuous on S. But then f attains its maximum and minimum on S, i.e. there is a point $\boldsymbol{x}^* \in S$ such that

$$||\boldsymbol{x}^*|| = \min_{\boldsymbol{x}\in S} ||\boldsymbol{x}||.$$

Moreover, since $\boldsymbol{x}^*$ is nonzero we have $\mu := ||\boldsymbol{x}^*|| > 0$. If $\boldsymbol{x} \in \mathbb{R}^n$ is nonzero then $\boldsymbol{x} = \boldsymbol{x}/||\boldsymbol{x}||_\infty \in S$. Thus

$$\mu \le ||\boldsymbol{x}|| = ||\frac{\boldsymbol{x}}{||\boldsymbol{x}||_\infty}|| = \frac{1}{||\boldsymbol{x}||_\infty}||\boldsymbol{x}||,$$

and this establishes the lower inequality. ∎

It can be shown that for the p norms we have for any q with $1 \le q \le p \le \infty$

$$||\boldsymbol{x}||_p \le ||\boldsymbol{x}||_q \le n^{1/q-1/p}||\boldsymbol{x}||_p, \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{A.9}$$

<

A.3 Vector spaces of functions

In $\mathbb{R}^m$ we have the operations $\boldsymbol{x} + \boldsymbol{y}$ and $a\boldsymbol{x}$ of vector addition and multiplication by a scalar $a \in \mathbb{R}$. Such operations can also be defined for functions. As an example, if $f(x) = x$, $g(x) = 1$, and a, b are real numbers then $af(x) + bg(x) = ax + b$. In general, if f and g are two functions defined on the same set I and $a \in \mathbb{R}$, then the sum $f + g$ and the product af are functions defined on I by

$$(f + g)(x) = f(x) + g(x),$$
$$(af(x) = af(x).$$

Two functions f and g defined on I are equal if $f(x) = g(x)$ for all $x \in I$. We say that f is the zero function, i.e. $f = 0$, if $f(x) = 0$ for all $x \in I$.

Definition A.10. *Suppose S is a collection of real valued or vector valued functions, all defined on the same set I. The collection S is called a vector space if $af + bg \in S$ for all $f, g \in S$ and all $a, b \in \mathbb{R}$. A subset T of S is called a subspace of S if T itself is a vector space.*

Example A.11. Vector spaces

- All polynomials π_d of degree at most d.
- All polynomials of all degrees.
- All trigonometric polynomials $a_0 + \sum_{k=1}^{d}(a_k \cos kx + b_k \sin kx$ of degree at most d.
- The set $C(I)$ of all continuous real valued functions defined on I.
- The set $C^r(I)$ of all real valued functions defined on I with continuous j'th derivative for $j = 0, 1, \ldots, r$.

Definition A.12. *A vector space S is said to be* finite dimesional *if*

$$S = \text{span}(\phi_1, \ldots, \phi_n) = \{\sum_{j=1}^{n} c_j \phi_j : c_j \in \mathbb{R}\},$$

for a finite number of functions $\phi_1, \ldots, \phi_n$ in S. The functions $\phi_1, \ldots, \phi_n$ are said to span *or* generate *S.*

Of the examples above the space $\pi_d = \text{span}(1, x, x^2, \ldots x^d)$ generated by the monomials $1, x, x^2, \ldots x^d$ is finite dimensional. Also the trigonometric polynomials are finite dimensional. The space of all polynomials of all degrees is not finite dimensional. To see this we observe that any finite set cannot generate the monomial x^{d+1} where d is the maximal degree of the elements in the spanning set. Finally we observe that $C(I)$ and $C^r(I)$ contain the space of polynomials of all degrees as a subspace. Hence they are not finite dimensional,

If $f \in S = \text{span}(\phi_1, \ldots, \phi_n)$ then $f = \sum_{j=1}^{n} c_j \phi_j$ for some $\boldsymbol{c} = (c_1, \ldots, c_n)$. With $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ we will often use the vector notation

$$f(x) = \boldsymbol{\phi}(x)^T \boldsymbol{c} \tag{A.10}$$

for f.

A.3.1 Linear independence and bases

All vector spaces in this section will be finite dimensional.

Definition A.13. *A set of functions $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ in a vector space S is said to be* linearly independent *on a subset J of I if $\boldsymbol{\phi}(x)^T \boldsymbol{c} = c_1\phi_1(x) + \cdots + c_n\phi_n(x) = 0$ for all $x \in J$ implies that $\boldsymbol{c} = \boldsymbol{0}$. If $J = I$ then we simply say that $\boldsymbol{\phi}$ is linearly independent.*

If $\boldsymbol{\phi}$ is linearly independent then the representation in (A.10) is unique. For if $f = \boldsymbol{\phi}^T \boldsymbol{c} = \boldsymbol{\phi}^T \boldsymbol{b}$ for some $\boldsymbol{c}, \boldsymbol{b} \in \mathbb{R}^n$ then $f = \boldsymbol{\phi}^T(\boldsymbol{c} - \boldsymbol{b}) = \boldsymbol{0}$. Since $\boldsymbol{\phi}$ is linearly independent we have $\boldsymbol{c} - \boldsymbol{b} = \boldsymbol{0}$, or $\boldsymbol{c} = \boldsymbol{b}$.

Definition A.14. *A set of functions $\boldsymbol{\phi}^T = (\phi_1, \ldots, \phi_n)$ in a vector space S is* a basis for *S if the following two conditions hold*

1. *$\boldsymbol{\phi}$ is linearly independent.*
2. *$S = \text{span}(\boldsymbol{\phi})$.*

Theorem A.15. *The monomials $1, x, x^2, \ldots x^d$ are linearly independent on any set $J \subset \mathbb{R}$ containing at least $d+1$ distinct points. In particular these functions form as basis for π_d.*

Proof. Let $x_0, \ldots, x_d$ be $d+1$ distinct points in J, and let $p(x) = c_0 + c_1 x + \cdots + c_d x^d = 0$ for all $x \in J$. Then $p(x_i) = 0$, for $i = 0, 1, \ldots, d$. Since a nonzero polynomial of degree d can have at most d zeros we conclude that p must be the zero polynomial. But then $c_k = p^{(k)}(0)/k! = 0$ for $k = 0, 1, \ldots, d$. It follows that the monomial is a basis for π_d since they span π_d by definition. ∎

To prove some basic results about bases in a vector space of functions it is convenient to introduce a matrix transforming one basis into another.

Lemma A.16. *Suppose S and T are finite dimensional vector spaces with $S \subset T$, and let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ be a basis for S and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)^T$ a basis for T. Then*

$$\boldsymbol{\phi} = \boldsymbol{A}^T \boldsymbol{\psi}, \tag{A.11}$$

for some matrix $\boldsymbol{A} \in \mathbb{R}^{m,n}$. If $f = \boldsymbol{\phi}^T \boldsymbol{c} \in S$ is given then $f = \boldsymbol{\psi}^T \boldsymbol{b}$ with

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{c}. \tag{A.12}$$

Moreover $\boldsymbol{A}$ has linearly independent columns.

Proof. Since $\phi_j \in T$ there are real numbers $a_{i,j}$ such that

$$\phi_j = \sum_{i=1}^{m} a_{i,j} \psi_i, \quad \text{for} \quad j = 1, \ldots, n,$$

This equation is simply the component version of (A.11). If $f \in S$ then $f \in T$ and $f = \boldsymbol{\psi}^T \boldsymbol{b}$ for some $\boldsymbol{b}$. By (A.11) we have $\boldsymbol{\phi}^T = \boldsymbol{\psi}^T \boldsymbol{A}$ and $f = \boldsymbol{\phi}^T \boldsymbol{c} = \boldsymbol{\psi}^T \boldsymbol{A}\boldsymbol{c}$ or $\boldsymbol{\psi}^T \boldsymbol{b} = \boldsymbol{\psi}^T \boldsymbol{A}\boldsymbol{c}$. Since $\boldsymbol{\psi}$ is linearly independent we get (A.12). Finally, to show that $\boldsymbol{A}$ has linearly independent columns suppose $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{0}$. Define $f \in S$ by $f = \boldsymbol{\phi}^T \boldsymbol{c}$. By (A.11) we have $f = \boldsymbol{\psi}^T \boldsymbol{A}\boldsymbol{c} = \boldsymbol{0}$. But then $f = \boldsymbol{\phi}^T \boldsymbol{c} = \boldsymbol{0}$. Since $\boldsymbol{\phi}$ is linearly independent we conclude that $\boldsymbol{c} = \boldsymbol{0}$. ∎

The matrix $\boldsymbol{A}$ in Lemma A.16 is called a *change of basis matrix*.

A basis for a vector space generated by n functions can have at most n elements.

Lemma A.17. *If $\boldsymbol{\psi} = (\psi_1 \ldots, \psi_k)^T$ is a linearly independent set in a vector space $S = \operatorname{span}(\phi_1, \ldots, \phi_n)$, then $k \leq n$.*

Proof. With $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ we have

$$\boldsymbol{\psi} = \boldsymbol{A}^T \boldsymbol{\phi}, \quad \text{for some} \quad \boldsymbol{A} \in \mathbb{R}^{n,k}.$$

If $k > n$ then $\boldsymbol{A}$ is a rectangular matrix with more columns than rows. From Corollary A.6 we know that the columns of such a matrix must be linearly dependent; I.e. there is some nonzero $\boldsymbol{c} \in \mathbb{R}^k$ such that $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{0}$. But then $\boldsymbol{\psi}^T \boldsymbol{c} = \boldsymbol{\phi}^T \boldsymbol{A}\boldsymbol{c} = \boldsymbol{0}$, for some nonzero $\boldsymbol{c}$. This implies that $\boldsymbol{\psi}$ is linearly dependent, a contradiction. We conclude that $k \leq n$. ∎

Lemma A.18. *Every basis for a vector space must have the same number of elements.*

Proof. Suppose $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)^T$ are two bases for the vector space. We need to show that $m = n$. Now

$$\boldsymbol{\phi} = \boldsymbol{A}^T \boldsymbol{\psi}, \quad \text{for some} \quad \boldsymbol{A} \in \mathbb{R}^{m,n},$$

$$\boldsymbol{\psi} = \boldsymbol{B}^T \boldsymbol{\phi}, \quad \text{for some} \quad \boldsymbol{B} \in \mathbb{R}^{n,m}.$$

By Lemma A.16 we know that both $\boldsymbol{A}$ and $\boldsymbol{B}$ have linearly independent columns. But then by Corollary A.6 we see that $m = n$. ∎

Definition A.19. *The number of elements in a basis in a vector space S is called the* dimension *of S, and is denoted by $\dim(S)$.*

The following lemma shows that every set of linearly independent functions in a vector space S can be extended to a basis for S. In particular every finite dimensional vector space has a basis.

Lemma A.20. *A set $\boldsymbol{\phi}^T = (\phi_1, \ldots, \phi_k)$ of linearly independent elements in a finite dimensional vector space S, can be extended to a basis $\boldsymbol{\psi}^T = (\psi_1, \ldots, \psi_m)$ for S.*

Proof. Let $S_k = \text{span}(\psi_1, \ldots, \psi_k)$ where $\psi_j = \phi_j$ for $j = 1, \ldots, k$. If $S_k = S$ then we set $m = k$ and stop. Otherwise there must be an element $\psi_{k+1} \in S$ such that $\psi_1, \ldots, \psi_{k+1}$ are linearly independent. We define a new vector space S_{k+1} by $S_{k+1} = \text{span}(\psi_1, \ldots, \psi_{k+1})$. If $S_{k+1} = S$ then we set $m = k + 1$ and stop the process. Otherwise we continue to generate vector spaces $S_{k+2}, S_{k+3}, \cdots$. Since S is finitely generated we must by Lemma A.17 eventually find some m such that $S_m = S$. ∎

The following simple, but useful lemma, shows that a spanning set must be a basis if it contains the correct number of elements.

Lemma A.21. *Suppose $S = \text{span}(\boldsymbol{\phi})$. If $\boldsymbol{\phi}$ contains $\dim(S)$ elements then $\boldsymbol{\phi}$ is a basis for S.*

Proof. Let $n = \dim(S)$ and suppose $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ is a linearly dependent set. Then there is one element, say ϕ_n which can be written as a linear combination of $\phi_1, \ldots, \phi_{n-1}$. But then $S = \text{span}(\phi_1, \ldots, \phi_{n-1})$ and $\dim(S) < n$ by Lemma A.17, a contradiction to the assumption that $\boldsymbol{\phi}$ is linearly dependent. ∎

A.4 Normed Vector Spaces

Suppose S is a vector space of functions. A *norm* $||\ ||$ $= ||f||$, is a function $\|\ \| : S \to [0, \infty)$ that satisfies for $f, g, \in S$, and $\alpha \in \mathbb{R}$ the following properties

$$\begin{array}{ll} 1. & ||f|| = 0 \quad \text{implies} \quad f = 0. \\ 2. & ||\alpha f|| = |\alpha| ||f||. \\ 3. & ||f + g|| \leq ||f|| + ||g||. \end{array} \tag{A.13}$$

Property 3 is known as the *Triangle Inequality*. The pair $(S, ||\ ||)$ is called a normed vector space (of functions).

In the rest of this section we assume that the functions in S are continuous, or at least piecewise continuous on some interval $[a, b]$.

Analogous to the p or ℓ^p norms for vectors in $\mathbb{R}^n$ we have the p or L^p norms for functions. They are defined for $1 \leq p \leq \infty$ and $f \in S$ by

$$\begin{array}{rl} ||f||_p = ||f||_{L^p[a,b]} & = \left(\int_a^b |f(x)|^p dx \right)^{1/p}, \quad p \geq 1, \\ ||f||_\infty = ||f||_{L^\infty[a,b]} & = \max_{a \leq x \leq b} |f(x)|. \end{array} \tag{A.14}$$

The 1,2, and ∞ norms are the most important.

We have for $1 \leq p \leq \infty$ and $f, g \in S$ the Hölder inequality

$$\int_a^b |f(x)g(x)| dx \leq ||f||_p ||g||_q, \quad \textit{where} \quad \frac{1}{p} + \frac{1}{q} = 1, \tag{A.15}$$

and the Minkowski inequality

$$||f + g||_p \leq ||f||_p + ||g||_p. \tag{A.16}$$

For $p = 2$ (A.15) is known as the Schwarz inequality, the Cauchy-Schwarz inequality, or the Buniakowski-Cauchy- Schwarz inequality.