UNIVERSITY OF STRATHCLYDE
MATHEMATICS & STATISTICS

MM406 Part I: Approximation theory

Contents

1 Introduction **2**
- 1.1 Motivation . . . 2
- 1.2 Linear spaces . . . 3
- 1.3 Norms . . . 4
- 1.4 Inner product spaces . . . 5
- 1.5 Hilbert spaces . . . 7
- 1.6 Exercises for Section 1 . . . 8

2 Best approximation **10**
- 2.1 Approximation by polynomials . . . 10
- 2.2 Minimax (best L_∞) approximation . . . 11
- 2.3 Best approximation in an inner product space . . . 12
- 2.4 Best approximation using an orthogonal basis . . . 15
- 2.5 Orthogonal polynomials in weighted L_2 spaces . . . 17
- 2.6 Best L_2-approximation in terms of trigonometric functions . . . 22
- 2.7 Exercises for Section 2 . . . 25

3 Interpolation **28**
- 3.1 Polynomial interpolation . . . 28
- 3.2 Interpolation by linear splines . . . 31
- 3.3 More smoothness: cubic splines . . . 34
- 3.4 Linear interpolation in 2D . . . 44
- 3.5 Exercises for Section 3 . . . 47

4 Numerical integration (quadrature) **49**
- 4.1 Gaussian quadrature in 1D . . . 49
- 4.2 2D quadrature . . . 50
- 4.3 Exercises for Section 4 . . . 51

1 Introduction

1.1 Motivation

Differential equations are very important in mathematical modelling, and often have to be solved numerically because there is no closed form analytic solution. The **finite element method** is a widely used way of obtaining a **numerical approximation** of the exact solution of differential boundary value problems (ODEs and elliptic PDEs), and is the focus of Part II of MM406.

Numerical approximation is also used in other contexts. For example:

(a) Suppose an experiment produces measurements of a quantity at fixed instants in time and that the quantity being measured depends continuously on time, t. The aim might be to model the data by a function $f(a_1, a_2, \ldots, a_n; t)$ of known form, but containing unknown parameters $\{a_i\}_{i=1}^n$. Generally the experiment produces more than n measurements and the objective is to select the parameters so that f gives a 'best' fit, in some sense, to the data.

(b) A bioengineer is given measurements of a knee joint. How does s/he represent the shape of the joint by means of a simple continuous function?

(c) A calculator manufacturer wants to 'implement' the cosine function, $\cos(x)$, on a chip. What is the simplest function made up of $+, -, *, \div$ operations that approximates $\cos(x)$ over $\left[0, \frac{\pi}{2}\right]$ to, say, 6 digits?

A few key points arise from the examples above.

(i) We often want the **best** fit from a family of possibilities.

(ii) To measure the quality of the final output we clearly need some method of measuring a function or a data set.

(iii) If the data are inaccurate (experimental errors, say) then there is no point in seeking a 'highly accurate' approximation. Most problems involve a trade-off between effort and accuracy.

In order to tackle problems like these we need to be able to characterise spaces of functions or vectors, and be able to measure their size. The relevant concepts are introduced below.

1.2 Linear spaces

Definition 1.1 A **linear space** (or **vector space**) $\mathcal{V}$ over the real numbers $\mathbb{R}$ is a set of elements closed with respect to addition and to multiplication by a scalar (real number). That is, its elements satisfy

1. $u,\, v \in \mathcal{V} \Rightarrow u + v \in \mathcal{V}$,
2. $v \in \mathcal{V},\, \alpha \in \mathbb{R} \Rightarrow \alpha v \in \mathcal{V}$.

Note that addition and multiplication must satisfy commutative, distributive and associative laws, and zero and inverse elements also exist.

Examples 1.1 Show that the following sets are all linear spaces, and give the zero element of each:

- the set $\mathbb{R}^n$ of all n-vectors;
- the set $C[a, b]$ of all continuous functions on $[a, b]$;
- the set $\mathbb{P}_n$ of all polynomials of degree less than or equal to n.

Definition 1.2 The set of elements $\{v_1,\, v_2,\, \ldots,\, v_n\}$ in the real linear space $\mathcal{V}$ is **linearly independent** if the only values of $c_i \in \mathbb{R}$ for which

$$\sum_{i=1}^{n} c_i\, v_i = 0$$

are $c_1 = c_2 = \cdots = c_n = 0$. A set of elements which is not linearly independent is called **linearly dependent**.

Examples 1.2 Determine whether or not the following sets of elements $\{\phi_1, \phi_2, \phi_3\}$ are linearly independent in $C[0, 1]$.
(a) $\phi_1(x) = x + 1$, $\phi_2(x) = x - 1$, $\phi_3(x) = x^2 + 1$;
(b) $\phi_1(x) = 1$, $\phi_2(x) = x - 1$, $\phi_3(x) = 3x + 2$.

Definition 1.3 A **basis** for the linear space $\mathcal{V}$ is a linearly independent set of functions $\{v_1,\, v_2,\, \ldots,\, v_n\} \in \mathcal{V}$ such that any element of $\mathcal{V}$ can be written as a linear combination of the v_i.

Examples 1.3 Show that $\left\{1, x, x^2\right\}$ forms a basis for $\mathbb{P}_2$. Does the set $\{\phi_1, \phi_2, \phi_3\}$ from (a) above also form a basis for $\mathbb{P}_2$?

A linear space is said to be **finite dimensional** if it has a finite number of basis elements.

Example:

- $\mathbb{R}^n$ and $\mathbb{P}_m$ are finite dimensional.
- $C[a,b]$ is not finite dimensional.

1.3 Norms

In order to judge how good a numerical approximation is we need to be able to measure the 'size' of quantities, and we use a **norm** to do this.

Definition 1.4 A **norm** on the real linear space $\mathcal{V}$ is a real-valued function on $\mathcal{V}$ whose value at $v \in \mathcal{V}$ is denoted by $\|v\|$. It satisfies the properties

(N1) $\|v\| \geq 0$, with $\|v\| = 0$ if and only if $v = 0$;

(N2) $\|\alpha v\| = |\alpha| \, \|v\|$, for any $\alpha \in \mathbb{R}$ and $v \in \mathcal{V}$; and

(N3) $\|u + v\| \leq \|u\| + \|v\|$ (the *triangle inequality*).

A linear space which has a norm defined on it is called a **normed linear space**.

Examples:

1. Let $v \in \mathcal{V} = \mathbb{R}^n$ (so $v = (v_1, v_2, \ldots v_n)^T$) and define the ℓ_p-norm by:

$$\|v\|_{\ell_p} = \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p} \qquad \text{if } 1 \leq p < \infty$$

and

$$\|v\|_{\ell_\infty} = \max_{1 \leq i \leq n} |v_i| \, .$$

It can be shown that $\| \cdot \|_{\ell_p}$ defines a norm for any $p \geq 1$ (see below and Exercises 1 for the important cases $p = 2, \ \infty$).

2. Let $u \in \mathcal{V} = C[a,b]$ (the set of continuous functions on the finite interval $[a,b]$) and define the L_p-norm by:

$$\|u\|_{L_p} = \left(\int_a^b |u(x)|^p \, dx \right)^{1/p} \qquad \text{if } 1 \leq p < \infty$$

and

$$\|u\|_{L_\infty} = \max_{a \leq x \leq b} |u(x)| \, .$$

It can be shown that $\|\cdot\|_{L_p}$ defines a norm for any $p \geq 1$ (see below and Exercises 1 for some important cases). Note that the L_p-norm is often written as $\|\cdot\|_{L_p} = \|\cdot\|_p$ if the context is clear enough to avoid confusion with $\|\cdot\|_{\ell_p}$.

Examples 1.4

1. Show that $\|\cdot\|_{\ell_\infty}$ defines a norm on $\mathbb{R}^n$.
2. Show that $\|\cdot\|_{L_1}$ defines a norm on $C[0,1]$.

The **distance** between two elements u and v of a normed linear space is defined to be $\|u-v\|$. Clearly, if v is an approximation of u, then the distance $\|u-v\|$ can serve as a measure of quality of the approximation. For example, if u, v are functions in $C[a,b]$ and $\|u-v\|_{L_\infty} < 0.1$, then

$$-0.1 \leq u(x) - v(x) \leq 0.1, \qquad \text{for all } x \in [a,b],$$

which shows that $v(x)$ approximates $u(x)$ everywhere in $[a,b]$ with error at most 0.1. If $\|u-v\|_{L_2}$ is small, it does not guarantee in general that $|u(x)-v(x)|$ is small for all x. However, it must be small for 'most' x.

1.4 Inner product spaces

Definition 1.5 An **inner product** on the real linear space $\mathcal{V}$ is a mapping

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$$

which satisfies the following properties

(IP1) $\langle v, v \rangle \geq 0$, with $\langle v, v \rangle = 0$ if and only if $v = 0$;

(IP2) $\langle u, v \rangle = \langle v, u \rangle$ for any u, $v \in \mathcal{V}$ **(commutativity)**; and

(IP3) $\langle \alpha\, u + \beta\, v, w \rangle = \alpha \langle u, w \rangle + \beta \langle v, w \rangle$ for any u, v, $w \in \mathcal{V}$ and any α, $\beta \in \mathbb{R}$ **(linearity)**.

A linear space which has an inner product defined on it is called an **inner product space**.

Examples:

1. Let u, $v \in \mathcal{V} = \mathbb{R}^n$ and define the **ℓ_2-inner product** by

$$\langle u, v \rangle_{\ell_2} = \sum_{i=1}^{n} u_i\, v_i\,.$$

2. Let $u,\, v \in \mathcal{V} = C[a,b]$ and define the $\boldsymbol{L_2}$**-inner product** by

$$\langle u, v\rangle_{L_2} = \int_a^b u(x)\, v(x)\, dx\,.$$

Because it is used so frequently (in both parts of MM406), we often simplify the notation and write this inner product as

$$\langle u, v\rangle_{L_2} = (u, v)\,. \tag{1.1}$$

3. Let $u,\, v \in \mathcal{V} = C[a,b]$, and let $\omega(x) \geq 0$ be a **weight function** (see below). The **weighted $\boldsymbol{L_2}$-inner product** of u and v is usually written as $(u,\, v)_\omega$ and is defined to be

$$(u,\, v)_\omega = \int_a^b \omega(x)\, u(x)\, v(x)\, dx\,.$$

We need to be careful with the definition of a weight function in order for this to satisfy the second part of (IP1).

Definition 1.6 A function ω is called a **weight function** on the interval (a,b) if $\omega(x) \geq 0$ for all $x \in (a,b)$ and $\omega(x) \not\equiv 0$ on any subinterval, and $\int_a^b \omega(x)\, dx$ exists.

Note that a weight function does not have to be bounded on (a,b), but it must be integrable. For example, one important weight function which we shall use later on is $\omega(x) = 1/\sqrt{1-x^2}$ on $(-1,1)$.

We now show that an inner product space is also a normed linear space with the norm defined by

$$\|v\| = \sqrt{\langle v, v\rangle}\,. \tag{1.2}$$

A norm which comes from an inner product is called an **induced norm**. 'Most' norms are not induced by inner products (e.g. see Exercises 1).

Examples 1.5 Show that $\|\cdot\|$ defined by (1.2) satisfies properties (N1) and (N2) of the norm definition.

To see that the triangle inequality (N3) is also satisfied, we first prove the fundamental **Cauchy-Schwarz inequality**

$$|\langle u, v\rangle| \leq \|u\|\, \|v\|\,. \tag{1.3}$$

Proof. By (IP1) we have $\langle u + \alpha\, v,\, u + \alpha\, v\rangle \geq 0$ for any $u,\, v \in \mathcal{V}$ and $\alpha \in \mathbb{R}$. Properties (IP2) and (IP3) give

$$\langle u + \alpha\, v,\, u + \alpha\, v\rangle = \langle u, u\rangle + 2\alpha\langle u, v\rangle + \alpha^2\langle v, v\rangle\,.$$

Hence, the quadratic polynomial $a\,\alpha^2 + b\,\alpha + c$, with $a = \langle v, v\rangle$, $b = 2\langle u, v\rangle$, $c = \langle u, u\rangle$, is non-negative for any value of α. It follows that $b^2 - 4ac \leq 0$, i.e. $(2\langle u, v\rangle)^2 - 4\langle v, v\rangle\langle u, u\rangle \leq 0$, which is equivalent to (1.3). □

By taking $\alpha = 1$ in the identity displayed in the above proof and using the Cauchy-Schwarz inequality, we get

$$\|u + v\|^2 = \|u\|^2 + 2\langle u, v\rangle + \|v\|^2 \leq \|u\|^2 + 2\|u\|\,\|v\| + \|v\|^2 = (\|u\| + \|v\|)^2$$

which proves the triangle inequality for the norm defined by (1.2).
Note that taking $\alpha = -1$ we get the identity

$$\|u - v\|^2 = \|u\|^2 - 2\langle u, v\rangle + \|v\|^2 \tag{1.4}$$

which is useful for computing distances in an inner product space.

Definition 1.7 We say that u, $v \in \mathcal{V}$ are **orthogonal** if their inner product is zero, i.e. if $\langle u, v\rangle = 0$.

Theorem 1.1 *A mutually orthogonal set of non-zero elements of an inner product space is linearly independent.*

Proof. Suppose that $\{v_i\}_{i=0}^n$ is mutually orthogonal (i.e. $\langle v_i, v_j\rangle = 0$ for $i \neq j$), and suppose that there are $c_i \in \mathbb{R}$ with $\sum_{i=0}^n c_i\, v_i = 0$. Taking the inner product with v_j gives

$$0 = \langle \sum_{i=0}^{n} c_i\, v_i,\ v_j\rangle = \sum_{i=0}^{n} c_i\, \langle v_i, v_j\rangle = c_j\, \langle v_j, v_j\rangle$$

because $\langle v_i, v_j\rangle = 0$ for $i \neq j$. Hence each $c_j = 0$ by (IP1). □

1.5 Hilbert spaces

A Hilbert space is a **complete inner product space**. This means that if the sequence $\{u_n\}_{n=1}^{\infty}$ is a Cauchy sequence with respect to the induced norm (i.e. for every positive number ε there is an integer N such that $\|u_n - u_m\| < \varepsilon$ whenever m, $n \geq N$), then the sequence converges to an element of the space. The space $C[a, b]$ is **not** a Hilbert space under the L_2 inner product, because it is possible to find a Cauchy sequence of continuous functions which converges to a discontinuous function (see Exercises 1). But there is a bigger space which is a Hilbert space. It consists of all functions which are **square integrable** on (a, b) and is called $L_2(a, b)$, i.e.

$$L_2(a, b) = \left\{ u : \int_a^b |u(x)|^2\, dx < \infty \right\}. \tag{1.5}$$

Example: The function

$$u(x) = \begin{cases} -2 & \text{if } x \in (0,1] \\ 5 & \text{if } x \in (1,2) \end{cases}$$

is in $L_2(0,2)$ but not in $C(0,2)$.

Other important Hilbert spaces which will be used in Part II are the spaces H^m where m is a non-negative integer. If $I \subset \mathbb{R}$ is an (open or closed) interval, then

- $H^0(I) = L_2(I)$;
- $H^1(I)$ is the space of functions u for which u, $u' \in L_2(I)$;
- $H^2(I)$ is the space of functions u for which u, u', $u'' \in L_2(I)$;

$$\cdots$$

- $H^m(I)$ is the space of functions u for which u and all derivatives of u up to the mth derivative are in $L_2(I)$.

1.6 Exercises for Section 1

1–1. Which of the following subsets of $\mathbb{R}^3$ form linear spaces (i.e. linear subspaces of $\mathbb{R}^3$)? (Here $\boldsymbol{x} = (x_1, x_2, x_3)$.)

(a) $\{\boldsymbol{x} : x_1 = x_2 \text{ and } x_3 = 0\}$;

(b) $\{\boldsymbol{x} : x_1 = x_2 + 1\ \}$;

(c) $\{\boldsymbol{x} : x_1,\ x_2,\ x_3 \text{ are all positive}\}$.

1–2. Determine (with a proof) whether or not the following sets are linear spaces.

(a) The set of real $n \times n$ matrices;

(b) The set of symmetric real $n \times n$ matrices;

(c) The set of real 2×2 matrices which have positive determinant;

(d) $\{f \in C[-1,1] : f(0) = 0\}$;

(e) $\{f \in C[-1,1] : f(0) = 1\}$.

1–3. Find $\|v\|_{\ell_1}$, $\|v\|_{\ell_2}$ and $\|v\|_{\ell_\infty}$ for

(a) $v = (3, -4, 0, 2)^T$; (b) $v = (2, -1, -3, 4)^T$.

1–4. Show that $\|\cdot\|_{\ell_2}$ defines a norm on $\mathbb{R}^n$.

1–5. Show that $\|\cdot\|_{L_2}$ and $\|\cdot\|_{L_\infty}$ both define norms on $C[0,1]$.

1–6. (a) For any $\varepsilon > 0$, find a continuous function $u_\varepsilon \in C[0,1]$ such that $\|u_\varepsilon\|_{L_2} < \varepsilon$, but $\|u_\varepsilon\|_{L_\infty} = 1$.

(b) Show that if $u,\ v \in C[a,b]$ and $|u(x) - v(x)| \geq 0.5$ on an interval of length r, then $\|u - v\|_{L_2} > 0.5\sqrt{r}$.

1–7. Which of the following functions $\omega(x)$ are valid **weight functions** on the given intervals? (Note that $H(x)$ is the Heaviside function.)

(a) x^2 on $(-1,1)$; (b) $x^{-1/2}$ on $(0,1)$; (c) 1 on $(-1,1)$;
(d) $\sin x$ on $(-\pi,\pi)$; (e) $|\sin x|$ on $(-\pi,\pi)$; (f) $H(x)$ on $(-1,1)$;
(g) e^{-x^2} on $\mathbb{R}$; (h) e^{-x} on $(0,\infty)$; (i) 1 on $(0,\infty)$.

1–8. Show that the induced norm $\|\cdot\|$ on an inner product space $\mathcal{V}$ satisfies the **parallelogram equality:**

$$\|u+v\|^2 + \|u-v\|^2 = 2\left(\|u\|^2 + \|v\|^2\right) \quad \text{for all } u,\ v \in \mathcal{V}\,.$$

1–9. Show that the norm ℓ_p on $\mathbb{R}^n$ is not an induced norm for $p \neq 2$. (Hint: set $u = (1,1,0,\ldots,0)$ and $v = (1,-1,0,\ldots,0)$, and show that they do not satisfy the above parallelogram equality.)

1–10. Let $u_n(x) = \sin(nx)$, where n is a positive integer. Show that u_m and u_n are orthogonal in $L_2(0,\pi)$ if $m \neq n$.

1–11. In Part II of MM406 you will see the bilinear form

$$a(u,v) = \int_0^1 u'(x)\,v'(x)\,dx$$

and the associated "energy norm" $\|u\|_a = \sqrt{a(u,u)}$. Show that this is in fact a **semi-norm** and not a norm, i.e. it satisfies $\|u\|_a \geq 0$ for all u and properties (N2) and (N3), but it is possible for $\|u\|_a = 0$ for a function u which is not identically zero.

1–12. Let $\alpha_n = 1/2 + 1/n$ and define the sequence $\{u_n\} \in C[0,1]$ by

$$u_n(x) = \begin{cases} 0 & \text{if } x \in [0,1/2] \\ n\,(x - 1/2) & \text{if } x \in (1/2,\alpha_n) \\ 1 & \text{if } x \in [\alpha_n,1] \end{cases}$$

Show that $\{u_n\}$ is a Cauchy sequence with respect to the L_2 norm, but that it converges to a function which is not in $C[0,1]$. (This proves that $C[0,1]$ is not a Hilbert space under the L_2-inner product.)

2 Best approximation

Abstract Problem. Let $\mathcal{A}$ be a finite dimensional subspace of a normed linear space $\mathcal{V}$ and let $v \in \mathcal{V}$. An element $a^* \in \mathcal{A}$ of minimum distance from v is called a **best approximation of v from $\mathcal{A}$**, and the problem of determining a^* is called a **best approximation problem**.
In terms of the norm we have

$$\|v - a^*\| \leq \|v - a\| \quad \forall a \in \mathcal{A}\,. \tag{2.1}$$

Examples 2.1 Let $\mathcal{V} = \mathbb{R}^2$, and $v = (1,1)^T$. Determine all best approximations of v from $\mathcal{A}$ in both the ℓ_2- and ℓ_∞-norms when: (i) $\mathcal{A} = \{(x,y)^T : x = 0\}$, (ii) $\mathcal{A} = \{(x,y)^T : y + x = 0\}$.

Examples 2.2 Let $\mathcal{V} = \mathbb{R}^3$, and $v = (1,2,4)^T$. Find best approximations of v in the ℓ_2- and ℓ_∞-norms using an element of the form $(\alpha,\alpha,\alpha)^T$.

As we can see from the first example above, it cannot be assumed in general that the best approximation a^* is *unique*. However we shall show that there is a unique best approximation in an inner product space. It can also be shown that there always exists a best approximation (which may not be unique) when $\mathcal{V}$ is a normed linear space and the approximation subspace $\mathcal{A}$ is finite dimensional, but we shall not prove this result.
We first review some results on polynomials which show why they can be useful in approximation theory, and then briefly look at the L_∞ best approximation problem, before returning to look at the inner product case in more detail.

2.1 Approximation by polynomials

In numerical approximation, extensive use is made of polynomials. Recall that we denote by $\mathbb{P}_n$ the linear space of real polynomials of degree at most n.
There are a number of reasons why polynomials are useful:

(a) they are easy to evaluate;

(b) they are easy to differentiate and integrate;

(c) Weierstrass's Theorem shows that they can approximate arbitrarily well.

Theorem 2.1 (Weierstrass) *Suppose $u \in C[a,b]$. Given any $\varepsilon > 0$, there exists a polynomial $q \in \bigcup_{n\geq 0} \mathbb{P}_n$ (whose degree generally depends on ε), such*

that

$$\|u - q\|_{L_\infty} < \varepsilon \,.$$

Similarly, let $1 \le p < \infty$*. For any* $u \in L_p(a,b)$ *and any* $\varepsilon > 0$ *there is a polynomial* q *such that*

$$\|u - q\|_{L_p} < \varepsilon \,.$$

Unfortunately Weierstrass's theorem is not of much practical use since for a particular u and ε the degree of q may be extremely high. The following theorem gives more information about the rate of convergence of the best approximating polynomials under additional assumptions on u.

Theorem 2.2 (Jackson) *Let* $1 \le p \le \infty$*. Suppose* u *is* r *times continuously differentiable (*$r \ge 1$*) on* $[a,b]$*. Then for any* $n > r$ *there is a polynomial* $q_n \in \mathbb{P}_n$ *such that*

$$\|u - q_n\|_{L_p} \le C_r \left(\frac{b-a}{n+1}\right)^r \|u^{(r)}\|_{L_p} \,,$$

where C_r *is a constant depending only on* r*. In particular,* C_r *may be taken equal to* $(\pi r)^r/(4^r r!)$ *.*

A high degree polynomial approximation may be very wiggly if the function u does not possess sufficiently high order derivatives (Runge phenomenon). For this reason, it is often preferred to restrict the degree of the approximating polynomial. This can involve different low-degree polynomial approximations in different sub-regions of $[a,b]$ as we shall see later when we introduce splines.

2.2 Minimax (best L_∞) approximation

Suppose that $u \in \mathcal{V} = C[a,b]$ and we wish to find the best approximation of u from $\mathcal{A} = \mathbb{P}_n$ in the L_∞-norm. That is we want to find a polynomial q_n of degree at most n (called the **minimax polynomial**) such that

$$\|u - q_n\|_{L_\infty} = \min_{q \in \mathbb{P}_n} \|u - q\|_{L_\infty} = \min_{q \in \mathbb{P}_n} \max_{x \in [a,b]} |u(x) - q(x)| \,.$$

It should be clear where the name **minimax** comes from: we seek to **mini**mise the **max**imum pointwise error.

Examples 2.3 Find the minimax approximation when $\mathcal{A} = \mathbb{P}_0$ (the space of constant functions).

There are very few functions for which there is a simple closed form for the minimax polynomial of (general) degree n. One problem which has been solved

is to find the minimax polynomial of degree n for the function $u(x) = x^{n+1}$. The solution is given in terms of **Chebyshev polynomials** (which will be defined later in this section). Fortunately the theory for the best approximation problem in an inner product space is far more complete, and we concentrate on this for the rest of the section. Note that the norm in an inner product space always means the induced norm, i.e. $\|v\| = \sqrt{\langle v, v\rangle}$.

2.3 Best approximation in an inner product space

We now give a theorem that characterises an important property satisfied by an element of an inner product space which is a best approximation. After this we shall show that there always is a best approximation, and it is unique (so it makes sense to talk about "the" best approximation in an inner product space).

Theorem 2.3 (Projection theorem) *Suppose that $\mathcal{A}$ is a finite dimensional subspace of the inner product space $\mathcal{V}$ and let $u \in \mathcal{V}$. Then $a^* \in \mathcal{A}$ is a best approximation of u from $\mathcal{A}$ iff (if and only if)*

$$\langle u - a^*, a\rangle = 0 \quad \forall a \in \mathcal{A} \tag{2.2}$$

i.e. $u - a^*$ is orthogonal to all elements of $\mathcal{A}$.

Proof.
(i) Necessity (i.e. any best approximation satisfies (2.2)):
Suppose that a^* is a best approximation of u from $\mathcal{A}$ and (for a contradiction) assume $\exists a \in \mathcal{A}$ such that

$$\langle u - a^*, a\rangle = \lambda \neq 0\,.$$

Clearly, $a \neq 0$. Without loss of generality we assume that $\lambda > 0$. (Otherwise replace a with $-a$.) We now employ a *perturbation argument*. Take $\varepsilon > 0$ and consider

$$\begin{aligned}
\|u - a^* - \varepsilon a\|^2 &= \|u - a^*\|^2 - 2\langle u - a^*, \varepsilon a\rangle + \|\varepsilon a\|^2 \quad \text{(by (1.4))} \\
&= \|u - a^*\|^2 - 2\varepsilon\,\lambda + \varepsilon^2\,\|a\|^2 \quad \text{(by definition of } \lambda) \\
&= \|u - a^*\|^2 + \varepsilon\|a\|^2\left(\varepsilon - \frac{2\lambda}{\|a\|^2}\right).
\end{aligned}$$

Now, if we have chosen ε small enough such that $\varepsilon < 2\lambda/\|a\|^2$, then the second term in the last sum is negative, which implies

$$\|u - a^* - \varepsilon a\|^2 < \|u - a^*\|^2\,.$$

Thus, $\hat{a} := a^* + \varepsilon a$ is in $\mathcal{A}$ (recall $\mathcal{A}$ is a linear subspace) and satisfies $\|u - \hat{a}\|^2 < \|u - a^*\|^2$. This contradicts the fact that a^* is the best approximation of u in $\mathcal{A}$. Thus our assumption that $\lambda \neq 0$ is incorrect, and we have shown that any best approximation satisfies (2.2).

(ii) Sufficiency (condition (2.2) guarantees that a^* is a best approximation):
Suppose $\langle u - a^*, a\rangle = 0 \ \ \forall a \in \mathcal{A}$. Then for any $a \in \mathcal{A}$ we have

$$\begin{aligned} \|u - a\|^2 &= \|u - a^* - (a - a^*)\|^2 \\ &= \|u - a^*\|^2 + \|a - a^*\|^2 - 2\langle u - a^*, a - a^*\rangle \,. \end{aligned}$$

Since $(a - a^*) \in \mathcal{A}$ the final term is zero by the assumption. Hence

$$\|u - a\|^2 \geq \|u - a^*\|^2$$

and it follows that a^* is a best approximation of u from $\mathcal{A}$. □

Corollary. If $\mathcal{A}$ is a finite dimensional subspace of $\mathcal{V}$ with basis $\{\phi_i\}_{i=0}^n$ then a^* is the best approximation of $u \in \mathcal{V}$ from $\mathcal{A}$ iff

$$\langle u - a^*, \phi_i\rangle = 0 \quad \text{for} \quad i = 0 : n \,. \tag{2.3}$$

Proof of Corollary. We need to show that (2.2) and (2.3) are equivalent, i.e.

$$\langle u - a^*, \phi_i\rangle = 0 \quad \text{for } i = 0 : n \quad \Longleftrightarrow \quad \langle u - a^*, a\rangle = 0 \quad \forall\, a \in \mathcal{A} \,.$$

($\Leftarrow$): is obvious, because each $\phi_i \in \mathcal{A}$.

($\Rightarrow$): The set $\{\phi_i\}_{i=0}^n$ is a basis for $\mathcal{A}$, so for any $a \in \mathcal{A}$ there are $\alpha_i \in \mathbb{R}$ with $a = \sum_{i=0}^n \alpha_i \, \phi_i$. Suppose that (2.3) holds, then the linearity of the inner product gives

$$\langle u - a^*, a\rangle = \sum_{i=0}^{n} \alpha_i \, \langle u - a^*, \phi_i\rangle = 0 \,,$$

which is (2.2). □

The above results say that $a^* \in \mathcal{A}$ is a best approximation of $u \in \mathcal{V}$ iff $u - a^*$ is orthogonal to the basis elements ϕ_i of $\mathcal{A}$. But it is not obvious that there *is* a best approximation, or how many elements of $\mathcal{A}$ might satisfy this condition. We now show how to explicitly construct *the* best approximation, and the construction shows (i) that there always is a best approximation in a finite dimensional inner product space (existence), and (ii) that there is only one (uniqueness).
For any $a^* \in \mathcal{A}$, there are constants $\alpha_0^*, \ldots \alpha_n^*$ with $a^* = \sum_{j=0}^n \alpha_j^* \, \phi_j$, and if a^* satisfies (2.3) then for each $i = 0 : n$,

$$\langle u, \phi_i\rangle = \langle \sum_{j=0}^{n} \alpha_j^* \phi_j, \phi_i\rangle = \sum_{j=0}^{n} \alpha_j^* \, \langle \phi_j, \phi_i\rangle = \sum_{j=0}^{n} \langle \phi_i, \phi_j\rangle \, \alpha_j^* \,.$$

That is, (2.3) is equivalent to the linear system

$$M\boldsymbol{\alpha}^* = \boldsymbol{b} \tag{2.4}$$

where $\boldsymbol{\alpha}^* = (\alpha_0^*,\, \alpha_1^*,\, \ldots,\, \alpha_n^*)^T \in \mathbb{R}^{n+1}$, $\boldsymbol{b} \in \mathbb{R}^{n+1}$ is the vector with components $b_i = \langle u, \phi_i \rangle$ for $i = 0 : n$, and $M \in \mathbb{R}^{(n+1)\times(n+1)}$ is the (symmetric) matrix with components $M_{ij} = \langle \phi_i, \phi_j \rangle$ for $i,\, j = 0 : n$. We know from linear algebra that if M is an invertible (nonsingular) matrix, then (2.4) has a solution and it is unique, and this corresponds to a unique best approximation a^*.

Claim: Matrix M is nonsingular (invertible).

Proof. If M is a singular matrix then there must exist a nonzero vector $\hat{\boldsymbol{c}} \in \mathbb{R}^{n+1}$ for which $M\,\hat{\boldsymbol{c}} = \boldsymbol{0}$, and so $\hat{\boldsymbol{c}}^T M\,\hat{\boldsymbol{c}} = 0$. But for any scalars $c_i \in \mathbb{R}$ which are not all zero, $a = \sum_{i=0}^n c_i\,\phi_i$ is a nonzero element of $\mathcal{A}$ and so $\langle a, a \rangle > 0$ by (IP1). By linearity,

$$\langle a, a \rangle = \sum_{i,j=0}^{n} c_i\,\langle \phi_i, \phi_j \rangle\,c_j = \sum_{i,j=0}^{n} c_i\,M_{ij}\,c_j = \sum_{i=0}^{n} c_i\,(M\,\boldsymbol{c})_i = \boldsymbol{c}^T M \boldsymbol{c}$$

where $\boldsymbol{c} = (c_0,\, c_1,\, \ldots,\, c_n)^T$. Thus we have shown that $\boldsymbol{c}^T M \boldsymbol{c} > 0$ for all nonzero $\boldsymbol{c} \in \mathbb{R}^{n+1}$ – i.e. M is a positive definite matrix, and hence it is not singular. □

We state this result as a theorem, and then summarise the steps which need to be carried out in order to construct the best approximation.

Theorem 2.4 *Suppose that $\mathcal{A}$ is a finite dimensional subspace of the inner product space $\mathcal{V}$ and let $u \in \mathcal{V}$. Then there exists a unique best approximation $a^* \in \mathcal{A}$ of u.*

Algorithm 2.1: Construction of the best approximation to u in $\mathcal{A}$

1. Find a basis $\{\phi_i\}_{i=0}^n$ for $\mathcal{A}$.
2. Calculate $\boldsymbol{b}$ and M (as defined above).
3. Solve the linear system $M\,\boldsymbol{\alpha}^* = \boldsymbol{b}$ for $\boldsymbol{\alpha}^* \in \mathbb{R}^{n+1}$.
4. Then $a^* = \sum_{i=0}^n \alpha_i^*\,\phi_i \in \mathcal{A}$ is the best approximation.

Examples 2.4 Find the best approximation to $u = \pi\,\sin(\pi\,x) \in L_2(0,1)$ from $\mathcal{A} = \mathbb{P}_1$.

Although the above construction will always work in theory, in practice there can be problems if the choice of basis functions gives a **badly conditioned** matrix M. (This means that the computed solution $\boldsymbol{\alpha}^*$, and hence a^*, could be very inaccurate.)

Example: Let $\mathcal{V} = L_2(0,1)$ and $\mathcal{A} = \mathbb{P}_n$, and choose the basis functions to be the **monomials** $\phi_i(x) = x^i$ for $i = 0:n$. In this case M is the so-called Hilbert matrix, a famous badly conditioned matrix whose condition number grows like $e^{3.5\,n}$. Condition numbers of M for $n = 1:13$ computed using `cond(hilb(n))` in MATLAB are shown in the following table.

n	condition number of M
1	1
2	19.2815
3	524.0568
4	1.5514×10^4
5	4.7661×10^5
6	1.4951×10^7
7	4.7537×10^8
8	1.5258×10^{10}
9	4.9315×10^{11}
10	1.6025×10^{13}
11	5.2247×10^{14}
12	1.8032×10^{16}
13	1.1716×10^{18}

Fortunately we can always eliminate this problem by using an **orthogonal** set of basis functions, and this also has the additional consequence of ensuring that M is a diagonal matrix, which makes it straightforward to compute $\boldsymbol{\alpha}^*$.

2.4 Best approximation using an orthogonal basis

The set of basis vectors $\{\phi_i\}_{i=0}^n$ for $\mathcal{A}$ is called an orthogonal basis if the basis functions are mutually orthogonal, i.e. if $\langle \phi_i, \phi_j \rangle = 0$ when $i \neq j$. In this case $M_{ij} = 0$ when $i \neq j$ and M is the **diagonal matrix** whose diagonal entries are $M_{ii} = \|\phi_i\|^2$. It follows from (2.3) that the best approximation is

$$a^* = \sum_{i=0}^{n} \frac{\langle u, \phi_i \rangle}{\langle \phi_i, \phi_i \rangle}\, \phi_i \,. \tag{2.5}$$

If we impose the further condition that the basis elements are **orthonormal** (i.e. $\langle \phi_i, \phi_i \rangle = 1$, so $\langle \phi_i, \phi_j \rangle = \delta_{i,j}$), then

$$a^* = \sum_{i=0}^{n} \langle u, \phi_i \rangle \, \phi_i \, . \tag{2.6}$$

Examples 2.5 Show that the set $\{\phi_0, \phi_1, \phi_2\} = \left\{ \frac{1}{\sqrt{2}}, \sqrt{\frac{3}{2}}\, x, \sqrt{\frac{5}{8}}\, (3x^2 - 1) \right\}$ is an orthonormal basis for $\mathbb{P}_2$ in $L_2(-1, 1)$.

Examples 2.6 If $\{\phi_i\}_{i=0}^n$ is an orthonormal basis for $\mathcal{A}$, then show that the accuracy of the approximation (2.6) is

$$\|u - a^*\|^2 = \|u\|^2 - \sum_{j=0}^{n} \langle u, \phi_j \rangle^2 \, .$$

Construction of an orthogonal basis for $\mathcal{A}$

Suppose that $\{\psi_j\}_{j=0}^n$ is a basis for $\mathcal{A}$ and let $\mathcal{A}_i$ be the subspace of $\mathcal{A}$ which is spanned by $\{\psi_j\}_{j=0}^i$. Then $\mathcal{A}_n = \mathcal{A}$ and $\mathcal{A}_{i-1} \subset \mathcal{A}_i$ for $i = 1 : n$. The **Gram-Schmidt** process may be used to generate an orthogonal basis $\{\phi_j\}_{j=0}^n$ for $\mathcal{A}$ as follows.

Algorithm 2.2: Gram-Schmidt construction of an orthogonal basis

1. Let $\phi_0 := \psi_0$ and set $i = 1$.
2. Let ψ_i^* be the best approximation to ψ_i out of $\mathcal{A}_{i-1}$, so that
   $\psi_i^* = \sum_{j=0}^{i-1} \dfrac{\langle \psi_i, \phi_j \rangle}{\langle \phi_j, \phi_j \rangle} \, \phi_j$ by (2.5).
3. Set $\phi_i := \psi_i - \psi_i^*$.
4. Set $i := i + 1$.
5. If $i \leq n$ go to 2, otherwise stop.

Proof that $\{\phi_j\}_{j=0}^n$ is an orthogonal basis for $\mathcal{A}$.
We use mathematical induction with **inductive hypothesis**
$(\text{IH})_i$: $\{\phi_j\}_{j=0}^{i-1}$ is an orthogonal basis for A_{i-1}.

Clearly $(\mathrm{IH})_1$ is true, since $\{\phi_0\}$ is an orthogonal basis for $\mathcal{A}_0$.

We now assume that $(\mathrm{IH})_i$ is true for some $i \geq 1$ and show that this implies $(\mathrm{IH})_{i+1}$ is also true. In this case, induction on i means that $(\mathrm{IH})_i$ is true for $i = 1 : n$.

If $(\mathrm{IH})_i$ is true, then $\{\phi_j\}_{j=0}^{i-1}$ is a mutually orthogonal set, and we only need to show that ϕ_i is orthogonal to each of its elements (since by Theorem 1.1 the set $\{\phi_j\}_{j=0}^{i}$ is linearly independent and hence a basis for $\mathcal{A}_i$).

Now, ψ_i^* is the best 2-norm approximation to ψ_i out of $\mathcal{A}_{i-1}$, so by (2.3),

$$\langle \psi_i - \psi_i^*, \phi_j \rangle = 0 \quad \text{for } j = 0 : i-1 \,.$$

That is

$$\langle \phi_i, \phi_j \rangle = 0 \quad \text{for } j = 0 : i-1 \,,$$

as required for $(\mathrm{IH})_{i+1}$. □

Examples 2.7 Apply the Gram-Schmidt process to generate an orthogonal basis for $\mathbb{P}_2 \subset L_2(-1, 1)$ starting from the basis 1, x, x^2.

2.5 Orthogonal polynomials in weighted L_2 spaces

Orthogonal elements are particularly useful in practice if $\mathcal{A} = \mathbb{P}_n$, i.e. the aim is to approximate a function $u \in \mathcal{V}$ by a polynomial.

We assume that the inner product satisfies

$$\langle u, x\, v \rangle = \langle x\, u, v \rangle \tag{2.7}$$

(where x is the polynomial variable), which is trivially the case for the weighted L_2-inner products

$$\langle u, v \rangle = (u, v)_\omega = \int_a^b \omega(x)\, u(x)\, v(x)\, dx \,.$$

We now give an algorithm for computing an orthogonal basis of *polynomials*. It relies, in particular, on (2.7) and on the notion of degree, and therefore it is not as general as the Gram-Schmidt method. However, it is easier to use than the Gram-Schmidt process since it exploits a three-term recurrence relation for the polynomials, and thus requires the computation of far fewer inner products in each step. Indeed, ϕ_{j+1} is given in terms of ϕ_j and ϕ_{j-1} rather than all of $\phi_0, \ldots, \phi_j$.

Recall that a polynomial p is said to be **monic** if its leading coefficient is 1, i.e.

$$p(x) = x^n + a_{n-1}\, x^{n-1} + \cdots + a_1\, x + a_0 \,.$$

Theorem 2.5 *Let* $\phi_0(x) = 1$, $\alpha_0 = \langle \phi_0, x\,\phi_0\rangle / \|\phi_0\|^2$ *and* $\phi_1(x) = (x-\alpha_0)\,\phi_0(x) = x - \alpha_0$. *Now for* $j \geq 1$ *recursively define*

$$\left.\begin{array}{rcl} \alpha_j &=& \langle \phi_j, x\,\phi_j\rangle / \|\phi_j\|^2 \\ \beta_j &=& \langle \phi_{j-1}, x\,\phi_j\rangle / \|\phi_{j-1}\|^2 \end{array}\right\} \tag{2.8}$$

$$\phi_{j+1}(x) = (x - \alpha_j)\,\phi_j(x) - \beta_j\,\phi_{j-1}(x)\,. \tag{2.9}$$

Then ϕ_j *is a* <u>monic</u> *polynomial of degree* j *and the set* $\{\phi_i\}_{i=0}^n$ *is orthogonal.*

Proof. The fact that ϕ_j is monic and has degree j follows from the construction. To prove orthogonality we use induction. Note that

$$\langle \phi_0, \phi_1\rangle = \langle \phi_0, (x-\alpha_0)\,\phi_0\rangle = \langle \phi_0, x\phi_0\rangle - \alpha_0\,\|\phi_0\|^2 = 0\,.$$

We use induction to prove the theorem, with inductive hypothesis:
$(\mathrm{IH})_j$: the set $\{\phi_i\}_{i=0}^j$ is orthogonal.
We have seen above that $\langle \phi_0, \phi_1\rangle = 0$ and so $(\mathrm{IH})_1$ is true, and we now show that if $(\mathrm{IH})_j$ holds then $(\mathrm{IH})_{j+1}$ also holds.
Suppose $(\mathrm{IH})_j$ holds for some $j \geq 1$. For $(\mathrm{IH})_{j+1}$ we need to show that $\langle \phi_i, \phi_{j+1}\rangle = 0$ for $i = 0 : j$. From (2.9) we have

$$\langle \phi_i, \phi_{j+1}\rangle = \langle \phi_i, x\phi_j\rangle - \alpha_j\,\langle \phi_i, \phi_j\rangle - \beta_j\,\langle \phi_i, \phi_{j-1}\rangle \tag{2.10}$$

and we consider three cases separately.

<u>Case I:</u> $i < j-1$
In this case $(\mathrm{IH})_j$ implies that both $\langle \phi_i, \phi_j\rangle$ and $\langle \phi_i, \phi_{j-1}\rangle$ are zero. Also (2.7) gives $\langle \phi_i, x\phi_j\rangle = \langle x\phi_i, \phi_j\rangle$ which is also zero by $(\mathrm{IH})_j$ because $x\phi_i(x)$ has degree less than j (and so is a linear combination of ϕ_k with $k = 0 : i+1$). Hence, each term on the RHS of (2.10) is zero, so $\langle \phi_i, \phi_{j+1}\rangle = 0$.

<u>Case II:</u> $i = j-1$
In this case (2.10) is

$$\begin{array}{rcl} \langle \phi_i, \phi_{j+1}\rangle &=& \langle \phi_{j-1}, x\phi_j\rangle - \alpha_j\,\langle \phi_{j-1}, \phi_j\rangle - \beta_j\,\|\phi_{j-1}\|^2 \\ &=& -\alpha_j\,\langle \phi_{j-1}, \phi_j\rangle \quad \text{by the definition of } \beta_j \text{ in (2.8)} \end{array}$$

which is zero by $(\mathrm{IH})_j$.

<u>Case III:</u> $i = j$
In this case (2.10) is

$$\begin{array}{rcl} \langle \phi_i, \phi_{j+1}\rangle &=& \langle \phi_j, x\phi_j\rangle - \alpha_j\,\|\phi_j\|^2 - \beta_j\,\langle \phi_j, \phi_{j-1}\rangle \\ &=& -\beta_j\,\langle \phi_j, \phi_{j-1}\rangle \quad \text{by the definition of } \alpha_j \text{ in (2.8)} \end{array}$$

and again this is zero by $(\text{IH})_j$.
Cases I, II and III show that $\langle \phi_i, \phi_{j+1} \rangle = 0$ for $i = 0 : j$, so $\{\phi_i\}_{i=0}^{j+1}$ is an orthogonal set, which means that $(\text{IH})_{j+1}$ holds. It then follows by induction that $\{\phi_i\}_{i=0}^{n}$ is an orthogonal set of polynomials. $\square$

Remarks:

1. The orthogonal set is unique for the specified inner product since each polynomial is monic.

2. To obtain an *orthonormal* basis use $\{\phi_i / \|\phi_i\|\}_{i=0}^{n}$.

3. For practical purposes it is convenient to replace the definition (2.8) for β_j by
$$\beta_j = \|\phi_j\|^2 / \|\phi_{j-1}\|^2 .$$
To see that this is equivalent, note from (2.8) that
$$\begin{aligned} \|\phi_{j-1}\|^2 \beta_j &= \langle \phi_{j-1}, x\phi_j \rangle = \langle x\phi_{j-1}, \phi_j \rangle \\ &= \langle \phi_j + \text{ polynomial of degree less than } j,\ \phi_j \rangle \\ &\qquad \text{since } x\phi_{j-1} \text{ is a monic poly of degree } j \\ &= \langle \phi_j, \phi_j \rangle + \langle \text{poly of degree} < j,\ \phi_j \rangle \\ &= \langle \phi_j, \phi_j \rangle \ \text{ from orthogonality} \\ &= \|\phi_j\|^2 . \end{aligned}$$
Hence, $\beta_j = \|\phi_j\|^2 / \|\phi_{j-1}\|^2$.

4. The standard orthogonal polynomials (Legendre, Chebyshev, Hermite and Laguerre) can all be obtained in this way using an appropriate weighted L_2 inner product (and scaling by a suitable constant).

Algorithm 2.3: Three-term recurrence for orthogonal polynomials

1. Let $\phi_0(x) = 1$, $\alpha_0 = \langle \phi_0, x\,\phi_0 \rangle / \|\phi_0\|^2$, $\phi_1(x) = (x - \alpha_0)\,\phi_0(x) = x - \alpha_0$ and set $j = 1$.

2. Set
$$\alpha_j = \langle \phi_j, x\,\phi_j \rangle / \|\phi_j\|^2 , \quad \beta_j = \|\phi_j\|^2 / \|\phi_{j-1}\|^2$$
and $\phi_{j+1}(x) = (x - \alpha_j)\,\phi_j(x) - \beta_j\,\phi_{j-1}(x)$.

3. Set j:= j + 1.

4. If $j < n - 1$ go to 2, otherwise stop.

Examples 2.8 Use Algorithm 2.3 to find a basis for $\mathbb{P}_2$ which is orthogonal with respect to the inner product

$$\langle u, v \rangle = \int_{-1}^{1} |x|\, u(x)\, v(x)\, dx$$

and hence obtain the best approximation to $\sqrt{|x|}$ from $\mathbb{P}_2$ with respect to the induced norm.

Legendre polynomials

These are orthogonal in $L_2(-1,1)$, i.e. use weight function $\omega(x) \equiv 1$. We use Algorithm 2.3 to find the first few basis functions. We have $\phi_0(x) = 1$, $\|\phi_0\|^2 = 2$, $\alpha_0 = 0$, $\phi_1(x) = x$, and

$j = 1$: $\alpha_1 = 0$, $\beta_1 = 1/3$, $\phi_2(x) = x^2 - 1/3$;

$j = 2$: $\alpha_2 = 0$, $\beta_2 = 4/15$, $\phi_3(x) = x^3 - 3/5$;

$j = 3$: $\alpha_3 = 0$, $\beta_3 = 9/35$, $\phi_4(x) = x^4 - (6/7)x^2 + 3/35$; etc.

Note that the usual formula for Legendre polynomials is a scaled version of the above. The nth degree Legendre polynomial is denoted by $P_n(x)$, and the scaling is chosen so that each $P_n(1) = 1$ – it can be shown that

$$P_n(x) = \frac{(2n)!}{(n!)^2\, 2^n}\, \phi_n(x)\,.$$

The first few Legendre polynomials are

$$\begin{aligned} &P_0(x) = 1\,, \quad P_1(x) = x\,, \\ &P_2(x) = \frac{3}{2}\,x^2 - \frac{1}{2}\,, \quad P_3(x) = \frac{5}{2}\,x^3 - \frac{3}{2}\,x\,, \\ &P_4(x) = \frac{35}{8}\,x^4 - \frac{15}{4}\,x^2 + \frac{3}{8}\,, \ldots \end{aligned}$$

and they satisfy the recurrence relation

$$(n+1)\, P_{n+1}(x) = (2n+1)\, x\, P_n(x) - n\, P_{n-1}(x) \quad \text{for } n \geq 1$$

and the orthogonality relation

$$\int_{-1}^{1} P_m(x)\, P_n(x)\, dx = \begin{cases} 0 & \text{if } m \neq n \\ \frac{2}{n+1} & \text{if } m = n. \end{cases}$$

They can also be obtained from the formula

$$P_n(x) = \frac{1}{2^n n!}\, \frac{d^n}{dx^n}\, \left(x^2 - 1\right)^n\,.$$

Chebyshev polynomials

The Chebyshev polynomial of the first kind of degree n is written as $T_n(x)$ and can be obtained from Algorithm 2.3 using the weighted inner product with weight function $\omega(x) = 1/\sqrt{1-x^2}$ on the interval $(-1, 1)$. They are then scaled so that

$$T_n(1) = 1 \quad \text{for } n \geq 0.$$

The first few are

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2\,x^2 - 1 \\ T_3(x) &= 4\,x^3 - 3\,x \\ \cdots &= \ldots \\ T_{n+1}(x) &= 2\,x\,T_n(x) - T_{n-1}(x) \quad \text{for } n \geq 1. \end{aligned}$$

Surprisingly they can be written as

$$T_n(x) = \cos\left(n\,\cos^{-1}(x)\right) \quad \text{for } n \geq 0 \tag{2.11}$$

which leads at once to the useful property

$$-1 \leq T_n(x) \leq 1 \quad \text{for all } x \in [-1, 1].$$

Using the substitution $x = \cos\theta$ gives

$$\langle T_n, T_n \rangle = \int_{-1}^{1} \frac{(T_n(x))^2}{\sqrt{1-x^2}}\,dx = \int_0^{\pi} \cos^2(n\theta)\,d\theta = \begin{cases} \pi & \text{if } n = 0 \\ \pi/2 & \text{if } n > 0 \end{cases}$$

which gives the orthogonality relation

$$\langle T_n, T_m \rangle = \begin{cases} 0 & \text{if } m \neq n \\ \pi & \text{if } m = n = 0 \\ \pi/2 & \text{if } m = n > 0\,. \end{cases}$$

Approximation on $[-1, 1]$ by Chebyshev polynomials

If p_n^* is the best approximation in $\mathbb{P}_n$ in terms of the weighted inner product with $\omega(x) = (1-x^2)^{-\frac{1}{2}}$ to a function $u(x)$ then it follows from (2.5) that

$$p_n^* = \sum_{j=0}^{n} c_j^*\,T_j(x)\,,$$

where

$$c_j^* = \frac{\langle u, T_j \rangle}{\langle T_j, T_j \rangle}\,.$$

The formula for $\langle T_j, T_j \rangle$ is given above and

$$\langle u, T_j \rangle = \int_{-1}^{1} \frac{u(x)\, T_j(x)}{\sqrt{1-x^2}}\, dx = \int_0^{\pi} u(\cos\theta)\, \cos(j\theta)\, d\theta\,.$$

Hence, we can write p_n^* as

$$p_n^*(x) = \frac{1}{2}\, d_0^*\, T_0(x) + \sum_{j=1}^{n} d_j^*\, T_j(x) \tag{2.12}$$

where

$$d_j^* = \frac{2}{\pi} \int_0^{\pi} u(\cos\theta)\, \cos(j\theta)\, d\theta \qquad \text{for } j \geq 0.$$

Example: To approximate $u(x) = e^x$ on $[-1, 1]$ by Chebyshev polynomials in $\mathbb{P}_4$ (i.e. this is the best approximation with the weighted L_2 norm with $\omega(x) = (1-x^2)^{-\frac{1}{2}}$) we have (2.12) with $n = 4$ with

$$d_j^* = \frac{2}{\pi} \int_0^{\pi} e^{\cos\theta}\, \cos(j\theta)\, d\theta\,.$$

Numerical integration for d_j^* yields

$$d_0^* = 2.532132,\ d_1^* = 1.130318,\ d_2^* = 0.271495,\ d_3^* = 0.044338,\ d_4^* = 0.005474.$$

This gives

$$p_4^*(x) = 1.000045 + 0.997308\, x + 0.499197\, x^2 + 0.177347\, x^3 + 0.043794\, x^4\,.$$

If we want the best approximation of u in the L_2-norm (i.e. with weight $\omega(x) \equiv 1$), then this is given in terms of Legendre polynomials by

$$q_4^*(x) = \sum_{j=0}^{4} c_j^*\, P_j(x)\,, \quad \text{where} \quad c_j^* = \frac{\langle u, P_j \rangle}{\|P_j\|^2}\,.$$

For $u(x) = e^x$ it can be shown (exercise) that

$$q_4^*(x) = 1.000031 + 0.997995\, x + 0.499350\, x^2 + 0.176139\, x^3 + 0.043600\, x^4\,.$$

(Note: It is useful to use Maple to calculate all the integrals, or if doing it by hand to use a reduction formula for $I_n = \int_{-1}^{1} x^n\, e^x\, dx$.)

2.6 Best L_2-approximation in terms of trigonometric functions

So far we have only considered approximating a function $u(x)$ by a sum of polynomials, but another important case is approximation in terms of $\sin kx$

and $\cos kx$. Suppose that u is a **periodic function** and for convenience we let the period be 2π. This means that

$$u(x + 2\pi) = u(x) \quad \text{for any } x.$$

Fix $n > 0$ and let $\mathcal{A}_n$ be the subset of $L_2(-\pi, \pi)$ which is spanned by the functions

$$\{1, \cos x, \sin x, \cos 2x, \sin 2x, \ldots, \cos nx, \sin nx\} .$$

Examples 2.9 Show that this set of functions is **orthogonal** in $L_2(-\pi, \pi)$. (In fact they are orthogonal on any contiguous interval of length 2π.)

Because an orthogonal set is linearly independent, these functions form a basis for $\mathcal{A}_n$, which hence has dimension $N = 2\,n + 1$. It follows from (2.5) that the best approximation of u in $\mathcal{A}_n$ is

$$u_n := \sum_{j=1}^{N} \frac{\langle u, \phi_j \rangle}{\langle \phi_j, \phi_j \rangle} \, \phi_j$$

where $\{\phi_j\}_{j=1}^{N} = \{1, \cos x, \sin x, \ldots, \cos nx, \sin nx\}$.
This gives:

$$u_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n} \Big(a_k \cos(kx) + b_k \sin(kx) \Big) , \tag{2.13}$$

where

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} u(x) \cos(kx)\, dx , \quad k = 0 : n , \tag{2.14}$$

and

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} u(x) \sin(kx)\, dx , \quad k = 1 : n . \tag{2.15}$$

The limit as $n \to \infty$ of (2.13) with (2.14) and (2.15) gives an infinite number of terms and is called the **Fourier series** of u.
If the function $u(x)$ is smooth, then very few terms are needed in order to get a good approximation, as shown in Figure 1.

Examples 2.10 Find the best approximation u_n of the function $u(x) = |x|$ in $L_2(-\pi, \pi)$.
This function and some approximations u_n are plotted in Figure 2. Note that the error is worst at the end of the interval and at $x = 0$, where $u(x)$ is not smooth.

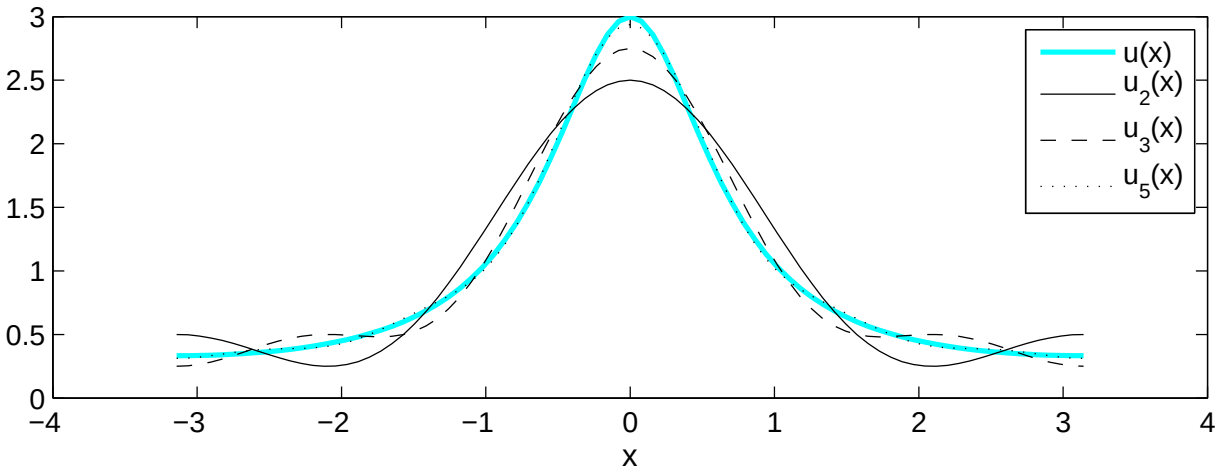

Figure 1: Plot of $u(x) = 3/(5-4\cos x)$ and its best approximation from $\mathcal{A}_n$ for $n = 2$, 3 and 5.

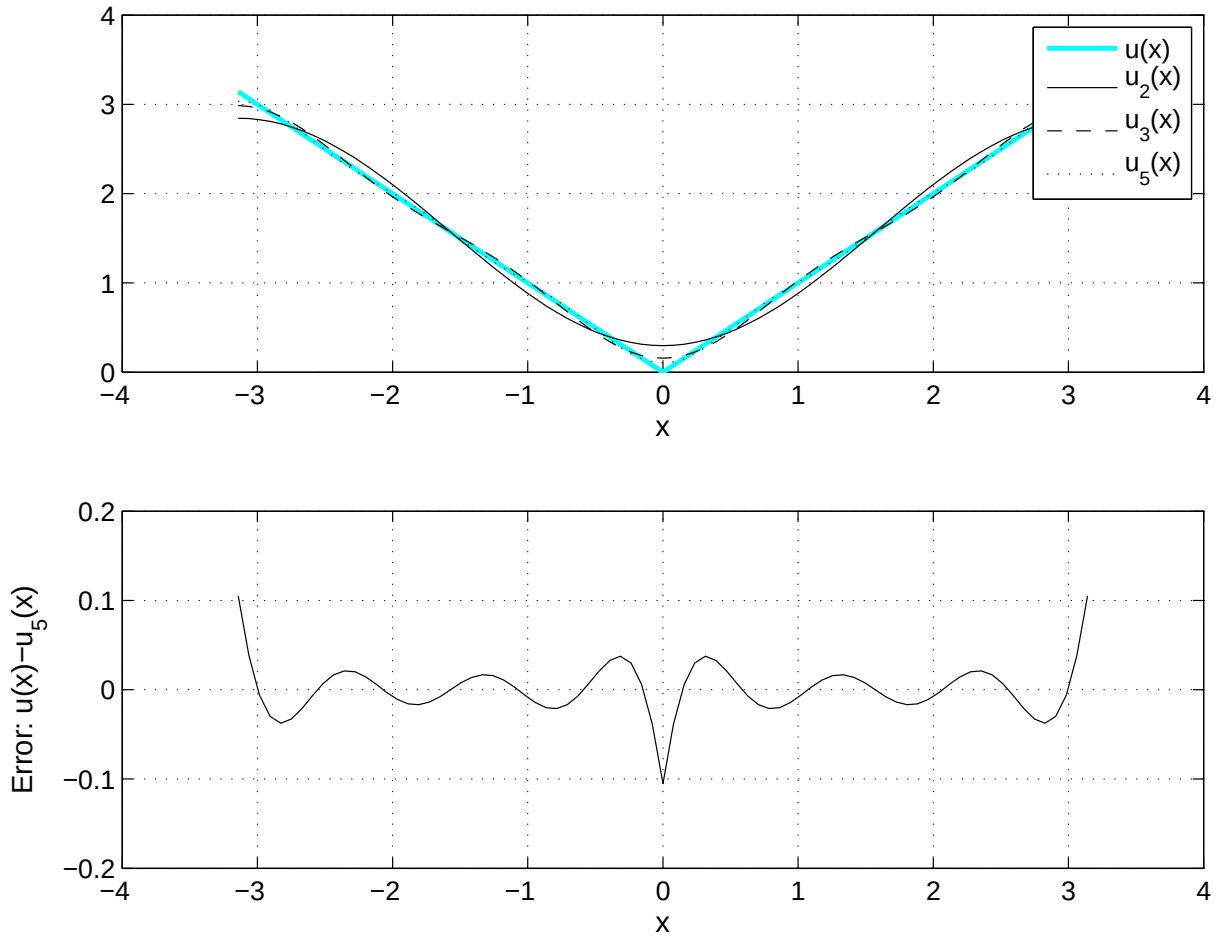

Figure 2: Plot of $u(x) = |x|$ and its best approximation from $\mathcal{A}_n$ for $n = 2$, 3 and 5. The lower plot shows the error $u(x) - u_5(x)$.

Figure 3 shows the approximation u_n when $n = 100$ for the discontinuous function

$$u(x) = \begin{cases} -1 & \text{when } x \in (-\pi, 0] \\ 1 & \text{when } x \in (0, \pi]. \end{cases} \tag{2.16}$$

The oscillations about the points of discontinuity ($x = 0$ and $x = \pm 1$) persist no matter how many terms are taken: this is called the **Gibbs phenomenon**.

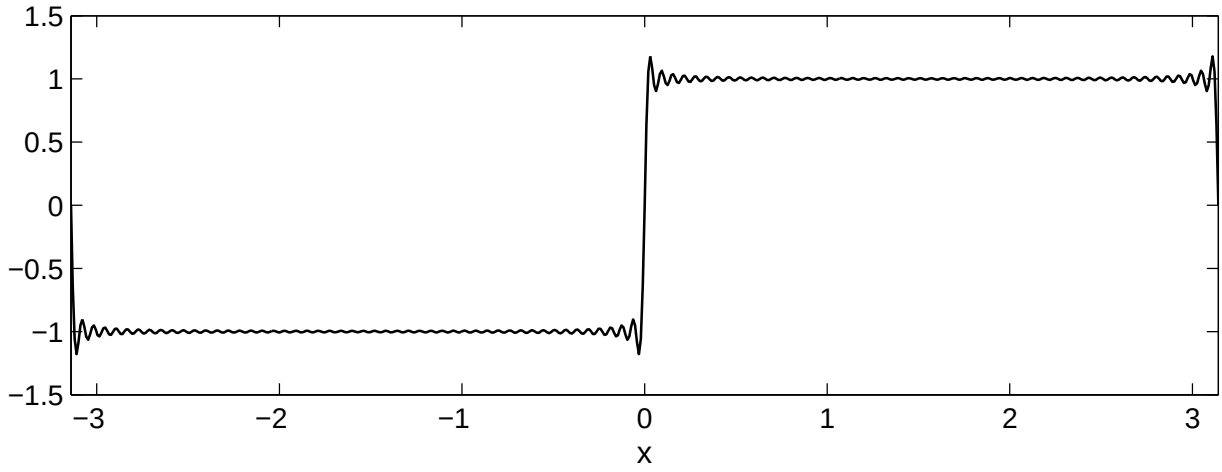

Figure 3: Plot of the best approximation to the discontinuous function u defined in (2.16) from $\mathcal{A}_n$ with $n = 100$.

2.7 Exercises for Section 2

2–1. Let $\mathcal{V} = \mathbb{R}^2$, and $v = (3, -1)^T$. Determine all best approximations of v from $\mathcal{A} = \{(x, y)^T : y = 0\}$ in both the ℓ_2- and ℓ_∞-norms.

2–2. Find the best approximation of $u(x)$ on the interval $[-1, 1]$ by a constant function in each of the L_2- and L_∞-norms, where $u(x)$ is defined by

$$u(x) = \begin{cases} 0, & \text{for } \varepsilon \le |x| \le 1, \\ 1 - \frac{|x|}{\varepsilon}, & \text{for } |x| < \epsilon\,. \end{cases}$$

Here, $0 < \varepsilon < 1$ is a constant.

2–3. Assume u is even (odd). Show that it has an even (odd) best L_p-norm approximation out of $\mathbb{P}_n$ on any interval $[-a, a]$ symmetric about the origin.

2–4. Find the error of the best L_2-norm approximation to $u(x) = x^3$ on $[-1, 1]$ out of $\mathbb{P}_2$, and verify Jackson's theorem for u with $r = 1$, $n = 2$, $C_1 = \pi/4$. Note that the computation of the best approximation is greatly simplified by using the result of the previous question.

2–5. Find the best approximation to the (even) function $u(x) = x^4$ in $\mathbb{P}_3$ in the L_2 norm on $(-1, 1)$.

2–6. Use Alg. 1 to find the best approximation to the (even) function $u(x) = \pi \cos(\pi x)$ in $L_2(-1, 1)$ from the spaces $\mathbb{P}_0$, $\mathbb{P}_1$ and $\mathbb{P}_2$.

2–7. Let $u(x) = e^x \in L_2(-1, 1)$ and use the orthonormal basis $\{\phi_j\}$ from Examples 2.5 in Alg. 1 to calculate p_2^*, the best approximation to u from $\mathbb{P}_2$. Show that p_2^* satisfies

$$\|u - p_2^*\|_{L_2} < 0.05\,.$$

2–8. Let $u(x) = x^4 \in L_2(-1, 1)$ and use the orthonormal basis $\{\phi_j\}$ from Examples 2.5 in Alg. 1 to find the best approximation from the space $\mathbb{P}_2$.

2–9. Apply the Gram-Schmidt process to generate an orthogonal basis for $\mathbb{P}_2 \subset L_2(0, 1)$ starting from the basis 1, x, x^2.

2–10. Apply the Gram-Schmidt process to generate an orthogonal basis for $\mathbb{P}_2$ with respect to the inner product

$$\langle u, v\rangle = \int_0^1 x\, u(x)\, v(x)\, dx$$

starting from the basis 1, x, x^2.

2–11. Use Algorithm 2.3 to find an orthogonal basis for $\mathbb{P}_2$ with respect to the inner product

$$\langle u, v\rangle = \int_0^1 x\, u(x)\, v(x)\, dx\,.$$

2–12. Use Algorithm 2.3 to find an orthogonal basis for $\mathbb{P}_2$ with respect to the inner product

$$\langle u, v\rangle = \int_{-1}^1 \omega(x)\, u(x)\, v(x)\, dx$$

with weight function $\omega(x) = \sqrt{1 - x^2}$. [if you do not use Maple for the integrals, then the following identities will be useful: $I_0 = \pi/2$, $I_1 = 0$, $I_n = I_{n-2}(n-1)/(n+2)$, where $I_n = \int_{-1}^1 x^n \sqrt{1 - x^2} dx$.]

2–13. Show that the Chebyshev polynomials defined by

$$T_n(x) = \cos(n \cos^{-1} x), \qquad -1 \le x \le 1,$$

satisfy the 3-term recurrence relation $T_0(x) = 1$, $T_1(x) = x$, $T_{n+1}(x) = 2\,x\,T_n(x) - T_{n-1}(x)$, $n \ge 1$. [Note: Replace x by $\cos\theta$ and use standard trigonometric idenities.] Hence, deduce that for $n \ge 1$, $T_n(x)$ is a polynomial of degree n with leading coefficient 2^{n-1}.

2–14. Find the best approximation to the function $u(x) = x^4$ in $\mathbb{P}_3$ in the induced norm corresponding to the L_2 inner product on $(-1, 1)$ with weight function $\omega(x) = (1 - x^2)^{-1/2}$.

2–15. The *Laguerre polynomials* $L_n(x)$ are orthogonal in the weighted $L_2(0, \infty)$ space with weight function $\omega(x) = e^{-x}$, and the leading coefficient of $L_n(x)$ is set to $(-1)^n/n!$. Use Algorithm 2.3 to calculate $L_n(x)$ for $n = 0 : 3$. (It would be sensible to use a package such as Maple to evaluate the integrals, but doing them by hand, then note that $\int_0^\infty x^n\, e^{-x}\, dx = n!$.)

Find the best approximation to $u(x) = x^4$ from $\mathbb{P}_2$ in the induced norm for this space.

2–16. The *Hermite polynomials* $H_n(x)$ are orthogonal in the weighted $L_2(\mathbb{R})$ space with weight function $\omega(x) = e^{-x^2}$, and the leading coefficient of $H_n(x)$ is set to 2^n. Use Algorithm 2.3 to calculate $H_n(x)$ for $n = 0 : 2$. (Note, you will probably need to use an algebraic manipulation package such as Maple to evaluate the integrals.)

2–17. Find the best trigonometric approximation $u_n(x)$ with $n = 2$ to the function $u(x) = \pi e^x$ in $L_2(-\pi, \pi)$. You may assume:

$$\int e^x\, \sin(kx)\, dx = \frac{e^x\, (\sin(kx) - k\, \cos(kx))}{1 + k^2}\,,$$

$$\int e^x\, \cos(kx)\, dx = \frac{e^x\, (k\, \sin(kx) + \cos(kx))}{1 + k^2}\,.$$

2–18. Find the best trigonometric approximation $u_n(x)$ with $n = 1$ to the function $u(x) = \pi\, (x - \pi)$ in $L_2(-\pi, \pi)$.

3 Interpolation

Consider the $n+1$ real-valued **data points** (x_k, y_k) for $k = 0 : n$ which are arranged so that the **nodes** satisfy $x_0 < x_1 < \cdots < x_n$. The **interpolation problem** is to find a function p which satisfies $p(x_k) = y_k$ for $k = 0 : n$. Such a function is said to **interpolate the data** – it passes through all $n+1$ data points. The Weierstrass and Jackson Theorems from Section 2 provide us with some evidence that polynomials might be a good place to start when seeking an interpolant, and we shall do just that.

3.1 Polynomial interpolation

The **polynomial interpolation problem** is as follows. Given $n \geq 0$ and the $n+1$ data points (x_k, y_k) for $k = 0 : n$ with distinct nodes, find the polynomial $p \in \mathbb{P}_n$ for which

$$p(x_k) = y_k \quad \text{for } k = 0 : n\,. \tag{3.1}$$

Theorem 3.1 (Lagrange) *The polynomial $p \in \mathbb{P}_n$ defined in (3.1) above exists and is unique.*

Proof. We give various different ways of finding a polynomial $p \in \mathbb{P}_n$ to satisfy (3.1) below, and this shows that it exists. Here we shall prove that it is unique. Suppose for a contradiction that (3.1) is satisfied by two different polynomials p and q in $\mathbb{P}_n$. Then $p(x_k) = y_k$ and $q(x_k) = y_k$ for $k = 0 : n$, and so if we set $u = p - q$, then $u \in \mathbb{P}_n$ satisfies

$$u(x_k) = 0 \quad \text{for } k = 0 : n\,.$$

That is, u is a polynomial of degree at most n which has $n+1$ zeros. The only way this is possible is for $u(x)$ to be zero everywhere, since polynomials of degree n have at most n distinct zeros (by the fundamental theorem of algebra). Hence $p(x) = q(x)$ for all x, and the assumption that there is more than one interpolating polynomial is wrong. □

If $p \in \mathbb{P}_n$ then

$$p_n(x) = a_0 + a_1\,x + a_2\,x^2 + \cdots + a_n\,x^n$$

and we can determine p by finding each of the coefficients a_j. The interpolation conditions (3.1) can be written as the linear system $M\,\boldsymbol{a} = \boldsymbol{y}$:

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

in unknown coefficients $\boldsymbol{a} = (a_0, \ldots, a_n)^T$. The coefficient matrix M is called a *Vandermonde* matrix. It is known that a Vandermonde matrix is always nonsingular (when the nodes x_j are distinct), and so p can be found by solving the system. However, the matrix M is **severely ill-conditioned** for larger n, and this approach is not recommended in practice.

An alternative (and much better) approach is to use Lagrange polynomial basis functions $l_j(x)$ for $\mathbb{P}_n$. Recall that these are defined on the node points $\{x_k\}_{k=0}^n$ and $l_j(x)$ is the (unique) nth degree polynomial with the property

$$l_j(x_k) = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

i.e. $l_j(x_k) = \delta_{j,k}$. It is written explicitly as

$$l_j(x) = \frac{(x - x_0)\,(x - x_1)\,\ldots\,(x - x_{j-1})(x - x_{j+1})\ldots(x - x_n)}{(x_j - x_0)(x_j - x_1)\ldots(x_j - x_{j-1})(x_j - x_{j+1})\ldots(x_j - x_n)}\,,$$

or more compactly as

$$l_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{(x - x_i)}{(x_j - x_i)}\,.$$

It follows from (3.2) that the interpolating polynomial $p(x)$ is

$$p(x) = \sum_{j=0}^{n} y_j\, l_j(x)\,. \tag{3.3}$$

Examples 3.1 Find the interpolating polynomial $p \in \mathbb{P}_2$ for the data $\{(0,1), (1,3), (3,2)\}$ and evaluate $p(2)$.

Theorem 3.2 (Interpolation error) *Suppose that $f \in C^{n+1}[a,b]$, let $a = x_0 < x_1 < \cdots < x_n = b$ and let $p \in \mathbb{P}_n$ be the polynomial which interpolates f at the nodes x_k, i.e.*

$$p(x_k) = f(x_k) \quad \textit{for } k = 0 : n\,.$$

Then for each $x \in [a,b]$ there exists $\xi \in (a,b)$ which depends on x, such that

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\pi_{n+1}(x)\,,$$

where

$$\pi_{n+1}(x) := (x - x_0)\ldots(x - x_n) = \prod_{k=0}^{n}(x - x_k)\,.$$

Runge's interpolation example: uniform vs Chebyshev spacing

Apart from ill-conditioning, there are other problems with polynomial interpolation. Unfortunately, if the degree of the polynomial is high, then the error between the interpolation polynomial and the function can be astonishingly large, even for harmless-looking functions and natural choices of interpolation nodes. C. Runge in 1901 came up with the following example to illustrate the dangers of polynomial interpolation. Consider the function

$$R(x) = \frac{1}{1 + 25x^2}$$

on the interval $[-1, 1]$. Let $p_n(x) \in \mathbb{P}_n$ denote the polynomial interpolant of $R(x)$ which uses the **equally spaced** points $-1 = x_0 < x_1 < \ldots < x_n = 1$. It can be shown that for $0.726 \approx< |x| < 1$ the polynomial $p_n(x)$ does not even converge to $R(x)$ as $n \to \infty$. The upper plots in Figure 4 illustrate this effect.

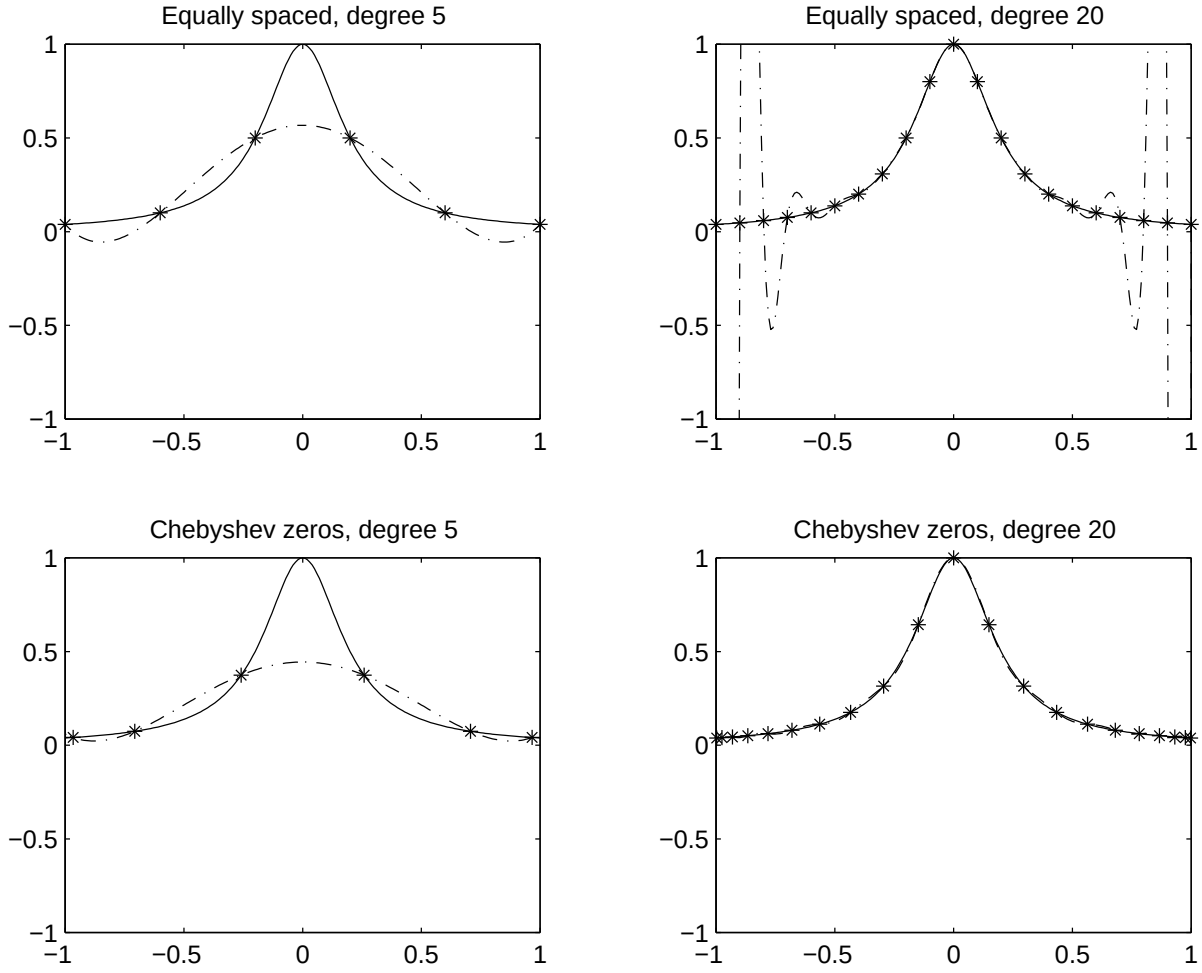

Figure 4: Polynomial interpolants on Runge's function.

Solid and dash-dot linetypes are used for the function $R(x)$ and the interpolant, respectively. The interpolation data is marked with asterisks. In the upper

right picture, which corresponds to $n = 20$, we see that severe oscillations have arisen in $p_n(x)$ between the interpolation points at the ends of the range. This is typical of high degree interpolants based on equally spaced points.

The lower pictures in Figure 4 illustrate the interpolants that arise when the interpolation points are taken as the zeros of the **Chebyshev polynomial** $T_{n+1}(x)$ of degree $n+1$. For this $R(x)$ it can be shown that for every $x \in [-1, 1]$, we have $p_n(x) \to R(x)$ as $n \to \infty$. It is indeed known that these Chebyshev nodes are in some sense (almost) optimal for interpolation.

However, in a practical problem of data interpolation we usually cannot choose the data locations freely – e.g. experimental measurements are often done at equidistant time intervals, and it is usually more reliable to use **splines** to interpolate data, rather than polynomials.

3.2 Interpolation by linear splines

An alternative and more flexible way of interpolating a set of data values $\{(x_k, y_k)\}_{k=0}^n$ with $x_0 < x_1 < \cdots < x_n$ is to look for a polynomials of low degree in each subinterval (x_{k-1}, x_k) which collectively interpolate the data. Such piecewise polynomial approximations are called **splines** and the endpoints x_k of the subintervals are called **knots**.

Definition 3.8 A spline of degree $r \geq 1$ on the knots $a = x_0 < x_1 < \cdots < x_n = b$ is a continuous function which is polynomial of degree r or less in each subinterval (x_{k-1}, x_k) for $k = 1 : n$.

Theorem 3.3 *The set of splines of degree $r \geq 1$ on the knots $a = x_0 < x_1 < \cdots < x_n = b$ is a linear subspace of $C[a, b]$.*

Proof. By definition all such splines are elements of $C[a, b]$, and so we just have to check that a linear combination of splines of degree r is also a spline of degree r. Given splines $S(x)$ and $T(x)$, each is a polynomial of degree at most r in $[x_{k-1}, x_k]$, and hence so is $\alpha S + \beta T$ for any real constants α and β. □

We shall concentrate on low-order cases: in this section we look at properties of linear splines ($r = 1$), and we then consider smoother interpolants, focussing on cubic splines ($r = 3$).

The simplest way to interpolate the data $\{(x_k, y_k)\}_{k=0}^n$ is just to join the data points by straight lines, and this is **piecewise linear interpolation**, or interpolation by a linear spline. It is illustrated in Figure 5 and the precise formula is given below.

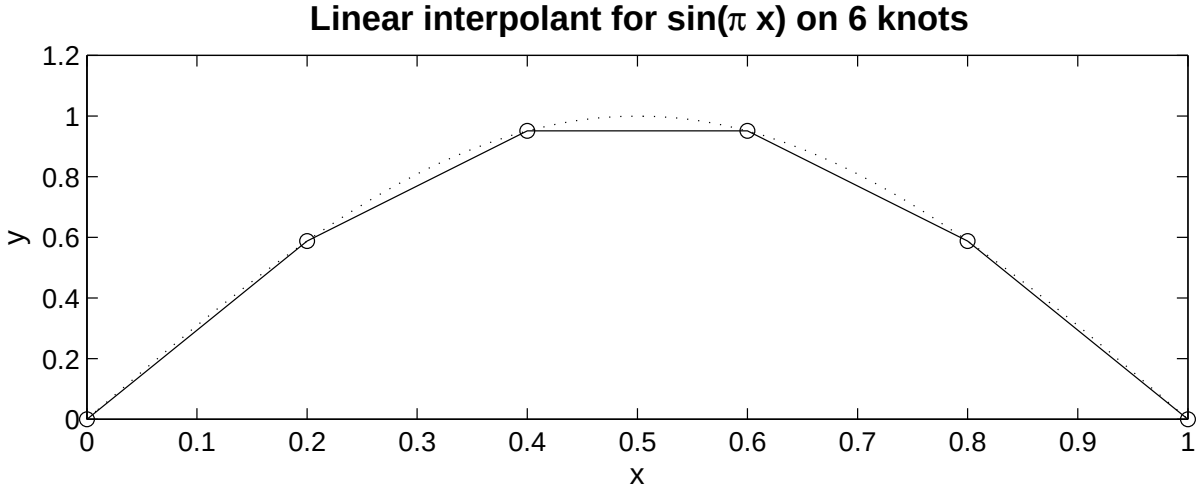

Figure 5: Piecewise linear spline interpolant (solid line) for the function $\sin(\pi x)$ (dotted line) on the 6 equally spaced points shown.

Definition 3.9 The linear spline which interpolates the data $\{(x_k, y_k)\}_{k=0}^n$ on the knots $a = x_0 < x_1 < \cdots < x_n = b$ is $S(x)$ defined on each subinterval by

$$S(x) = \left(\frac{x_k - x}{x_k - x_{k-1}}\right) y_{k-1} + \left(\frac{x - x_{k-1}}{x_k - x_{k-1}}\right) y_k \quad \text{when } x \in [x_{k-1}, x_k].$$

Examples 3.2 Show $S(x_k) = y_k$ for $k = 0 : n$, and hence that $S \in C[a, b]$.

Note that $S(x)$ can also be defined in terms of piecewise linear **basis functions**. We define the *hat function* ϕ (shown in Figure 6) by

$$\phi(x) = \begin{cases} 1 - |x| & \text{if } |x| \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The linear spline basis functions are $\{\phi_k(x)\}_{k=0}^n$ given in terms of ϕ by:

$$\phi_k(x) = \begin{cases} \phi\left((x - x_k)/(x_{k+1} - x_k)\right) & \text{if } x \ge x_k \\ \phi\left((x - x_k)/(x_k - x_{k-1})\right) & \text{if } x \le x_k\,. \end{cases}$$

Note that $\phi_k(x_k) = 1$ and ϕ_k is zero at all the other knots, and an alternative definition is

$$\phi_k(x) = \begin{cases} (x - x_{k-1})/(x_k - x_{k-1}) & \text{for } x \in [x_{k-1},\, x_k] \\ (x_{k+1} - x)/(x_{k+1} - x_k) & \text{for } x \in (x_k,\, x_{k+1}] \\ 0 & \text{otherwise.} \end{cases}$$

Examples 3.3 Show that the linear spline interpolant is $S(x) = \sum_{j=0}^{n} y_j\, \phi_j(x)\,.$

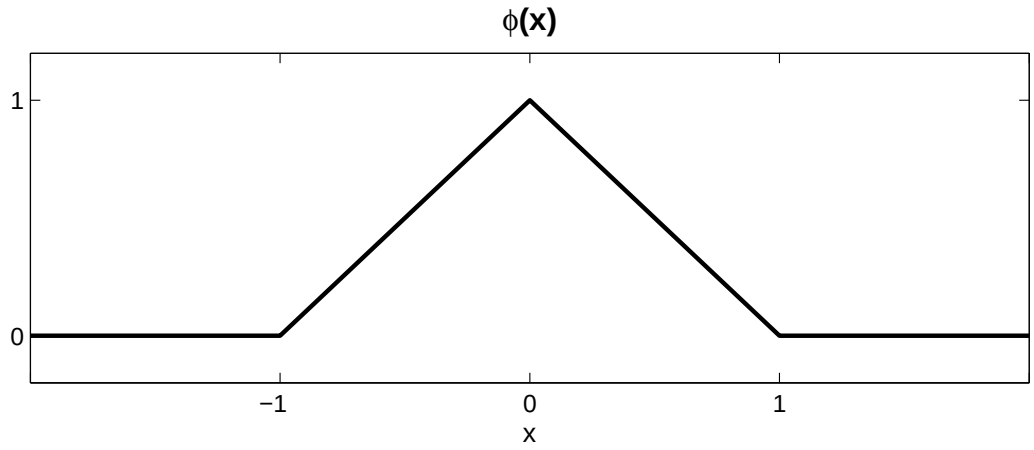

Figure 6: The hat function $\phi(x)$.

The following result gives a bound for the maximum error between a function and its piecewise linear interpolant.

Theorem 3.4 *Suppose that $f \in C^2[a,b]$ and let S be the linear spline which interpolates f at the knots $a = x_0 < x_1 < \cdots < x_n = b$. Then*

$$\|f - S\|_{L_\infty} \leq \frac{1}{8}\, h^2\, \|f''\|_{L_\infty}\,,$$

where $h = \max_k (x_k - x_{k-1})$.

Proof. Consider a subinterval $[x_{k-1}, x_k] \subset [a,b]$. In this subinterval $S(x)$ is a linear polynomial, so it follows from Theorem 3.2 that there is some ξ for which

$$f(x) - S(x) = \frac{f''(\xi)}{2}\,(x - x_{k-1})\,(x - x_k) \quad \text{for } x \in [x_{k-1}, x_k]\,.$$

Note that $(x - x_{k-1})\,(x - x_k)$ is non-positive on the subinterval and achieves its minimum value of $-(x_k - x_{k-1})^2/4$ at $x = (x_{k-1} + x_k)/2$. Hence if $x \in [x_{k-1}, x_k]$ then

$$|f(x) - S(x)| \leq \frac{(x_k - x_{k-1})^2}{8} \max_{\xi \in [x_{k-1}, x_k]} |f''(\xi)| \leq \frac{h^2}{8}\, \|f''\|_{L_\infty}\,,$$

as required. □

If S is a spline on $[a,b]$, then $S'(x)$ is a (discontinous) piecewise constant function, and so $S \notin C^1[a,b]$. However, it can be shown that it is in the space $H^1(a,b)$ defined in Section 1.5. The following result shows that it is the interpolant with smallest energy (semi-)norm.

Theorem 3.5 *Suppose that S is the linear spline which interpolates the data $\{(x_k, y_k)\}_{k=0}^n$ at the knots $a = x_0 < x_1 < \cdots < x_n = b$. Then for any $v \in H^1(a,b)$ which also interpolates the data at these knots,*

$$\|S'\|_{L_2} \leq \|v'\|_{L_2}\,.$$

Proof. The interpolation condition is $v(x_k) = S(x_k) = y_k$ for $k = 0 : n$. Note that

$$\|v'\|_{L_2}^2 = \|v' - S'\|_{L_2}^2 + \|S'\|_{L_2}^2 + 2\,\langle v' - S',\, S'\rangle_{L_2}$$

and look at the last term on the RHS:

$$\langle v'-S',\, S'\rangle_{L_2} = \int_a^b \{v'(x) - S'(x)\}\; S'(x)\,dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \{v'(x) - S'(x)\}\; S'(x)\,dx\,.$$

Integrating by parts gives

$$\int_{x_{k-1}}^{x_k} \{v'(x) - S'(x)\}\, S'(x)\,dx = \Big[\{v(x) - S(x)\}\, S'(x)\Big]_{x_{k-1}}^{x_k} - \int_{x_{k-1}}^{x_k} \{v(x) - S(x)\}\, S''(x)\,dx$$

and each of these terms is zero – the first because $v(x) = S(x)$ when x is a knot (interpolation), and the second because S is a linear function in the subinterval $[x_{k-1}, x_k]$, and so its second derivative is identically zero in this subinterval. Thus

$$\|v'\|_{L_2}^2 = \|v' - S'\|_{L_2}^2 + \|S'\|_{L_2}^2 \geq \|S'\|_{L_2}^2$$

as required. □

Remark: The proof is slightly more subtle than it looks: the identity

$$\int_a^b \{v'(x) - S'(x)\}\; S'(x)\,dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \{v'(x) - S'(x)\}\; S'(x)\,dx$$

only holds because the function $\{v'(x) - S'(x)\}\; S'(x)$ is integrable on (a, b) because both S and v are in the space $H^1(a, b)$. It does **not** hold in general for functions S and v which are not in $H^1(a, b)$.

3.3 More smoothness: cubic splines

Piecewise linear interpolation has the advantage over polynomial interpolation in that it does not introduce oscillations, but it has the drawback in not being smooth; the first derivative is discontinuous at each knot. We can get around this difficulty by using higher degree piecewise interpolants.

We use a **counting argument** to determine the degree of continuity which can be obtained from a spline S which is a piecewise polynomial interpolant on an interval $[a, b]$. Suppose that there are $n + 1$ knots, $a = x_0 < x_1 < \cdots < x_n = b$, and that there is a piecewise polynomial interpolant of degree up to r on each subinterval $[x_{k-1}, x_k]$. Each of these has possible $r + 1$ degrees of freedom (the coefficients of x^j for $j = 0 : r$) and there are n intervals, which gives

$$\text{total degrees of freedom } = n \times (r + 1)\,.$$

Now look at the constraints. The first is that the spline **interpolates** the given data, i.e. $S(x)$ is prescribed at both ends of each subinterval, which gives $2\,n$ constraints. This ensures that the spline is continuous, but if we also want it to be in $C^1[a,b]$ we need to make sure that the derivative of S matches at each internal knot: this gives $n-1$ constraints for each derivative we wish to match. So for $S \in C^m[a,b]$,

$$\text{total number of constraints } = \underbrace{2\,n}_{\text{interpolation}} + \underbrace{m \times (n-1)}_{\text{derivatives}} = (m+2)\,n - m\,.$$

So to ensure that $S \in C^m[a,b]$ we need to use piecewise polynomials of degree r for which

$$(r+1)\,n \geq (m+2)\,n - m$$

for arbitrarily large n – i.e. we can take $r = m+1$, and there will then be an 'extra' m (or $r-1$) free degrees of freedom.

Examples:

1. A linear spline ($r = 1$) can only have $m = 0$, and there are no 'extra' degrees of freedom.
2. If S is a cubic spline ($r = 3$) then $S \in C^2[a,b]$, and there are two 'extra' degrees of freedom.

We shall focus on cubic splines in this section, and give their full definition below. Figure 7 shows the cubic spline interpolant for the function from Figure 5. Note how much smoother it is than the linear spline: indeed the word *spline* comes from a device used by architects to produce smooth curves before the era of computers.

Definition 3.10 The function $S(x)$ is a **cubic spline** with respect to the knots $a = x_0 < x_1 < \cdots < x_n = b$, if

(i) S is a polynomial of degree at most three on each interval $[x_{k-1}, x_k]$ for $k = 1 : n$, and

(ii) $S \in C^2[a,b]$ (that is, S, S' and S'' are continuous, or S is twice continuously differentiable).

The set of all cubic splines with respect to a given knot sequence $X = [x_0, x_1, \ldots, x_n]$ will be denoted $\mathcal{S}_X$.

Remark: As shown in Theorem 3.3, the set $\mathcal{S}_X$ is a linear subspace of $C^2[a,b]$.

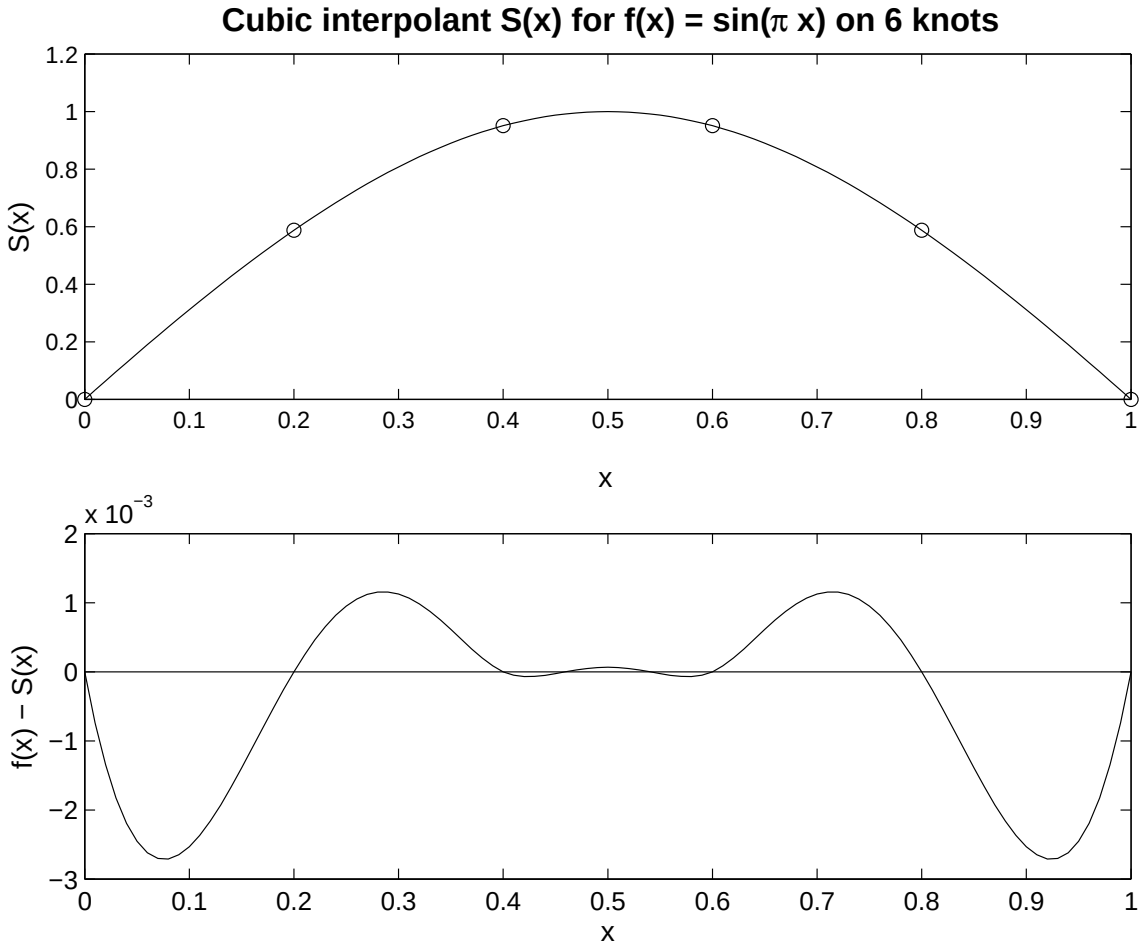

Figure 7: TOP PLOT: cubic spline interpolant $S(x)$ for the function $f(x) = \sin(\pi x)$ on the 6 equally spaced points shown. BOTTOM PLOT: interpolation error $f(x) - S(x)$.

Before considering the question of existence of an interpolating spline, we need to determine the **dimension** of the space $\mathcal{S}_X$, and we do this by explicitly constructing a basis for $\mathcal{S}_X$.

Define

$$(x - x_k)_+^j = \begin{cases} (x - x_k)^j & \text{if } x \geq x_k \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, 2, \ldots,$$

to be **truncated power functions**. Clearly, $(x - x_k)_+^j$ is a polynomial in $\mathbb{P}_j$ on each of the half-axes $(-\infty, x_k)$ and (x_k, ∞). Moreover, it is continuous and it has continuous derivatives at x_k up to the order $j - 1$. Hence $(x - x_k)_+^3$ belongs to $\mathcal{S}_X$.

Theorem 3.6 *Let $X = [x_0, x_1, \ldots, x_n]$ with $x_0 = a$ and $x_n = b$. Every spline $S \in \mathcal{S}_X$ has the following **truncated power expansion**:*

$$S(x) = p_3(x) + \sum_{k=1}^{n-1} \gamma_k \, (x - x_k)_+^3, \qquad x \in [a, b], \tag{3.4}$$

where $p_3 \in \mathbb{P}_3$ is the polynomial which coincides with S on $[x_0, x_1]$, and

$$\gamma_k = a_{k+1} - a_k, \qquad k = 1 : n-1,$$

with a_k being the coefficient of x^3 in the polynomial coinciding with S on $[x_{k-1}, x_k]$.

Proof. Let $S \in \mathcal{S}_X$. Each piece S_k of S can be written as

$$S_k(x) = a_k\, x^3 + b_k\, x^2 + c_k\, x + d_k, \qquad x \in [x_{k-1}, x_k], \quad k = 1 : n,$$

for some real coefficients a_k, b_k, c_k and d_k. Note that S_k''' is a constant:

$$S_k'''(x) = 6\, a_k\,.$$

By definition $p_3(x) = a_1\, x^3 + b_1\, x^2 + c_1\, x + d_1$, and we set

$$\tilde{S}(x) := a_1\, x^3 + b_1\, x^2 + c_1\, x + d_1 + \sum_{j=1}^{n-1} \gamma_j\, (x - x_j)_+^3, \qquad x \in [a, b],$$

with $\gamma_j = a_{j+1} - a_j$. Since the truncated powers $(x - x_j)_+^3$ belong to $\mathcal{S}_X$, $\tilde{S}$ is also an element of $\mathcal{S}_X$. The Theorem will follow if we show that $\tilde{S}$ coincides with S.
We let $\tilde{S}_k$ denote the restriction of $\tilde{S}$ to $[x_{k-1}, x_k]$, i.e.

$$\tilde{S}(x) = \tilde{S}_k(x) = a_1\, x^3 + b_1\, x^2 + c_1\, x + d_1 + \sum_{j=1}^{k-1} \gamma_j\, (x - x_j)^3, \qquad x \in [x_{k-1}, x_k],$$

which is a cubic polynomial. In particular, $\tilde{S}_1(x) = a_1\, x^3 + b_1\, x^2 + c_1\, x + d_1 = S_1(x)$ on $[x_0, x_1]$. Arguing by induction, we assume that $k \geq 2$ and $\tilde{S}_{k-1} = S_{k-1}$. Since both $\tilde{S}$ and S are continuous at x_{k-1}, we have

$$\tilde{S}_k(x_{k-1}) = \tilde{S}_{k-1}(x_{k-1}) = S_{k-1}(x_{k-1}) = S_k(x_{k-1})\,.$$

It similarly follows that

$$\tilde{S}_k'(x_{k-1}) = S_k'(x_{k-1}), \quad \tilde{S}_k''(x_{k-1}) = S_k''(x_{k-1})$$

because both $\tilde{S}$ and S are twice continuously differentiable at x_{k-1}. Moreover, $\tilde{S}_k'''$ is the same constant as S_k''':

$$\tilde{S}_k'''(x) = 6\, a_1 \, + \sum_{j=1}^{k-1} 6\, \gamma_j = 6 \left(a_1 + \sum_{j=1}^{k-1} (a_{j+1} - a_j) \right) = 6\, a_k = S_k'''(x)\,.$$

Expanding the two polynomials S_k and $\tilde{S}_k$ into their Taylor series about $x = x_{k-1}$, we get

$$\tilde{S}_k(x) = \sum_{j=0}^{3} \frac{\tilde{S}_k^{(j)}(x_{k-1})}{j!} (x - x_{k-1})^j = \sum_{j=0}^{3} \frac{S_k^{(j)}(x_{k-1})}{j!} (x - x_{k-1})^j = S_k(x)\,,$$

which implies $\tilde{S}_k = S_k$, concluding the induction step. Thus, $\tilde{S}_k = S_k$ for $k = 1 : n$, and hence $\tilde{S} = S$. □

Theorem 3.7 *The dimension of the space $\mathcal{S}_X$ of all cubic splines with respect to a knot sequence $X = [x_0, x_1, \ldots, x_n]$ is $n + 3$.*

Proof. We will show that the functions

$$1,\ x,\ x^2,\ x^3,\ (x - x_1)_+^3,\ \ldots,\ (x - x_{n-1})_+^3 \tag{3.5}$$

form a basis for $\mathcal{S}_X$. Since there are $n+3$ of these functions, the statement that $\dim \mathcal{S}_X = n + 3$ follows immediately.

The polynomials 1, x, x^2 and x^3 are all in the space $\mathcal{S}_X$, since each of them is a spline, all of whose pieces come from the same polynomial. As discussed before, each $(x - x_k)_+^3$ also belongs to $\mathcal{S}_X$ and Theorem 3.6 shows that every spline $S \in \mathcal{S}_X$ can be expressed as a linear combination of the functions (3.5). Therefore, it only remains to show that the the functions (3.5) are *linearly independent*. To do this, we form a linear combination of them, set

$$S(x) := a_1 x^3 + b_1 x^2 + c_1 x + d_1 + \sum_{k=1}^{n-1} \gamma_k (x - x_k)_+^3$$

and assume that it is identically zero, i.e. $S(x) = 0$ for all $x \in [a, b]$. Note that the sum term is zero for $x \in [x_0, x_1]$ and so the polynomial $a_1 x^3 + b_1 x^2 + c_1 x + d_1$ coincides with $S(x) \equiv 0$ on $[x_0, x_1]$; hence $a_1 = b_1 = c_1 = d_1 = 0$ and

$$0 = S(x) = \sum_{k=1}^{n-1} \gamma_k (x - x_k)_+^3\,.$$

Setting $\hat{x}_1 := \frac{1}{2}(x_1 + x_2)$ then gives $0 = S(\hat{x}_1) = \frac{\gamma_1}{8}(x_2 - x_1)^3$, and so $\gamma_1 = 0$ because $x_1 \neq x_2$. Proceeding by induction we assume that $\gamma_1 = \cdots = \gamma_k = 0$ for a given $k < n - 1$. Then $0 = S(\hat{x}_{k+1}) = \frac{\gamma_{k+1}}{8}(x_{k+2} - x_{k+1})^3$, where $\hat{x}_{k+1} := (x_{k+1} + x_{k+2})/2$ which implies that $\gamma_{k+1} = 0$. This shows that $\gamma_k = 0$ for $k = 1 : n - 1$, which concludes the proof that the functions (3.5) are linearly independent. □

Examples 3.4 Let $X = [-1, -1/2, 0, 1/2, 1]$. Check whether the following functions belong to the space $\mathcal{S}_X$, and find the truncated power expansions for those which do:

$$f_1(x) = x^2 - x^3, \quad f_2(x) = e^x x^2, \quad f_3(x) = |x|^3, \quad f_4(x) = (x - 1/2)_+^2 .$$

End conditions

Since the dimension of $\mathcal{S}_X$ is $n + 3$, the $n + 1$ interpolation conditions

$$s(x_j) = y_j, \qquad j = 0 : n,$$

do not determine the interpolating spline completely, and we need to impose two additional constraints to take up the two extraneous degrees of freedom. Three common ways to do this are:

1. **Clamped boundary**: if $S(x)$ is interpolating a function $f(x)$ on $[a, b]$ then set $S'(a) = f'(a)$ and $S'(b) = f'(b)$. Obviously this is not possible if there is no derivative information on f, or if the data points do not come from a known function.
2. **Not a knot condition:** make $S'''(x)$ continuous at the knots x_1 and x_{n-1} (this is the default for the `spline` function in Matlab).
3. **Natural (or free) boundary:** set $S''(a) = S''(b) = 0$. These splines posses an important 'minimum norm' property, which we investigate further below.

Definition 3.11 Given knots $a = x_0 < x_1 < \cdots < x_n = b$ and data values $\{y_k\}_{k=0}^n$, the **natural cubic spline** is $S \in \mathcal{S}_X$, $X = [x_0, \ldots, x_n]$, satisfying

(i) $S(x_i) = y_i$ for $i = 0 : n$;

(ii) $S''(a) = S''(b) = 0$.

We need to show that for any set of interpolation conditions that there is a natural cubic spline, and that it is unique. We shall see that these properties follow from the result below, which is the key to the properties of the natural cubic spline (and is similar in flavour to Theorem 3.5).

Theorem 3.8 *Suppose that $S \in \mathcal{S}_X$ is such that $S''(a) = S''(b) = 0$. For any function $g \in C^2[a, b]$ that satisfies $g(x_k) = S(x_k)$ for $k = 0 : n$, we have*

$$\int_a^b [g''(x)]^2 \, dx - \int_a^b [S''(x)]^2 \, dx = \int_a^b [g''(x) - S''(x)]^2 \, dx \, . \tag{3.6}$$

Proof. Note that

$$\int_a^b [g''(x)]^2\,dx - \int_a^b [S''(x)]^2\,dx = \int_a^b [g''(x)-S''(x)]^2\,dx + 2\int_a^b S''(x)\,[g''(x)-S''(x)]\,dx \tag{3.7}$$

and integrating the final term by parts gives

$$\int_a^b S''(x)\,[g''(x)-S''(x)]\,dx = \Big[S''(x)\,[g'(x)-S'(x)]\Big]_a^b - \int_a^b S'''(x)\,[g'(x)-S'(x)]\,dx\,. \tag{3.8}$$

S''' is a piecewise constant, say $S'''(x) = \alpha_k$ for $x \in [x_{k-1}, x_k]$. Hence

$$\begin{aligned}
\int_a^b S'''(x)\,[g'(x)-s'(x)]\,dx &= \sum_{k=1}^n \int_{x_{k-1}}^{x_k} S'''(x)\,[g'(x)-S'(x)]\,dx \\
&= \sum_{k=1}^n \alpha_k \int_{x_{k-1}}^{x_k} [g'(x)-S'(x)]\,dx \\
&= \sum_{k=1}^n \alpha_k \Big[g(x)-S(x)\Big]_{x_{k-1}}^{x_k} \\
&= 0 \qquad \text{since } g(x_k) = S(x_k) \text{ for } k = 0:n\,.
\end{aligned}$$

Also,

$$\Big[S''(x)\,[g'(x)-S'(x)]\Big]_a^b = 0$$

since $S''(a) = S''(b) = 0$.
Thus the right hand side of (3.8) is zero, and equation (3.7) implies (3.6). □

The following result, which is an immediate consequence of Theorem 3.8, shows that the natural cubic spline has the smallest 2-norm of the second derivative among all twice differentiable functions that interpolate the same data.

Theorem 3.9 *Given knots $a = x_0 < x_1 < \cdots < x_n = b$ and data $\{y_k\}_{k=0}^n$, let S be the natural cubic spline which interpolates the data. For all $g \in C^2[a,b]$ which satisfies $g(x_k) = y_k$ for $k = 0:n$ (i.e. g is any interpolating function in $C^2[a,b]$) we have*

$$\int_a^b [g''(x)]^2\,dx \geq \int_a^b [S''(x)]^2\,dx.$$

We now show that Theorem 3.8 guarantees the existence and uniqueness of natural cubic splines.

Theorem 3.10 *The natural cubic spline for any choice of knots* $a = x_0 < x_1 < \cdots < x_n = b$ *and data values* $\{y_i\}_{i=0}^n$*, exists and is unique.*

Proof. Let $\phi_1, \ldots, \phi_{n+3}$ be a basis for $\mathcal{S}_X$, for example the truncated power basis (3.5). Any $S \in \mathcal{S}_X$ is uniquely represented by the vector $\boldsymbol{c} = [c_1, \ldots, c_{n+3}]^T$ of its coefficients in the expansion

$$S = \sum_{j=1}^{n+3} c_j \, \phi_j \, .$$

Since conditions (i) and (ii) in the natural cubic spline definition are both linear conditions, they are equivalent to $n+3$ linear equations in the $n+3$ unknowns $c_1, \ldots, c_{n+3}$. Let Φ be the matrix of this linear system. We thus want to show that the linear equation $\Phi \, \boldsymbol{c} = [y_0, \ldots, y_n, 0, 0]^T$ is uniquely solvable for any values of $y_0, \ldots, y_n$. This is guaranteed if Φ is non-singular. According to a well known result from linear algebra, Φ is non-singular if and only if

$$\Phi \, \boldsymbol{c} = 0 \quad \text{implies} \quad \boldsymbol{c} = \boldsymbol{0} \, .$$

The latter means the following:

$$S(x_k) = 0, \quad k = 0 : n, \quad \text{and} \quad S''(a) = S''(b) = 0 \qquad \text{implies} \qquad S \equiv 0 \, .$$

Thus, it remains to show that this is true. Let us assume that $S(x_k) = 0$ for $k = 0 : n$ and $S''(a) = S''(b) = 0$. Then the function $g(x) \equiv 0$ satisfies the conditions of Theorem 3.8. Therefore

$$-\int_a^b [S''(x)]^2 \, dx = \int_a^b [S''(x)]^2 \, dx \, ,$$

which implies $\int_a^b [S''(x)]^2 \, dx = 0$. Since S'' is continuous, this can only happen if $S''(x) \equiv 0$, i.e. S is a linear function. But the only linear function satisfying $S(a) = S(b) = 0$ is $S \equiv 0$. □

Remark: Theorem 3.9 shows that if we measure 'wiggliness' of a function g by $\int_a^b [g''(x)]^2 \, dx$, then the natural cubic spline is the least wiggly of all interpolants in $C^2[a, b]$. This may explain its superior performance which is shown Figure 8. The dash-dot line illustrates the degree 10 polynomial interpolant to the 11 data points (marked with asterisks) $x = [0 : 10]$ and $y = [2, 1, 4, 6, 10, 6, 6, 3, 3, 2, 3]$. The dashed and solid lines show respectively the linear and cubic spline interpolants.

Construction of the natural cubic spline interpolant

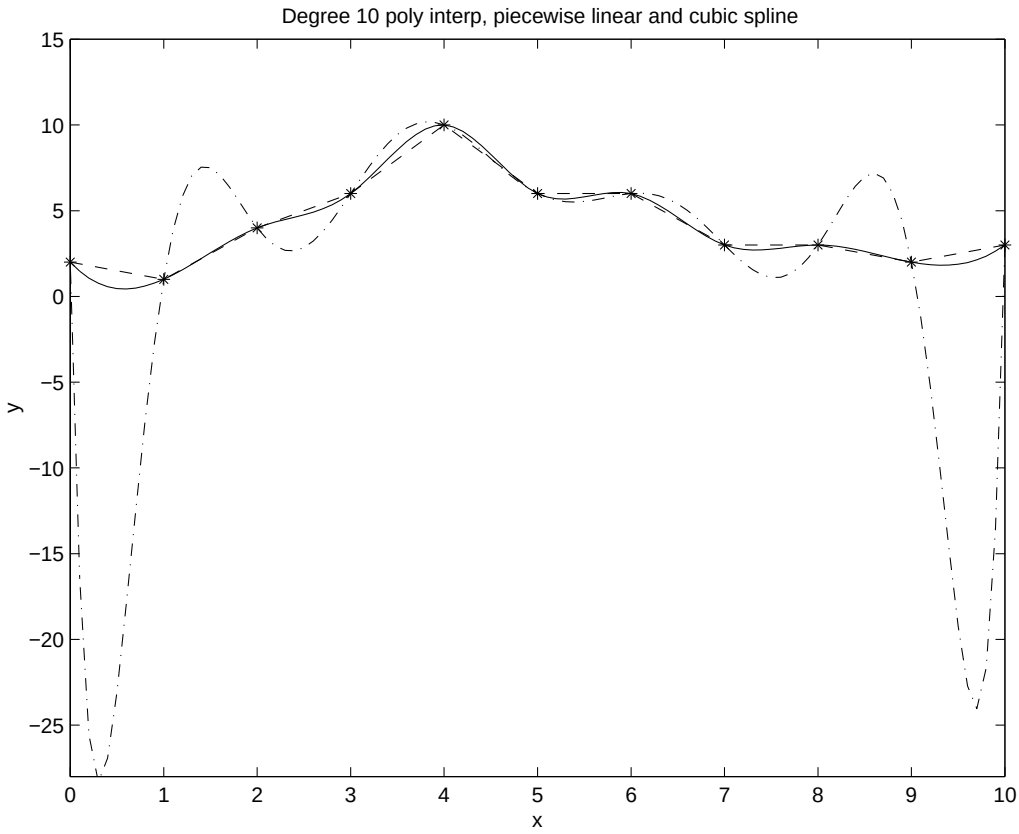

Figure 8: Degree 10 polynomial interpolant (dash-dot), piecewise linear interpolant (dash) and cubic spline (solid). See text for details of the data.

We now show that the natural cubic spline interpolant can be computed efficiently even for very large n.

Suppose the knots are $a = x_0 < x_1 < \cdots < x_n = b$ and the data is $\{y_k\}_{k=0}^n$. Let $S(x)$ be the natural cubic spline which interpolates the data on $[a, b]$. It is a piecewise cubic function and we let S_k denote the cubic on the kth subinterval, i.e. $S_k(x) = S(x)$ for $x \in [x_{k-1}, x_k]$. We also set $h_k = x_k - x_{k-1}$ (it is the length of the interval that S_k is defined on) and let $\{z_k\}_{k=0}^n$ denote the second derivative values $\{S''(x_k)\}_{k=0}^n$ (note that this definition makes sense because S'' is a continuous function). They will be free parameters until we determine them uniquely by solving a linear system of equations.

The function S'' is continuous and piecewise linear, so for $x \in [x_{k-1}, x_k]$ we have

$$S_k''(x) = \left(\frac{x_k - x}{h_k}\right) z_{k-1} + \left(\frac{x - x_{k-1}}{h_k}\right) z_k$$

and integrating this twice gives

$$S_k(x) = c_k\,(x_k - x) + d_k\,(x - x_{k-1}) + \frac{z_{k-1}}{6\,h_k}\,(x_k - x)^3 + \frac{z_k}{6\,h_k}\,(x - x_{k-1})^3 \quad (3.9)$$

for some (arbitrary) constants c_k and d_k. To satisfy the interpolation conditions

$$S(x_k) = y_k, \qquad k = 0 : n$$

we impose $S_k(x_{k-1}) = y_{k-1}$ and $S_k(x_k) = y_k$ on (3.9), and this gives

$$c_k = \frac{y_{k-1}}{h_k} - \frac{z_{k-1}\,h_k}{6}\,, \quad d_k = \frac{y_k}{h_k} - \frac{z_k\,h_k}{6}\,.$$

Hence, we may write (3.9) as

$$\begin{aligned} S_k(x) &= \left(\frac{y_{k-1}}{h_k} - \frac{z_{k-1}\,h_k}{6}\right)(x_k - x) + \left(\frac{y_k}{h_k} - \frac{z_k\,h_k}{6}\right)(x - x_{k-1}) \\ &\quad + \frac{z_{k-1}}{6\,h_k}\,(x_k - x)^3 + \frac{z_k}{6\,h_k}\,(x - x_{k-1})^3\,. \end{aligned} \tag{3.10}$$

Note that the piecewise cubic whose chunks are the S_k solves the interpolation problem for *any* choice of the $\{z_k\}_{k=0}^n$, and by construction it is continuous since the values at x_k of both pieces meeting there are the same as they are equal to y_k. The same is true for the *second derivatives* since the common value of S_k'' and S_{k+1}'' at x_k is z_k. The condition that the spline is natural, i.e.

$$S''(a) = S''(b) = 0$$

simply means that

$$z_0 = z_n = 0\,.$$

The only property we still need to ensure is the *continuity of the first derivatives* at each interior knot x_k, $k = 1 : n-1$. We will see that this is achieved for one and only one choice of the (still free) parameter values $z_1, \ldots, z_{n-1}$.

Differentiating (3.10) gives

$$S_k'(x) = \frac{y_k - y_{k-1}}{h_k} - \frac{h_k}{6}\,(z_k - z_{k-1}) - \frac{z_{k-1}}{2\,h_k}\,(x_k - x)^2 + \frac{z_k}{2\,h_k}\,(x - x_{k-1})^2 \tag{3.11}$$

and the continuity of S' is imposed as $S_k'(x_k) = S_{k+1}'(x_k)$ for $k = 1 : n-1$. This gives

$$h_k\,z_{k-1} + 2\,(h_k + h_{k+1})\,z_k + h_{k+1}\,z_{k+1} = b_k \quad \text{for } k = 1 : n-1,$$

where

$$b_k := 6\left(\frac{y_{k+1} - y_k}{h_{k+1}} - \frac{y_k - y_{k-1}}{h_k}\right).$$

This represents $n-1$ linear equations in the $n-1$ unknowns $\{z_k\}_{k=1}^{n-1}$ and can be written as the matrix–vector system

$$\begin{pmatrix} \gamma_1 & h_2 & 0 & \cdots & \cdots & \cdots & 0 \\ h_2 & \gamma_2 & h_3 & 0 & \cdots & \cdots & \vdots \\ 0 & h_3 & \gamma_3 & h_4 & 0 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & h_{n-1} \\ 0 & \cdots & \cdots & \cdots & 0 & h_{n-1} & \gamma_{n-1} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ \vdots \\ \vdots \\ z_{n-1} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ \vdots \\ \vdots \\ b_{n-1} \end{pmatrix},$$

where $\gamma_k = 2(h_k + h_{k+1})$. Since $\gamma_i > h_i + h_{i+1}$, the coefficient matrix is *diagonally dominant* and hence nonsingular. Thus, there exists a unique solution to the problem. (In fact, we already know this by Theorem 3.10.) As soon as the solution $\{z_k\}_{k=1}^{n-1}$ is computed, the polynomial pieces of the spline are given explicitly by (3.10) taking into account that $z_0 = z_n = 0$.

Finally, we note that the matrix is **tridiagonal** and there are efficient methods for solving such systems. In fact Gaussian elimination without pivoting is quite fast and effective. The number of floating point operations needed in the computation grows *linearly* with the number n of data points.

Examples 3.5 Compute the natural cubic spline S which solves the interpolation problem $S(0) = 1,\ S(1) = 2,\ S(2) = 0,\ S(3) = 1$.

3.4 Linear interpolation in 2D

In this section we look at the problem of interpolating functions (or data) defined on regions of $\mathbb{R}^2$. We shall restrict attention to linear interpolation on one triangle and piecewise linear interpolation on a region formed from two or more triangles. The key to the process is to find the linear basis functions for the **reference triangle** $\mathcal{T}$ shown in Figure 9. We shall map between the reference triangle and an arbitrary triangle in $\mathbb{R}^2$, so we need to use a different coordinate system for each.

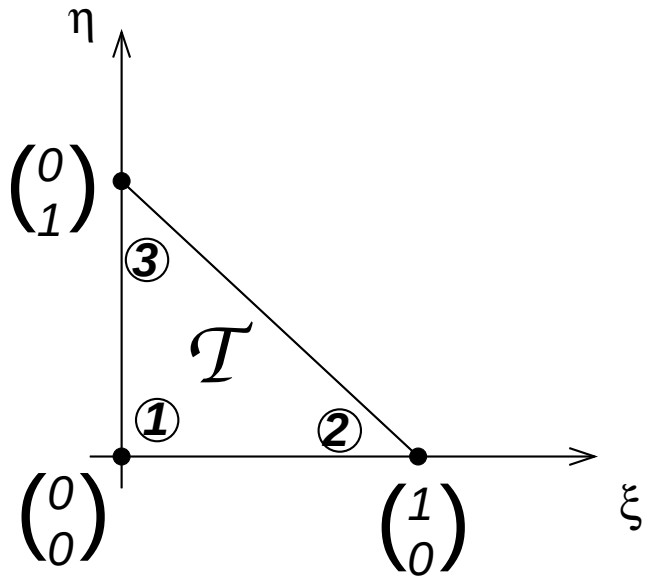

Figure 9: The reference triangle $\mathcal{T}$ in the (ξ, η)-plane.

The reference triangle $\mathcal{T}$ has nodes $\boldsymbol{\nu}_i = (\xi_i, \eta_i)^T$ for $i = 1:3$ where

$$\boldsymbol{\nu}_1 = (0,0)^T, \quad \boldsymbol{\nu}_2 = (1,0)^T, \quad \boldsymbol{\nu}_3 = (0,1)^T.$$

We want to find three basis functions $\Phi_k(\xi,\eta)$, $k = 1:3$, one associated with each node, which are **affine functions** of ξ and η and satisfy

$$\Phi_k(\xi_j,\eta_j) = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k, \end{cases}$$

i.e. $\Phi_k(\xi_j,\eta_j) = \delta_{j,k}$. It is straightforward to verify that the following three functions satisfy these conditions:

$$\Phi_1(\xi,\eta) = 1-\xi-\eta\,, \quad \Phi_2(\xi,\eta) = \xi\,, \quad \Phi_3(\xi,\eta) = \eta\,. \tag{3.12}$$

Just as in 1D, linear interpolation involves the product of the function value at a node with the corresponding basis function.

Definition 3.12 If $U(\xi,\eta)$ is a continuous function defined on the reference triangle $\mathcal{T}$, then the **linear interpolant** of U on $\mathcal{T}$ is

$$U_L(\xi,\eta) := \sum_{k=1}^{3} U(\xi_k,\eta_k)\,\Phi_k(\xi,\eta)\,.$$

Examples 3.6 Show that if U is an affine function on $\mathcal{T}$ then $U = U_L$.

We need to be able to calculate the linear interpolant on an arbitrary triangle T, and to do this we first map $\mathcal{T}$ to T.

Mapping the reference triangle to $T \in \mathbb{R}^2$

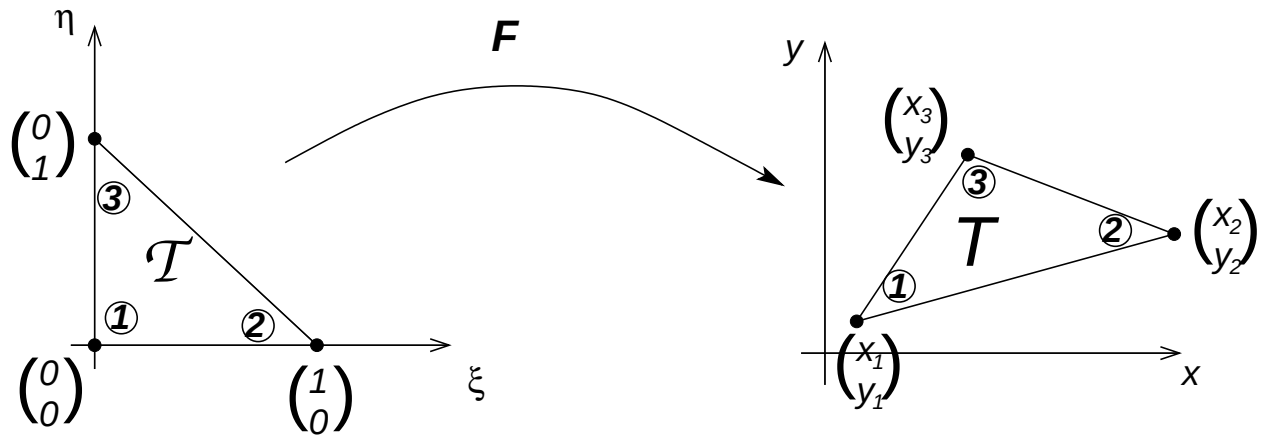

Figure 10: The reference triangle $\mathcal{T}$ is mapped to T in the (x,y)-plane.

Suppose that the situation is as shown in Figure 10: we wish to find an **affine** map $\boldsymbol{F}$ which takes the reference triangle $\mathcal{T}$ in the (ξ,η)-plane to the given

triangle T in the (x, y)-plane, with each node of $\mathcal{T}$ being mapped to the node of T with the same number. That is, $\boldsymbol{F} : \mathcal{T} \to T$ must satisfy

$$\boldsymbol{F}(\xi_k, \eta_k) = \boldsymbol{r}_k$$

for $k = 1 : 3$, where $\boldsymbol{r}_k = (x_k, y_k)^T$ are the nodes of T. It is straightforward to verify that the mapping $\boldsymbol{F}(\boldsymbol{\nu}) = \boldsymbol{F}(\xi, \eta)$ given by

$$\boldsymbol{F}(\xi, \eta) = (1 - \xi - \eta)\, \boldsymbol{r}_1 + \xi\, \boldsymbol{r}_2 + \eta\, \boldsymbol{r}_3 = \sum_{k=1}^{3} \Phi_k(\xi, \eta)\, \boldsymbol{r}_k \tag{3.13}$$

does what is required. Note that

$$\boldsymbol{F}(\boldsymbol{\nu}) = \widehat{J}\, \boldsymbol{\nu} + \boldsymbol{r}_1 \tag{3.14}$$

where

$$\widehat{J} = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}, \tag{3.15}$$

(it is the transpose of the **Jacobian matrix** J which will be seen in Part II).

Construction of linear basis functions on T

We constructed the linear interpolant on $\mathcal{T}$ in terms of the 'nodal' linear basis functions Φ_k, and we want to find linear basis functions ϕ_k on T which have the same properties, i.e. $\phi_k(x_j, y_j) = \delta_{j,k}$. We can use the mapping $\boldsymbol{F}$ between the two triangles in order to do this. Set

$$\phi_k(x, y) = \phi_k(\boldsymbol{r}) = \Phi_k(\boldsymbol{F}^{-1}(\boldsymbol{r})) \quad \text{for } k = 1 : 3$$

where $\boldsymbol{F}^{-1}$ is the **inverse** of the map $\boldsymbol{F}$, i.e. $\boldsymbol{F}^{-1} : T \to \mathcal{T}$ is defined by

$$\boldsymbol{F}^{-1}(\boldsymbol{r}) = \boldsymbol{\nu} \quad \Leftrightarrow \quad \boldsymbol{F}(\boldsymbol{\nu}) = \boldsymbol{r}\,.$$

It follows from (3.14) that $\boldsymbol{F}^{-1}(\boldsymbol{r}) = \widehat{J}^{-1}\,(\boldsymbol{r} - \boldsymbol{r}_1)$ and so

$$\phi_k(x, y) = \Phi_k(\widehat{J}^{-1}\,(\boldsymbol{r} - \boldsymbol{r}_1)) \quad \text{for } k = 1 : 3. \tag{3.16}$$

Examples 3.7 Find $\boldsymbol{F}$, $\boldsymbol{F}^{-1}$ and the functions ϕ_k when:

1. $\boldsymbol{r}_1 = (0,0)^T$, $\boldsymbol{r}_2 = (h,0)^T$, $\boldsymbol{r}_3 = (0,h)^T$;
2. $\boldsymbol{r}_1 = (1,1)^T$, $\boldsymbol{r}_2 = (0,1)^T$, $\boldsymbol{r}_3 = (1,0)^T$.

Note that it is not necessary to ever actually evaluate the ϕ_k, and it is usual to evaluate the linear interpolant in terms of the reference functions Φ_k.

Definition 3.13 If $u(x,y)$ is a continuous function defined on the triangle T, then the linear interpolant of u on T is

$$u_L(x,y) := \sum_{k=1}^{3} u(x_k,y_k)\,\Phi_k(\widehat{J}^{-1}(\boldsymbol{r}-\boldsymbol{r}_1))\,. \tag{3.17}$$

Piecewise linear interpolation on two or more triangles

If a region is formed from two or more triangles, then the **piecewise linear interpolant** of a continuous function u on the region is determined by mapping each triangle onto the reference triangle and applying (3.17).

Examples 3.8 Find the piecewise linear interpolant of the function $u(x,y) = (x+y)^2$ on the square of side 1 whose bottom left corner is at the origin, using the triangles whose common edge is $y = 1-x$.

Remark: The piecewise linear interpolant is continuous by construction: if two triangles have a common edge, then the value of the interpolant on this edge (on either triangle) is determined solely by the nodal values, and these are the same for each triangle because both triangles interpolate the given function at their nodes. In the example above, the triangles intersect along the line $y = 1-x$, and so $x+y = 1$ at any point along the common edge, giving $u_L = 1$ on this edge for both T_1 and T_2.

3.5 Exercises for Section 3

3–1. Find the interpolating polynomial $p \in \mathbb{P}_3$ for the data $\{(0,1),\,(1,3),\,(3,2),(4,4)\}$ and evaluate $p(2)$.

3–2. Find the polynomial $p \in \mathbb{P}_2$ which interpolates the data $\{(1.0,1.0),\,(3.0,3.5),(7.0,8.0)\}$ and evaluate $p(5.0)$ to 2 significant figures.

3–3. Let $f(x) = x^2$ and calculate $S(x)$, the linear spline which interpolates f at the knots $x_0 = 0$, $x_1 = 2$ and $x_2 = 4$, and evaluate $S(1)$ and $S(3)$. Use Theorem 3.4 to bound the maximum error $|f(x) - S(x)|$ for $x \in [0, 4]$.

3–4. Repeat the previous question for the function $f(x) = x^3$.

3–5. Let $X = [-1, -1/2, 0, 1/2, 1]$. Check whether the following functions belong to the space $\mathcal{S}_X$, and find the truncated power expansions for those which do:

$f_1(x) = -1 + x^2 + 3x^4$, $f_2(x) = \sqrt{x^2 + x + 1/4}$, $f_3(x) = 5x^3 - 3x^2 + 1$, $f_4(x) = |x|$, $f_5(x) = |x|^3 - |x + 1/2|^3$, $f_6(x) = \cos^2 x$, and

$$f_7(x) = \left\{ \begin{array}{cc} (x+1/2)^3 & \text{if } -1 \le x < -1/2 \\ 0 & \text{if } -1/2 \le x < 1/2 \\ -3(x-1/2)^3 & \text{if } 1/2 \le x \le 1. \end{array} \right\}$$

3–6. A **quadratic spline** with respect to a given knot sequence $X = [x_0, x_1, \ldots, x_n]$ is a function $S \in C^1[a, b]$ which coincides with a quadratic polynomial on each interval $[x_{k-1}, x_k]$, $k = 1 : n$. Show that any quadratic spline has a truncated power expansion $S(x) = p_2(x) + \sum_{i=1}^{n-1} \gamma_i \, (x - x_i)_+^2$, where p_2 is the quadratic polynomial that coincides with S on $[x_0, x_1]$, and $\gamma_k = a_{k+1} - a_k$, $k = 1 : n - 1$, with a_k being the leading coefficient of the polynomial that coincides with S on $[x_{k-1}, x_k]$.

3–7. Deduce from the previous question that the dimension of the space of quadratic splines with respect to a fixed knot sequence is $n + 2$.

3–8. Compute the natural cubic spline S that solves the interpolation problem $S(0) = 0$, $S(1) = -2$, $S(2) = 1$.

3–9. Find the natural cubic spline $S(x)$ which interpolates the function $f(x) = x^4$ at the knots $[-1, 0, 1]$. (Hint: you can reduce the algebra by noting that since both f and the knots are symmetric about $x = 0$, then so is S.)

3–10. Find the maps $\boldsymbol{F}$ and $\boldsymbol{F}^{-1}$ and the basis functions ϕ_k for the triangle T_1 with nodes $\boldsymbol{r}_1 = (0, 1)^T$, $\boldsymbol{r}_2 = (h, 1)^T$, $\boldsymbol{r}_3 = (0, 1 + h)^T$.

3–11. Find the maps $\boldsymbol{F}$ and $\boldsymbol{F}^{-1}$ and the basis functions ϕ_k for the triangle T_2 with nodes $\boldsymbol{r}_1 = (0, 1)^T$, $\boldsymbol{r}_2 = (-h, 1)^T$, $\boldsymbol{r}_3 = (0, 1 + h)^T$.

3–12. Find the piecewise linear interpolant of the function $u(x, y) = x^2 + y^2$ on the region which consists of triangles T_1 and T_2 (defined above) stuck together along the line $x = 0$.

4 Numerical integration (quadrature)

Part II of MM406 is concerned with using the finite element method to find the numerical solution of differential equations, and this involves approximating integrals in 1D and 2D (i.e. single and double integrals). Here we outline ways in which this is done in practice.

4.1 Gaussian quadrature in 1D

There are a number of ways in which integrals can be evaluated numerically, but we shall focus on **Gaussian quadrature** formulae. They are based on the standard interval $[-1, 1]$ and the m-point Gauss rule takes the form:

$$\int_{-1}^{1} g(\xi)\, d\xi \approx \sum_{i=1}^{m} \omega_i\, g(\xi_i)\,, \tag{4.1}$$

where ξ_i, $i = 1 : m$ are called the **abscissae** or **nodes** and $\omega_i > 0$ are called the **weights**. The first few formulae are:

m	abscissa(e)	weight(s)
1	0	2
2	$\pm\dfrac{1}{\sqrt{3}}$	1
3	0	$8/9$
	$\pm\sqrt{3/5}$	$5/9$

So for example, if $m = 3$ then

$$\int_{-1}^{1} g(\xi)\, d\xi \approx \frac{5}{9}\, g(-\sqrt{3/5}) + \frac{8}{9}\, g(0) + \frac{5}{9}\, g(\sqrt{3/5})\,.$$

The abscissae x_i for the m-point formula are the zeros of the **Legendre polynomial** $P_m(x)$, and the formula (4.1) is **exact** for all polynomials of degree up to $2m - 1$, i.e. if $r \le 2m - 1$ is a non-negative integer, then

$$\int_{-1}^{1} \xi^r\, d\xi = \sum_{i=1}^{m} \omega_i\, \xi_i^r\,.$$

So 1-point Gaussian quadrature integrates exactly all linear functions, 2-point Gaussian quadrature formula integrates all cubic polynomials etc.

Examples 4.1 Show that (4.1) with $m = 2$ integrates all polynomials of degree up to three exactly, but is not exact for all polynomials in $\mathbb{P}_n$ for $n > 3$.

In practice we need to evaluate integrals on a general interval $[a, b]$, and to do this we change the variables of integration in order to use (4.1).
The change of variables

$$x = \frac{(b-a)}{2}\,\xi + \frac{(b+a)}{2} \tag{4.2}$$

maps $x \in [a, b]$ to $\xi \in [-1, 1]$, so

$$\int_a^b f(x)\,dx = \frac{(b-a)}{2}\int_{-1}^{1} f\left(\frac{(b-a)}{2}\,\xi + \frac{(b+a)}{2}\right)\,d\xi$$

and the right-hand integral can be approximated by applying the formula (4.1) to the function

$$g(\xi) = \frac{(b-a)}{2}\,f\left(\frac{(b-a)}{2}\,\xi + \frac{(b+a)}{2}\right)\,.$$

Examples 4.2 Use 2-point Gaussian quadrature and the change of variables (4.2) to approximate the integral $\displaystyle\int_0^1 e^x\,dx\,.$

4.2 2D quadrature

The idea here is similar: we first give a quadrature rule on a reference element (here we use the reference triangle $\mathcal{T}$ from Section 3.4), and then map into the reference element in order to approximate an integral on an arbitrary triangle.
In order to approximate an integral over the reference triangle $\mathcal{T}$ (shown in Figure 9) we use the **edge midpoints** as the abscissae and the corresponding weights are all equal to 1/6, i.e. the rule is

$$\iint_{\mathcal{T}} g(\xi,\eta)\,d\xi\,d\eta \approx \left(g(\tfrac{1}{2}\,,\,\tfrac{1}{2}) + g(0,\tfrac{1}{2}) + g(\tfrac{1}{2},0)\right)/6\,. \tag{4.3}$$

Examples 4.3 Show that the 2D quadrature rule (4.3) is exact for all polynomials in ξ and η of degree up to 2.

We need to evaluate integrals over an arbitrary triangle T with nodes $\boldsymbol{r}_k = (x_k, y_k)^T$ for $k = 1:3$, and this can be done by using the mapping $\boldsymbol{F}$ defined in (3.13). If $(x, y)^T = \boldsymbol{F}(\xi, \eta)$ then

$$x = (1-\xi-\eta)\,x_1 + \xi\,x_2 + \eta\,x_3\,, \quad y = (1-\xi-\eta)\,y_1 + \xi\,y_2 + \eta\,y_3$$

and

$$\iint_T f(x,y)\,dx\,dy = \iint_{\mathcal{T}} f(\boldsymbol{F}(\xi,\eta)) \left|\frac{\partial(x,y)}{\partial(\xi,\eta)}\right| d\xi d\eta$$

and we can apply formula (4.3) to the right-hand integral in order to approximate the integral of f over T. Note that

$$\frac{\partial(x,y)}{\partial(\xi,\eta)} = \begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{pmatrix} = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} = \widehat{J}$$

defined in (3.15) and

$$\left|\frac{\partial(x,y)}{\partial(\xi,\eta)}\right| = \det(\widehat{J}) = 2\,A_{123}$$

where A_{123} is the area of the triangle with nodes $\boldsymbol{r_1}$, $\boldsymbol{r_2}$ and $\boldsymbol{r_3}$, as shown in Figure 10. Formula (4.3) then gives

$$\iint_T f(x,y)\,dx\,dy \approx \frac{A_{123}}{3}\left[f\left(\frac{\boldsymbol{r}_1+\boldsymbol{r}_2}{2}\right) + f\left(\frac{\boldsymbol{r}_2+\boldsymbol{r}_3}{2}\right) + f\left(\frac{\boldsymbol{r}_1+\boldsymbol{r}_3}{2}\right)\right]. \tag{4.4}$$

An integral over a region which contains more than one triangle can be evaluated using this formula as the sum of the integral over each individual triangle.

Examples 4.4 Use (4.4) to evaluate $\iint_T f(\boldsymbol{r})\,dx\,dy$ when T is the triangle with nodes $(0,0)^T$, $(h,0)^T$ and $(0,h)^T$, and $f(\boldsymbol{r}) = x^2 + y^2$.

4.3 Exercises for Section 4

4–1. Show that formula (4.1) is *exact* for $g(\xi) = \xi^r$ when r is any odd positive integer if the abscissae are symmetrically arranged about $\xi = 0$ and the weights corresponding to $\pm\xi_k$ are equal.

4–2. Use m-point Gaussian quadrature for $m = 1,\ 2,\ 3$ to approximate the integral of $g(\xi) = \xi^6$ over the interval $[-1, 1]$, and use the value of the exact integral to work out the error of each approximation.

4–3. Use m-point Gaussian quadrature for $m = 1,\ 2,\ 3$ to approximate the integral of $f(x) = \sin x$ over the interval $[0, \pi/2]$, and use the value of the exact integral to work out the error of each approximation.

4–4. Use formula (4.4) to approximate $\iint_T f(x,y)\,dx\,dy$ when T is the triangle with nodes $(0,0)^T$, $(h,0)^T$ and $(0,h)^T$, and $f(x,y) = x^3 + y^3$.

4–5. Use formula (4.4) (twice) to evaluate $\iint_R f(x,y)\,dx\,dy$ when R is the square of side h whose bottom left-hand corner is at the origin, and $f(x,y) = x^2 + y^2$.

4–6. Use formula (4.4) to approximate $\iint_T f(x,y)\,dx\,dy$ when T is the triangle with nodes $(0,0)^T$, $(2,0)^T$ and $(1,2)^T$, and $f(x,y) = x + y^3$.

4–7. Use formula (4.4) (twice) to approximate $\iint_R f(x,y)\,dx\,dy$ when R is the square which is made from the two triangles with nodes at $(0,0)^T$, $(h,0)^T$, $(0,h)^T$ and $(h,h)^T$, $(0,h)^T$, $(h,0)^T$ with $h = 0.1$, and $f(x,y) = e^{x+2y}$.