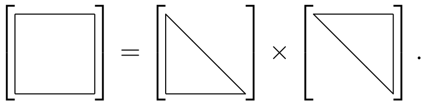
Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 1[1]

1 LU factorization of matrices

1.1 Definition and applications

Let A be a real $n \times n$ matrix. We say that the $n \times n$ matrices L and U are an *LU factorization* of A if ***(1)*** L is lower triangular (i.e., $L_{i,j} = 0$, $i < j$); ***(2)*** U is upper triangular, $U_{i,j} = 0$, $i > j$; and ***(3)*** $A = LU$. Therefore the factorization takes the form



Application 1 Calculation of a determinant: $\det A = (\det L)(\det U) = (\prod_{k=1}^{n} L_{k,k}) \cdot (\prod_{k=1}^{n} U_{k,k})$.

Application 2 Testing for nonsingularity: $A = LU$ is nonsingular iff all the diagonal elements of L and U are nonzero.

Application 3 Solution of linear systems: Let $A = LU$ and suppose we wish to solve $A\boldsymbol{x} = \boldsymbol{b}$. This is the same as $L(U\boldsymbol{x}) = \boldsymbol{b}$, which we decompose into $L\boldsymbol{y} = \boldsymbol{b}$, $U\boldsymbol{x} = \boldsymbol{y}$. Both latter systems are triangular and can be calculated easily. Thus, $L_{1,1}y_1 = b_1$ gives y_1, next $L_{2,1}y_1 + L_{2,2}y_2 = b_2$ yields y_2 etc. Having found $\boldsymbol{y}$, we solve for $\boldsymbol{x}$ by reversing the order: $U_{n,n}x_n = y_n$ gives x_n, $U_{n-1,n-1}x_{n-1} + U_{n-1,n}x_n = y_{n-1}$ produces x_{n-1} and so on. This requires $\mathcal{O}(n^2)$ computational operations (usually we only bother to count multiplications/divisions).

Application 4 The inverse of A: It is straightforward to devise a direct way of calculating the inverse of triangular matrices, subsequently forming $A^{-1} = U^{-1}L^{-1}$.

Why not Cramer's rule? For the uninitiated, a recursive definition of a determinant may seem to be a good method for its calculation (and perhaps even for the solution of linear systems with Cramer's rule). Unfortunately, the number of operations increases like $n!$. Thus, on a 10^9 flop/sec. computer

$$n = 10 \quad \Rightarrow \quad 10^{-4} \text{ sec.}, \qquad n = 20 \quad \Rightarrow \quad 17 \text{ min}, \qquad n = 30 \quad \Rightarrow \quad 4 \times 10^5 \text{ years.}$$

1.2 The calculation of LU factorization

We denote the *columns* of L by $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_n$ and the *rows* of U by $\boldsymbol{u}_1^\top, \boldsymbol{u}_2^\top, \ldots, \boldsymbol{u}_n^\top$. Hence

$$A = LU = [\ \boldsymbol{l}_1 \quad \boldsymbol{l}_2 \quad \cdots \quad \boldsymbol{l}_n\ ] \begin{bmatrix} \boldsymbol{u}_1^\top \\ \boldsymbol{u}_2^\top \\ \vdots \\ \boldsymbol{u}_n^\top \end{bmatrix} = \sum_{k=1}^{n} \boldsymbol{l}_k \boldsymbol{u}_k^\top. \tag{1.1}$$

Since the first $k-1$ components of $\boldsymbol{l}_k$ and $\boldsymbol{u}_k$ are all zero, each rank-one matrix $\boldsymbol{l}_k\boldsymbol{u}_k^\top$ has zeros in its first $k-1$ rows and columns.

Assume that the factorization exists (hence the diagonal elements of L are nonzero) and that A is nonsingular. Since $\boldsymbol{l}_k\boldsymbol{u}_k^\top$ stays the same if we replace $\boldsymbol{l}_k \to \alpha\boldsymbol{l}_k$, $\boldsymbol{u}_k \to \alpha^{-1}\boldsymbol{u}_k$, where $\alpha \neq 0$, we may assume w.l.o.g. that all diagonal elements of L equal one. In other words, the kth row of $\boldsymbol{l}_k\boldsymbol{u}_k^\top$ is $\boldsymbol{u}_k^\top$ and its kth column is $U_{k,k}$ times $\boldsymbol{l}_k$.

[1] Please email all corrections and suggestions to these notes to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/`.

We begin our calculation by extracting $\boldsymbol{l}_1$ and $\boldsymbol{u}_1^\top$ from A, and then proceed similarly to extract $\boldsymbol{l}_2$ and $\boldsymbol{u}_2^\top$, etc.

First we note that since the leading $k-1$ elements of $\boldsymbol{l}_k$ and $\boldsymbol{u}_k$ are zero for $k \geq 2$, it follows from (1.1) that $\boldsymbol{u}_1^\top$ is the first row of A and $\boldsymbol{l}_1$ is the first column of A, divided by $A_{1,1}$ (so that $L_{1,1} = 1$).

Next, having found $\boldsymbol{l}_1$ and $\boldsymbol{u}_1$, we form the matrix $A - \boldsymbol{l}_1\boldsymbol{u}_1^\top = \sum_{k=2}^{n} \boldsymbol{l}_k\boldsymbol{u}_k^\top$. The first row & column of A are zero and it follows that $\boldsymbol{u}_2^\top$ is the second row of $A - \boldsymbol{l}_1\boldsymbol{u}_1^\top$, while $\boldsymbol{l}_2$ is its second column, scaled so that $L_{2,2} = 1$.

The LU algorithm: Set $A_0 := A$. For all $k = 1, 2, \ldots, n$ set $\boldsymbol{u}_k^\top$ to the kth row of A_{k-1} and $\boldsymbol{l}_k$ to the kth column of A_{k-1}, scaled so that $L_{k,k} = 1$. Further, calculate $A_k := A_{k-1} - \boldsymbol{l}_k\boldsymbol{u}_k^\top$ before incrementing k.

Note that all elements in the first k rows & columns of A_k are zero. Hence, we can use the storage of the original A to accumulate L and U. The full LU factorization requires $\mathcal{O}(n^3)$ computational operations.

1.3 Relation to Gaussian elimination

The equation $A_k = A_{k-1} - \boldsymbol{l}_k\boldsymbol{u}_k^\top$ has the property that the jth row of A_k is the jth row of A_{k-1} minus $L_{j,k}$ times $\boldsymbol{u}_k^\top$ (the kth row of A_{k-1}). Moreover, the multipliers $L_{k,k}, L_{k+1,k}, \ldots, L_{n,k}$ are chosen so that the outcome of this *elementary row operation* is that the kth column of A_k is zero. This construction is analogous to Gaussian elimination for solving $A\boldsymbol{x} = \boldsymbol{b}$. An important difference is that in LU we do not consider the right hand side $\boldsymbol{b}$ until the factorization is complete. This is useful e.g. when there are many right hand sides, in particular if not all the $\boldsymbol{b}$'s are known at the outset: in Gaussian elimination the solution for each new $\boldsymbol{b}$ would require $\mathcal{O}(n^3)$ computational operations, whereas with LU factorization $\mathcal{O}(n^3)$ operations are required for the initial factorization, but then the solution for each new $\boldsymbol{b}$ only requires $\mathcal{O}(n^2)$ operations.

1.4 Pivoting

Naive LU factorization fails when, for example, $A_{1,1} = 0$. The remedy is to exchange rows of A, a technique called *column pivoting.* This is equivalent to picking a suitable equation for eliminating the first unknown in Gaussian elimination. Specifically, column pivoting means that, having obtained A_{k-1}, we exchange two rows of A_{k-1} so that the element of largest magnitude in the kth column is in the 'pivotal position' (k,k). In other words,

$$|(A_{k-1})_{k,k}| = \max\{|(A_{k-1})_{j,k}| \ : \ j = 1, 2, \ldots, n\}.$$

Of course, the same exchange is required in the portion of L that has been formed already (i.e., the first $k-1$ columns). Also, we need to record the permutation of rows to solve for the right hand side and/or to compute the determinant. (The exchange of rows can be regarded as the pre-multiplication of the relevant matrix by a permutation matrix.)

Column pivoting copes with zeros at the pivot position, except when the whole kth column of A_{k-1} is zero – in that case it is usual to let $\boldsymbol{l}_k$ be the kth unit vector while, as before, choose $\boldsymbol{u}_k^\top$ as the kth row of A_k). Such a choice preserves the condition that the matrix $\boldsymbol{l}_k\boldsymbol{u}_k^\top$ has the same kth row and column as A_{k-1}. Thus $A_k := A_{k-1} - \boldsymbol{l}_k\boldsymbol{u}_k^\top$ still has zeros in its kth row and column as required.

An important advantage of column pivoting is that every element of L has magnitude at most one. This avoids not just division by zero but also tends to reduce the chance of very large numbers occuring during the factorization, a phenomenon that might lead to *ill conditioning* and to accumulation of *roundoff error.*

In *row pivoting* one exchanges columns of A_{k-1}, rather than rows (sic!), whereas *total pivoting* corresponds to exchange of both rows and columns, so that the modulus of the pivotal element $(A_{k-1})_{k,k}$ is maximised.

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 2[1]

2 Factorization of structured matrices

2.1 Symmetric matrices

Let A be an $n \times n$ symmetric matrix (i.e., $A_{k,\ell} = A_{\ell,k}$). An analogue of LU factorization takes advantage of symmetry: we express A in the form of the product $LDL^\top$, where L is $n \times n$ lower triangular, with ones on its diagonal, whereas D is a diagonal matrix. Subject to its existence, we can write this factorization as

$$A = \begin{bmatrix} \boldsymbol{l}_1 & \boldsymbol{l}_2 & \cdots \boldsymbol{l}_n \end{bmatrix} \begin{bmatrix} D_{1,1} & 0 & \cdots & 0 \\ 0 & D_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & D_{n,n} \end{bmatrix} \begin{bmatrix} \boldsymbol{l}_1^\top \\ \boldsymbol{l}_2^\top \\ \vdots \\ \boldsymbol{l}_n^\top \end{bmatrix} = \sum_{k=1}^{n} D_{k,k} \boldsymbol{l}_k \boldsymbol{l}_k^\top$$

where, as before, $\boldsymbol{l}_k$ is the kth column of L.
The analogy with the algorithm of Section 1.2 becomes obvious by letting $U = DL^\top$, but the present form lends itself better to exploitation of symmetry. Specifically, to compute this factorization, we let $A_0 = A$ and for $k = 1, 2, \ldots, n$ let $\boldsymbol{l}_k$ be the multiple of the kth column of A_{k-1} such that $L_{k,k} = 1$. Set $D_{k,k} = (A_{k-1})_{k,k}$ and form $A_k = A_{k-1} - D_{k,k}\boldsymbol{l}_k\boldsymbol{l}_k^\top$.

Example Let $A = A_0 = \begin{bmatrix} 2 & 4 \\ 4 & 11 \end{bmatrix}$. Hence $\boldsymbol{l}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $D_{1,1} = 2$ and

$$A_1 = A_0 - D_{1,1}\boldsymbol{l}_1\boldsymbol{l}_1^\top = \begin{bmatrix} 2 & 4 \\ 4 & 11 \end{bmatrix} - 2 \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 3 \end{bmatrix}.$$

We deduce that $\boldsymbol{l}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $D_{2,2} = 3$ and $A = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$.

2.2 Symmetric positive definite matrices

Recall that A is positive definite if $\boldsymbol{x}^\top A\boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$.

Theorem Let A be a real $n \times n$ symmetric matrix. It is positive definite if and only if it has an $LDL^\top$ factorization in which the diagonal elements of D are all positive.

Proof. Suppose that $A = LDL^\top$ and let $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$. Since L is nonsingular, $\boldsymbol{y} := L^\top \boldsymbol{x} \neq \boldsymbol{0}$. Then $\boldsymbol{x}^\top A\boldsymbol{x} = \boldsymbol{y}^\top D\boldsymbol{y} = \sum_{k=1}^{n} D_{k,k} y_k^2 > 0$, hence A is positive definite.
Conversely, suppose that A is positive definite. We wish to demonstrate that an $LDL^\top$ factorization exists. We denote by $\boldsymbol{e}_k \in \mathbb{R}^n$ the kth unit (a.k.a. coordinate) vector. Hence $\boldsymbol{e}_1^\top A\boldsymbol{e}_1 = A_{1,1} > 0$ and $\boldsymbol{l}_1$ & $D_{1,1}$ are well defined. We now show that $(A_{k-1})_{k,k} > 0$ for $k = 1, 2, \ldots$. The result is true for $k = 1$ and we continue by induction (hence may assume that $A_{k-1} = A - \sum_{j=1}^{k-1} D_{j,j}\boldsymbol{l}_j\boldsymbol{l}_j^\top$ has been computed successfully).
We define $\boldsymbol{x} \in \mathbb{R}^n$ as follows. The bottom $n-k$ components are zero, $x_k = 1$ and $x_1, x_2, \ldots, x_{k-1}$ are calculated in a reverse order, each x_j being chosen so that $\boldsymbol{l}_j^\top \boldsymbol{x} = 0$ for $j = k-1, k-2, \ldots, 1$. In other words, since $0 = \boldsymbol{l}_j^\top \boldsymbol{x} = \sum_{i=1}^{n} L_{i,j}x_i = \sum_{i=j}^{k} L_{i,j}x_i$, we let $x_j = -\sum_{i=j+1}^{k} L_{i,j}x_i$, $j = k-1, k-2, \ldots, 1$.

[1] Corrections and suggestions to these notes should be emailed to A.Iserles@damtp.cam.ac.uk. All handouts are available on the WWW at the URL http://www.damtp.cam.ac.uk/user/na/PartIB/.

Since the first $k-1$ rows & columns of A_{k-1} vanish, our choice implies that $(A_{k-1})_{k,k} = \boldsymbol{x}^\top A_{k-1}\boldsymbol{x}$. Thus, from the definition of A_{k-1} and the choice of $\boldsymbol{x}$,

$$(A_{k-1})_{k,k} = \boldsymbol{x}^\top A_{k-1}\boldsymbol{x} = \boldsymbol{x}^\top \left(A - \sum_{j=1}^{k-1} D_{j,j}\boldsymbol{l}_j\boldsymbol{l}_j^\top \right) \boldsymbol{x} = \boldsymbol{x}^\top A\boldsymbol{x} - \sum_{j=1}^{k-1} D_{j,j}(\boldsymbol{l}_j^\top \boldsymbol{x})^2 = \boldsymbol{x}^\top A\boldsymbol{x} > 0,$$

as required. Hence $(A_{k-1})_{k,k} > 0$, $k = 1, 2, \ldots, n$, and the factorization exists. □

Conclusion It is possible to check if a symmetric matrix is positive definite by trying to form its $LDL^\top$ factorization.

Cholesky factorization Define $D^{1/2}$ as the diagonal matrix whose (k,k) element is $D_{k,k}^{1/2}$, hence $D^{1/2}D^{1/2} = D$. Then, A being positive definite, we can write

$$A = (LD^{1/2})(D^{1/2}L^\top) = (LD^{1/2})(LD^{1/2})^\top.$$

In other words, letting $\tilde{L} := LD^{1/2}$, we obtain the *Cholesky factorization* $A = \tilde{L}\tilde{L}^\top$.

2.3 Sparse matrices

Frequently it is required to solve *very* large systems $A\boldsymbol{x} = \boldsymbol{b}$ ($n = 10^5$ is considered small in this context!) where nearly all the elements of A are zero. Such a matrix is called *sparse* and efficient solution of $A\boldsymbol{x} = \boldsymbol{b}$ should exploit sparsity. In particular, we wish the matrices L and U to inherit as much as possible of the sparsity of A. The only tool at our disposal at the moment is the freedom to exchange rows and columns to minimise *fill-in.* To this end the following theorem is useful.

Theorem Let $A = LU$ be an LU factorization (without pivoting) of a sparse matrix. Then all leading zeros in the rows of A to the left of the diagonal are inherited by L and all the leading zeros in the columns of A above the diagonal are inherited by U.

Proof Follows from Question 2 on Examples' Sheet 1. □

This theorem suggests that if one requires a factorization of a sparse matrix then one might try to reorder its rows and columns by a preliminary calculation so that many of the zero elements are leading zero elements in rows and columns. This will reduce the fill-in.

Example 1 The LU factorisation of

$$\begin{bmatrix} -3 & 1 & 1 & 2 & 0 \\ 1 & -3 & 0 & 0 & 1 \\ 1 & 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 3 & 0 \\ 0 & 1 & 0 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -\frac{1}{3} & 1 & 0 & 0 & 0 \\ -\frac{1}{3} & -\frac{1}{8} & 1 & 0 & 0 \\ -\frac{2}{3} & -\frac{1}{4} & \frac{6}{19} & 1 & 0 \\ 0 & -\frac{3}{8} & \frac{1}{19} & \frac{4}{81} & 1 \end{bmatrix} \begin{bmatrix} -3 & 1 & 1 & 2 & 0 \\ 0 & -\frac{8}{3} & \frac{1}{3} & \frac{2}{3} & 1 \\ 0 & 0 & \frac{19}{8} & \frac{3}{4} & \frac{1}{8} \\ 0 & 0 & 0 & \frac{81}{19} & \frac{4}{19} \\ 0 & 0 & 0 & 0 & \frac{272}{81} \end{bmatrix},$$

has significant fill-in. However, reordering (symmetrically) rows and columns $1 \leftrightarrow 3$, $2 \leftrightarrow 4$ and $4 \leftrightarrow 5$ yields

$$\begin{bmatrix} 2 & 0 & 1 & 0 & 0 \\ 0 & 3 & 2 & 0 & 0 \\ 1 & 2 & -3 & 0 & 1 \\ 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 1 & 1 & -3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{2}{3} & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{6}{29} & \frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 1 & 0 & 0 \\ 0 & 3 & 2 & 0 & 0 \\ 0 & 0 & -\frac{29}{6} & 0 & 1 \\ 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & -\frac{272}{87} \end{bmatrix}.$$

Example 2 If the nonzeros of A occur only on the diagonal, in one row and in one column, then the full row and column should be placed at the bottom and on the right of A, respectively.

General treatment of orderings that minimise sparsity can be addressed using *graph theory,* but this is well outside the scope of an undergraduate course.

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 3[1]

Band matrices The matrix A is a *band matrix* if there exists an integer $r < n$ such that $A_{i,j} = 0$ for $|i-j| > r$, $i,j = 1,2,\ldots,n$. In other words, all the nonzero elements of A reside in a band of width $2r+1$ along the main diagonal. In that case, according to the statement from the end of the last lecture, $A = LU$ implies that $L_{i,j} = U_{i,j} = 0 \; \forall \; |i-j| > r$ and sparsity structure is inherited by the factorization.
In general, the expense of calculating an LU factorization of an $n \times n$ *dense* matrix A is $\mathcal{O}(n^3)$ operations and the expense of solving $A\boldsymbol{x} = \boldsymbol{b}$, provided that the factorization is known, is $\mathcal{O}(n^2)$. However, in the case of a banded A, we need just $\mathcal{O}(r^2 n)$ operations to factorize and $\mathcal{O}(rn)$ operations to solve a linear system. If $r \ll n$ this represents a very substantial saving!

General sparse matrices are crucial to a wide range of applications, e.g. the solution of partial differential equations. There exists a wealth of methods for their solution. One approach is efficient factorization, that minimizes *fill in.* Yet another is to use iterative methods, our next topic. There also exists a substantial body of other, highly effective methods, e.g. Fast Fourier Transforms and multigrid techniques (cf. Part II course in Numerical Analysis), fast multipole techniques and much more.

3 Iterative methods for linear systems

3.1 Basic iterative schemes

Solution of $A\boldsymbol{x} = \boldsymbol{b}$ by factorization is frequently very expensive for large n, even if we exploit sparsity. An alternative is to use *iterative methods.* Such methods are very efficient and have been subjected to intensive attention in the last few decades. An example of an iterative scheme is to write $A = B - C$, where ***(1)*** B & C are $n \times n$ matrices; ***(2)*** B is nonsingular; ***(3)*** the system $B\boldsymbol{x} = \boldsymbol{c}$ is *easy to solve* and ***(4)*** the matrix C is somehow 'small' in comparison with B. We write the original system in the form $B\boldsymbol{x} = C\boldsymbol{x} + \boldsymbol{b}$ and consider solving it by iteration. Choose an arbitrary $\boldsymbol{x}_0 \in \mathbb{R}^n$ and define $\boldsymbol{x}_{m+1}$, $m = 0,1,\ldots$, by solving

$$B\boldsymbol{x}_{m+1} = C\boldsymbol{x}_m + \boldsymbol{b}. \tag{3.1}$$

Provided that B is, for example, banded, the solution of (3.1) is cheap (and the LU factorization of B can be re-used – an example of why the LU formalism is superior to Gaussian elimination). Often the sequence $\{\boldsymbol{x}_m\}_{m=0}^{\infty}$ converges to the solution of $A\boldsymbol{x} = \boldsymbol{b}$.

The Jacobi iteration We write $A = A_{\rm D} - A_{\rm L} - A_{\rm U}$, where $A_{\rm L}$ is strictly lower triangular, $A_{\rm D}$ is diagonal and $A_{\rm U}$ is strictly upper triangular. Suppose that no diagonal element of A is zero. The *Jacobi iteration* is

$$A_{\rm D}\boldsymbol{x}_{m+1} = (A_{\rm L} + A_{\rm U})\boldsymbol{x}_m + \boldsymbol{b}, \qquad m = 0,1,\ldots. \tag{3.2}$$

The Gauss–Seidel iteration In the above notation, it takes the form

$$(A_{\rm D} - A_{\rm L})\boldsymbol{x}_{m+1} = A_{\rm U}\boldsymbol{x}_m + \boldsymbol{b}, \qquad m = 0,1,\ldots. \tag{3.3}$$

Note that $A_{\rm L} + A_{\rm D}$ is lower triangular, hence the solution of (3.3) is cheap.

[1] Corrections and suggestions to these notes should be emailed to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/`.

3.2 Necessary and sufficient conditions for convergence

Suppose that A is nonsingular and denote by $\boldsymbol{x}^*$ the solution of $A\boldsymbol{x}=\boldsymbol{b}$. Having written $A=B-C$, we examine the iterative scheme (3.1). (Note that both (3.2) and (3.3) can be cast in this form.) Our goal is to identify conditions so that $\boldsymbol{x}_m \to \boldsymbol{x}^*$, regardless of the choice of $\boldsymbol{x}_0 \in \mathbb{R}^n$.

Subtract $B\boldsymbol{x}^*=C\boldsymbol{x}^*+\boldsymbol{b}$ from (3.1). This gives $B(\boldsymbol{x}_{m+1}-\boldsymbol{x}^*)=C(\boldsymbol{x}_m-\boldsymbol{x}^*)$, hence $B\boldsymbol{\varepsilon}_{m+1}=C\boldsymbol{\varepsilon}_m$, where $\boldsymbol{\varepsilon}_m := \boldsymbol{x}_m-\boldsymbol{x}^*$ is the error in the mth iterate. Since B is nonsingular (otherwise we cannot execute (3.1) in the first place), it follows that

$$\boldsymbol{\varepsilon}_{m+1}=H\boldsymbol{\varepsilon}_m=\cdots=H^{m+1}\boldsymbol{\varepsilon}_0, \quad m=0,1,\ldots \quad \text{where} \quad H=B^{-1}C \quad \text{is the } \textit{iteration matrix}.. \tag{3.4}$$

This indicates that the errors tend to zero as $m\to\infty$ (regardless of the choice of $\boldsymbol{x}_0$) provided that $\lim_{m\to\infty} H^m = O$.

We employ the notation $\rho(P)$ for the magnitude of the largest (in absolute value) eigenvalue of the $n\times n$ matrix P. The quantity $\rho(P)$ is called the *spectral radius* of the matrix P. (*Note:* Recall that, even if P is real, its eigenvalues might be complex.)

Theorem $\lim_{m\to\infty}\boldsymbol{x}_m=\boldsymbol{x}^*$ for all $\boldsymbol{x}_0\in\mathbb{R}^n$ if and only if $\rho(H)<1$.

Proof. We commence with the case $\rho(H)\ge 1$ and wish to demonstrate that $\boldsymbol{\varepsilon}_m$ need not tend to $\boldsymbol{0}$. Let λ be an eigenvalue of H such that $|\lambda|=\rho(H)$ and let $\boldsymbol{w}$ be a corresponding eigenvector, $H\boldsymbol{w}=\lambda\boldsymbol{w}$. If $\boldsymbol{w}$ is real, we choose $\boldsymbol{x}_0=\boldsymbol{x}^*+\boldsymbol{w}$, hence $\boldsymbol{\varepsilon}_0=\boldsymbol{w}$. It follows at once by induction that $\boldsymbol{\varepsilon}_m=\lambda^m\boldsymbol{w}$, and this cannot tend to zero since $|\lambda|\ge 1$.

If $\lambda\in\mathbb{C}\setminus\mathbb{R}$ then $\boldsymbol{w}$ is complex. Moreover, also $\bar\lambda\ne\lambda$ is an eigenvalue and $\bar{\boldsymbol{w}}$ is its eigenvector (the bar denotes complex conjugation). Note that $\boldsymbol{w}$ and $\bar{\boldsymbol{w}}$ are linearly independent (otherwise they would have corresponded to the same eigenvalue). We denote the *Euclidean length* of $\boldsymbol{p}\in\mathbb{C}^n$ by

$$\|\boldsymbol{p}\|=\left(\sum_{k=1}^n |p_k|^2\right)^{1/2}.$$

Note that $\|\boldsymbol{p}\|$ is a continuous function of the components of $\boldsymbol{p}$. Hence, $\|z\boldsymbol{w}+\bar z\bar{\boldsymbol{w}}\|$ is a continuous function of the complex variable z. It is a consequence of the linear independence of $\boldsymbol{w}$ and $\bar{\boldsymbol{w}}$ and of the theorem that a continuous function attains its minimum in a closed interval that

$$\inf_{-\pi\le\theta\le\pi}\|\mathrm{e}^{\mathrm{i}\theta}\boldsymbol{w}+\mathrm{e}^{-\mathrm{i}\theta}\bar{\boldsymbol{w}}\| = \min_{-\pi\le\theta\le\pi}\|\mathrm{e}^{\mathrm{i}\theta}\boldsymbol{w}+\mathrm{e}^{-\mathrm{i}\theta}\bar{\boldsymbol{w}}\|=\nu,$$

say, is positive. ($\nu=0$ would have implied $\mathrm{e}^{\mathrm{i}\theta^*}\boldsymbol{w}+\mathrm{e}^{-\mathrm{i}\theta^*}\bar{\boldsymbol{w}}=\boldsymbol{0}$ for some θ^*.) By homogeneity, it is true for every $z\in\mathbb{C}$ that

$$\|z\boldsymbol{w}+\bar z\bar{\boldsymbol{w}}\|\ge\nu|z|. \tag{3.5}$$

We let $\boldsymbol{x}_0=\boldsymbol{x}^*+\boldsymbol{w}+\bar{\boldsymbol{w}}$, hence $\boldsymbol{\varepsilon}_0=\boldsymbol{w}+\bar{\boldsymbol{w}}$. (Note that everything in sight is real: this was precisely the purpose of our construction!) We have by induction on (3.1) that

$$\boldsymbol{\varepsilon}_m=\lambda^m\boldsymbol{w}+\bar\lambda^m\bar{\boldsymbol{w}}, \qquad m=0,1,\ldots.$$

Setting $z=\lambda^m$, (3.5) implies that $\|\boldsymbol{\varepsilon}_m\|\ge\nu|\lambda^m|\ge\nu$. Hence the sequence $\{\boldsymbol{\varepsilon}_m\}_{m=0}^\infty$ is bounded away from zero and $\boldsymbol{\varepsilon}_m\not\to\boldsymbol{0}$. This completes the proof of the 'only if' part of the theorem.

Mathematical Tripos Part IB: Easter 2006
Numerical Analysis – Lecture 4[1]

Recap of Theorem $\lim_{m\to\infty} \boldsymbol{x}_m = \boldsymbol{x}^*$ for all $\boldsymbol{x}_0 \in \mathbb{R}^n$ if and only if $\rho(H) < 1$.

. . . proof. We consider next the case of $\rho(H) < 1$. Assume for simplicity that H possesses n linearly independent eigenvectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n$, say. Hence $H\boldsymbol{w}_j = \lambda_j \boldsymbol{w}_j$, $|\lambda_j| < 1$, $j = 1, 2, \ldots, n$. Linear independence means that every $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ can be expressed as a linear combination of the eigenvectors. Therefore, given $\boldsymbol{x}_0 \in \mathbb{R}^n$, there exist $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{C}$ such that $\boldsymbol{\varepsilon}_0 = \boldsymbol{x}_0 - \boldsymbol{x}^* = \sum_{j=1}^n \alpha_j \boldsymbol{w}_j$. Thus,

$$\boldsymbol{\varepsilon}_1 = H\boldsymbol{\varepsilon}_0 = \sum_{j=1}^n \alpha_j \lambda_j \boldsymbol{w}_j \qquad \text{and, by induction,} \qquad \boldsymbol{\varepsilon}_m = \sum_{j=1}^n \alpha_j \lambda_j^m \boldsymbol{w}_j$$

for all $m = 0, 1, \ldots$. Since $\rho(H) < 1$, it follows that $\lim_{m\to\infty} \boldsymbol{\varepsilon}_m = \boldsymbol{0}$, as required. □

The 'missing' case Suppose that $\rho(H) < 1$ but that H does not have n linearly independent eigenvalues. This occurs, for example, for the matrix $H = \begin{bmatrix} a & b \\ 0 & a \end{bmatrix}$, where $b \neq 0$ and $|a| < 1$. The eigenvalues of H are both a, but it is an easy exercise to verify that all eigenvectors are necessarily multiples of $\boldsymbol{e}_1$. Moreover, $H^m = \begin{bmatrix} a^m & ma^{m-1}b \\ 0 & a^m \end{bmatrix}$ (prove!), therefore $|a| < 1$ implies $H^m \to O$.

4 QR factorization of matrices

4.1 Scalar products, norms and orthogonality

We first revise a few definitions. $\mathbb{R}^n$ is the linear space of all real n-tuples.

- For all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ we define the *scalar product*

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \langle \boldsymbol{v}, \boldsymbol{u} \rangle = \sum_{j=1}^n u_j v_j = \boldsymbol{u}^\top \boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{u}.$$

- If $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$ then $\langle \alpha\boldsymbol{u} + \beta\boldsymbol{w}, \boldsymbol{v} \rangle = \alpha\langle \boldsymbol{u}, \boldsymbol{v} \rangle + \beta\langle \boldsymbol{w}, \boldsymbol{v} \rangle$.
- The *norm* (a.k.a. the *Euclidean length*) of $\boldsymbol{u} \in \mathbb{R}^n$ is $\|\boldsymbol{u}\| = \left(\sum_{j=1}^n u_j^2\right)^{1/2} = \langle \boldsymbol{u}, \boldsymbol{u} \rangle^{1/2} \geq 0$.
- For $\boldsymbol{u} \in \mathbb{R}^n$, $\|\boldsymbol{u}\| = 0$ iff $\boldsymbol{u} = \boldsymbol{0}$.
- We say that $\boldsymbol{u} \in \mathbb{R}^n$ and $\boldsymbol{v} \in \mathbb{R}^n$ are *orthogonal* to each other if $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$.
- The vectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m \in \mathbb{R}^n$ are *orthonormal* if

$$\langle \boldsymbol{q}_k, \boldsymbol{q}_\ell \rangle = \begin{cases} 1, & k = \ell, \\ 0, & k \neq \ell, \end{cases} \qquad k, \ell = 1, 2, \ldots, m.$$

- An $n \times n$ real matrix Q is *orthogonal* if all its columns are orthonormal. Since $(Q^\top Q)_{k,\ell} = \langle \boldsymbol{q}_k, \boldsymbol{q}_\ell \rangle$, this implies that $Q^\top Q = I$ (I is the *unit matrix*). Hence $Q^{-1} = Q^\top$ and $QQ^\top = QQ^{-1} = I$. We conclude that the rows of an orthogonal matrix are also orthonormal, and that $Q^\top$ is an orthogonal matrix. Further, $1 = \det I = \det(QQ^\top) = \det Q \det Q^\top = (\det Q)^2$, and thus we deduce that $\det Q = \pm 1$, and that an orthogonal matrix is nonsingular.

[1] Corrections and suggestions to these notes should be emailed to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/`.

Proposition If P, Q are orthogonal then so is PQ.

Proof. Since $P^\top P = Q^\top Q = I$, we have $(PQ)^\top(PQ) = (Q^\top P^\top)(PQ) = Q^\top(P^\top P)Q = Q^\top Q = I$, hence PQ is orthogonal. □

Proposition Let $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m \in \mathbb{R}^n$ be orthonormal. Then $m \le n$.

Proof. We argue by contradiction. Suppose that $m \ge n+1$ and let Q be the orthogonal matrix whose columns are $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$. Since Q is nonsingular and $\boldsymbol{q}_m \ne \boldsymbol{0}$, there exists a nonzero solution to the linear system $Q\boldsymbol{a} = \boldsymbol{q}_m$, hence $\boldsymbol{q}_m = \sum_{j=1}^n a_j \boldsymbol{q}_j$. But

$$0 = \langle \boldsymbol{q}_\ell, \boldsymbol{q}_m \rangle = \left\langle \boldsymbol{q}_\ell, \sum_{j=1}^n a_j \boldsymbol{q}_j \right\rangle = \sum_{j=1}^n a_j \langle \boldsymbol{q}_\ell, \boldsymbol{q}_j \rangle = a_\ell, \qquad \ell = 1, 2, \ldots, n,$$

hence $\boldsymbol{a} = \boldsymbol{0}$, a contradiction. We deduce that $m \le n$. □

Lemma Let $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m \in \mathbb{R}^n$ be orthonormal and $m \le n-1$. Then there exists $\boldsymbol{q}_{m+1} \in \mathbb{R}^n$ such that $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{m+1}$ are orthonormal.

Proof. We construct $\boldsymbol{q}_{m+1}$. Let Q be the $n \times m$ matrix whose columns are $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m$. Since

$$\sum_{k=1}^n \sum_{j=1}^m Q_{k,j}^2 = \sum_{j=1}^m \|\boldsymbol{q}_j\|^2 = m < n,$$

it follows that $\exists\ \ell \in \{1, 2, \ldots, n\}$ such that $\sum_{j=1}^m Q_{\ell,j}^2 < 1$. We let $\boldsymbol{w} = \boldsymbol{e}_\ell - \sum_{j=1}^m \langle \boldsymbol{q}_j, \boldsymbol{e}_\ell \rangle \boldsymbol{q}_j$. Then for $i = 1, 2, \ldots, m$

$$\langle \boldsymbol{q}_i, \boldsymbol{w} \rangle = \langle \boldsymbol{q}_i, \boldsymbol{e}_\ell \rangle - \sum_{j=1}^m \langle \boldsymbol{q}_j, \boldsymbol{e}_\ell \rangle \langle \boldsymbol{q}_i, \boldsymbol{q}_j \rangle = 0,$$

i.e. by design $\boldsymbol{w}$ is orthogonal to $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m$. Further, since $Q_{\ell,j} = \langle \boldsymbol{q}_j, \boldsymbol{e}_\ell \rangle$, we have

$$\|\boldsymbol{w}\|^2 = \langle \boldsymbol{w}, \boldsymbol{w} \rangle = \langle \boldsymbol{e}_\ell, \boldsymbol{e}_\ell \rangle - 2\sum_{j=1}^m \langle \boldsymbol{q}_j, \boldsymbol{e}_\ell \rangle \langle \boldsymbol{e}_\ell, \boldsymbol{q}_j \rangle + \sum_{j=1}^m \langle \boldsymbol{q}_j, \boldsymbol{e}_\ell \rangle \sum_{k=1}^m \langle \boldsymbol{q}_k, \boldsymbol{e}_\ell \rangle \langle \boldsymbol{q}_j, \boldsymbol{q}_k \rangle = 1 - \sum_{j=1}^m Q_{\ell,j}^2 > 0.$$

Thus we define $\boldsymbol{q}_{m+1} = \boldsymbol{w}/\|\boldsymbol{w}\|$. □

4.2 The QR factorization

The QR factorization of an $m \times n$ matrix A has the form $A = QR$, where Q is an $m \times m$ *orthogonal* matrix and R is an $m \times n$ *upper triangular* matrix (i.e., $R_{i,j} = 0$ for $i > j$). We will demonstrate in the sequel that every matrix has a (non-unique) QR factorization.

An application Let $m = n$ and A be nonsingular. We can solve $A\boldsymbol{x} = \boldsymbol{b}$ by calculating the QR factorization of A and solving first $Q\boldsymbol{y} = \boldsymbol{b}$ (hence $\boldsymbol{y} = Q^\top \boldsymbol{b}$) and then $R\boldsymbol{x} = \boldsymbol{y}$ (a triangular system!).

Interpretation of the QR factorization Let $m \ge n$ and denote the columns of A and Q by $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$ and $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m$ respectively. Since

$$[\ \boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_n\ ] = [\ \boldsymbol{q}_1 \quad \boldsymbol{q}_2 \quad \cdots \quad \boldsymbol{q}_m\ ] \begin{bmatrix} R_{1,1} & R_{1,2} & \cdots & R_{1,n} \\ 0 & R_{2,2} & & \vdots \\ \vdots & \ddots & \ddots & \\ & & 0 & R_{n,n} \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix},$$

we have $\boldsymbol{a}_k = \sum_{j=1}^k R_{j,k} \boldsymbol{q}_j$, $k = 1, 2, \ldots, n$. In other words, Q has the property that each kth column of A can be expressed as a linear combination of the first k columns of Q.

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 5[1]

4.3 The Gram–Schmidt algorithm

Given a nonzero $m \times n$ matrix A with the columns $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \in \mathbb{R}^m$, we construct Q & R where Q is orthogonal, R upper-triangular and $A = QR$: in other words,

$$\sum_{k=1}^{\ell} R_{k,\ell}\boldsymbol{q}_\ell = \boldsymbol{a}_k, \quad k = 1, 2, \ldots, n, \qquad \text{where} \qquad A = [\ \boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_n\ ]. \tag{4.1}$$

Assuming $\boldsymbol{a}_1 \neq \boldsymbol{0}$, we derive $\boldsymbol{q}_1$ and $R_{1,1}$ from the equation (4.1) for $k = 1$. Since $\|\boldsymbol{q}_1\| = 1$, we let $\boldsymbol{q}_1 = \boldsymbol{a}_1/\|\boldsymbol{a}_1\|$, $R_{1,1} = \|\boldsymbol{a}_1\|$.
Next we form the vector $\boldsymbol{b} = \boldsymbol{a}_2 - \langle \boldsymbol{q}_1, \boldsymbol{a}_2\rangle \boldsymbol{q}_1$. It is orthogonal to $\boldsymbol{q}_1$, since

$$\langle \boldsymbol{q}_1, \boldsymbol{a}_2 - \langle \boldsymbol{q}_1, \boldsymbol{a}_2\rangle \boldsymbol{q}_1\rangle = \langle \boldsymbol{q}_1, \boldsymbol{a}_2\rangle - \langle \boldsymbol{q}_1, \boldsymbol{a}_2\rangle\langle \boldsymbol{q}_1, \boldsymbol{q}_1\rangle = 0.$$

If $\boldsymbol{b} \neq \boldsymbol{0}$, we set $\boldsymbol{q}_2 = \boldsymbol{b}/\|\boldsymbol{b}\|$, hence $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ are orthonormal. Moreover,

$$\langle \boldsymbol{q}_1, \boldsymbol{a}_2\rangle \boldsymbol{q}_1 + \|\boldsymbol{b}\|\boldsymbol{q}_2 = \langle \boldsymbol{q}_1, \boldsymbol{a}_2\rangle \boldsymbol{q}_1 + \boldsymbol{b} = \boldsymbol{a}_2,$$

hence, to obey (4.1) for $k = 2$, we let $R_{1,2} = \langle \boldsymbol{q}_1, \boldsymbol{a}_2\rangle$, $R_{2,2} = \|\boldsymbol{b}\|$.

The Gram–Schmidt algorithm The above idea can be extended to all columns of A.

Step 1 Set $k := 0$, $j := 0$ (k is the number of columns of Q that have been already formed and j is the number of columns of A that have been already considered, clearly $k \leq j$);

Step 2 Increase j by 1. If $k = 0$ then set $\boldsymbol{b} := \boldsymbol{a}_j$, otherwise (i.e., when $k \geq 1$) set $R_{i,j} := \langle \boldsymbol{q}_i, \boldsymbol{a}_j\rangle$, $i = 1, 2, \ldots, k$, and $\boldsymbol{b} := \boldsymbol{a}_j - \sum_{i=1}^{k}\langle \boldsymbol{q}_i, \boldsymbol{a}_j\rangle \boldsymbol{q}_i$. *[Note: $\boldsymbol{b}$ is orthogonal to $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_k$.]*

Step 3 If $\boldsymbol{b} \neq \boldsymbol{0}$ increase k by 1. Subsequently, set $\boldsymbol{q}_k := \boldsymbol{b}/\|\boldsymbol{b}\|$, $R_{k,j} := \|\boldsymbol{b}\|$ and $R_{i,j} := 0$ for $i \geq k+1$. *[Note: Hence, each column of Q has unit length, as required, $\boldsymbol{a}_j = \sum_{i=1}^{k} R_{i,j}\boldsymbol{q}_j$ and R is upper triangular, because $k \leq j$.]*

Step 4 Terminate if $j = n$, otherwise go to ***Step 2.***

Previous lecture ⇒ Since the columns of Q are orthonormal, there are at most m of them, i.e. the final value of k can't exceed m. If it is less then m then a previous lemma demonstrates that we can add columns so that Q becomes $m \times m$ and orthogonal.

The disadvantage of Gram–Schmidt is its *ill-conditioning*. Since we are using finite arithmetic, even small imprecisions in the calculation of inner products rapidly lead to effective loss of orthogonality. Thus, errors accumulate fast and even for moderate values of m it is no longer true that the computed off-diagonal elements of $Q^\top Q$ are very small in magnitude.

On the other hand, orthogonality conditions are preserved well when one generates a new orthogonal matrix by computing the product of two given orthogonal matrices. Therefore algorithms that express Q as a product of simple orthogonal matrices are highly useful. This suggests an alternative way forward.

4.4 Orthogonal transformations

Given real, $m \times n$ matrix $A_0 = A$, we seek a sequence $\Omega_1, \Omega_2, \ldots, \Omega_k$ of $m \times m$ orthogonal matrices such that the matrix $A_i := \Omega_i A_{i-1}$ has more zero elements below the main diagonal than A_{i-1} for $i = 1, 2, \ldots, k$ and so that the manner of insertion of such zeros is such that A_k is upper triangular. We then let $R = A_k$, therefore

$$\Omega_k \Omega_{k-1} \cdots \Omega_2 \Omega_1 A = R$$

[1] Corrections and suggestions to these notes should be emailed to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/`.

and $Q = (\Omega_k\Omega_{k-1}\cdots\Omega_1)^{-1} = (\Omega_k\Omega_{k-1}\cdots\Omega_1)^\top = \Omega_1^\top\Omega_2^\top\cdots\Omega_k^\top$. Hence $A = QR$, where Q is orthogonal and R upper triangular.

4.5 Givens rotations

We say that an $m \times m$ orthogonal matrix Ω_j is a *Givens rotation* if it coincides with the unit matrix, except for four elements. Specifically, we use the notation $\Omega^{[p,q]}$, where $1 \le p < q \le m$ for a matrix such that

$$\Omega^{[p,q]}_{p,p} = \Omega^{[p,q]}_{q,q} = \cos\theta, \qquad \Omega^{[p,q]}_{p,q} = \sin\theta, \qquad \Omega^{[p,q]}_{q,p} = -\sin\theta$$

for some $\theta \in [-\pi, \pi]$. The remaining elements of $\Omega^{[p,q]}$ are those of a unit matrix. For example,

$$m = 4 \quad \Longrightarrow \quad \Omega^{[1,2]} = \begin{bmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \Omega^{[2,4]} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & 0 & \sin\theta \\ 0 & 0 & 1 & 0 \\ 0 & -\sin\theta & 0 & \cos\theta \end{bmatrix}.$$

Geometrically, such matrices correspond to the underlying coordinate system being rigidly rotated along a two-dimensional plane (in mechanics this is called an *Euler rotation*). It is trivial to confirm that they are orthogonal.

Theorem Let A be an $m \times n$ matrix. Then, for every $1 \le p < q \le m$, $i \in \{p, q\}$ and $1 \le j \le n$, there exists $\theta \in [-\pi, \pi]$ such that $(\Omega^{[p,q]}A)_{i,j} = 0$. Moreover, all the rows of $\Omega^{[p,q]}A$, except for the pth and the qth, are the same as the corresponding rows of A, whereas the pth and the qth rows are linear combinations of the 'old' pth and qth rows.

Proof. Let $i = q$. If $A_{p,j} = A_{q,j} = 0$ then any θ will do, otherwise we let

$$\cos\theta := A_{p,j}/\sqrt{A^2_{p,j} + A^2_{q,j}}, \qquad \sin\theta := A_{q,j}/\sqrt{A^2_{p,j} + A^2_{q,j}}.$$

Hence

$$(\Omega^{[p,q]}A)_{q,k} = -(\sin\theta)A_{p,k} + (\cos\theta)A_{q,k}, \quad k = 1, 2, \ldots, n \qquad \Rightarrow \qquad (\Omega^{[p,q]})_{q,j} = 0.$$

Likewise, when $i = p$ we let $\cos\theta := A_{q,j}/\sqrt{A^2_{p,j} + A^2_{q,j}}$, $\sin\theta := -A_{p,j}/\sqrt{A^2_{p,j} + A^2_{q,j}}$.

The last two statements of the theorem are an immediate consequence of the structure of $\Omega^{[p,q]}$. □

An example: Suppose that A is 3×3. We can force zeros underneath the main diagonal as follows.

1 First pick $\Omega^{[1,2]}$ so that $(\Omega^{[1,2]}A)_{2,1} = 0 \quad \Rightarrow \quad \Omega^{[1,2]}A = \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ \times & \times & \times \end{bmatrix}$.

2 Next pick $\Omega^{[1,3]}$ so that $(\Omega^{[1,3]}\Omega^{[1,2]}A)_{3,1} = 0$. Note that multiplication by $\Omega^{[1,3]}$ doesn't alter the second row, hence $(\Omega^{[1,3]}\Omega^{[1,2]}A)_{2,1}$ remains zero $\quad \Rightarrow \quad \Omega^{[1,3]}\Omega^{[1,2]}A = \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix}$.

3 Finally, pick $\Omega^{[2,3]}$ so that $(\Omega^{[2,3]}\Omega^{[1,3]}\Omega^{[1,2]}A)_{3,2} = 0$. Since both second and third row of $\Omega^{[1,3]}\Omega^{[1,2]}A$ have a leading zero, their linear combination preserves these zeros, hence also

$$(\Omega^{[2,3]}\Omega^{[1,3]}\Omega^{[1,2]}A)_{2,1} = (\Omega^{[2,3]}\Omega^{[1,3]}\Omega^{[1,2]}A)_{3,1} = 0.$$

It follows that $\Omega^{[2,3]}\Omega^{[1,3]}\Omega^{[1,2]}A$ is upper triangular. Therefore

$$R = \Omega^{[2,3]}\Omega^{[1,3]}\Omega^{[1,2]}A = \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix}, \qquad Q = (\Omega^{[2,3]}\Omega^{[1,3]}\Omega^{[1,2]})^\top.$$

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 6[1]

The Givens algorithm Given $m \times n$ matrix A, let ℓ_i be the number of leading zeros in the ith row of A, $i = 1, 2, \ldots, m$.

Step 1 Stop if the (integer) sequence $\{\ell_1, \ell_2, \ldots, \ell_m\}$ increases monotonically, the increase being strictly monotone for $\ell_i \le n$.

Step 2 Pick any two integers $1 \le p < q \le m$ such that either $\ell_p > \ell_q$ or $\ell_p = \ell_q < n$.

Step 3 Replace A by $\Omega^{[p,q]}A$, using the Givens rotation that annihilates the $(q, \ell_q + 1)$ element. Update the values of ℓ_p and ℓ_q and go to *Step 1.*

The final matrix A is upper triangular and also has the property that the number of leading zeros in each row increases *strictly monotonically* until all the rows of A are zero – a matrix of this form is said to be in *standard form*. This end result, as we recall, is the required matrix R.

The cost There are less than mn rotations and each rotation replaces two rows by their linear combinations, hence the total cost is $\mathcal{O}(mn^2)$.

If we wish to obtain explicitly an orthogonal Q s.t. $A = QR$ then we commence by letting Ω be the $m \times m$ unit matrix and, each time A is premultiplied by $\Omega^{[p,q]}$, we also premultiply Ω by the same rotation. Hence the final Ω is the product of all the rotations, in correct order, and we let $Q = \Omega^\top$. The extra cost is $\mathcal{O}(m^2 n)$. However, in most applications we don't need Q but, instead, just the action of $Q^\top$ on a given vector (recall: solution of linear systems!). This can be accomplished by multiplying the vector by successive rotations, the cost being $\mathcal{O}(mn)$.

4.6 Householder transformations

Let $\boldsymbol{u} \in \mathbb{R}^m \setminus \{\boldsymbol{0}\}$. The $m \times m$ matrix

$$I - 2\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^2}$$

is called a *Householder transformation* (or a *Householder reflection*). Each such matrix is symmetric and orthogonal, since

$$\left(I - 2\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^2}\right)^\top \left(I - 2\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^2}\right) = \left(I - 2\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^2}\right)^2 = I - 4\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^2} + 4\frac{\boldsymbol{u}(\boldsymbol{u}^\top\boldsymbol{u})\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^4} = I.$$

Householder transformations offer an alternative to Given rotations in the calculation of a QR factorization.

Deriving the first column of R Our goal is to multiply an $m \times n$ matrix A by a sequence of Householder transformations so that each product induces zeros under the diagonal in an entire successive column. To start with, we seek a reflection that transforms the first nonzero column of A to a multiple of $\boldsymbol{e}_1$.

Let $\boldsymbol{a} \in \mathbb{R}^m$ be the first nonzero column of A. We wish to choose $\boldsymbol{u} \in \mathbb{R}^m$ s.t. the bottom $m-1$ entries of

$$\left(I - 2\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^2}\right)\boldsymbol{a} = \boldsymbol{a} - 2\frac{\boldsymbol{u}^\top\boldsymbol{a}}{\|\boldsymbol{u}\|^2}\boldsymbol{u}$$

[1] Corrections and suggestions to these notes should be emailed to A.Iserles@damtp.cam.ac.uk. All handouts are available on the WWW at the URL http://www.damtp.cam.ac.uk/user/na/PartIB/.

vanish and, in addition, we normalise $\boldsymbol{u}$ so that $2\boldsymbol{u}^\top\boldsymbol{a} = \|\boldsymbol{u}\|^2$ (recall that $\boldsymbol{a} \neq \boldsymbol{0}$). Therefore $u_i = a_i$, $i = 2, \ldots, m$ and the normalisation implies that

$$2u_1a_1 + 2\sum_{i=2}^{m} a_i^2 = u_1^2 + \sum_{i=2}^{m} a_i^2 \quad \Rightarrow \quad u_1^2 - 2u_1a_1 + a_1^2 - \sum_{i=1}^{m} a_i^2 = 0 \quad \Rightarrow \quad u_1 = a_1 \pm \|\boldsymbol{a}\|.$$

It is usual to let the sign be the same as the sign of a_1, since $\|\boldsymbol{u}\| \ll 1$ might lead to a division by a tiny number, hence to numerical difficulties.

For large m we do not execute explicit matrix multiplication. Instead, to calculate

$$\left(I - 2\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^2}\right) A = A - 2\frac{\boldsymbol{u}(\boldsymbol{u}^\top A)}{\|\boldsymbol{u}\|^2},$$

we first evaluate $\boldsymbol{w}^\top := \boldsymbol{u}^\top A$, subsequently forming $A - \frac{2}{\|\boldsymbol{u}\|^2}\boldsymbol{u}\boldsymbol{w}^\top$.

Subsequent columns of R Suppose that $\boldsymbol{a}$ is the first column of A that isn't compatible with standard form (previous columns have been, presumably, already dealt with by Householder transformations) and that the standard form requires to bring the $k+1, \ldots, m$ components to zero. Hence, nonzero elements in previous columns must be confined to the first $k-1$ rows and we want them to be unamended by the reflection. Thus, we let the first $k-1$ components of $\boldsymbol{u}$ be zero and choose $u_k = a_k \pm \left(\sum_{i=k}^{m} a_i^2\right)^{1/2}$ and $u_i = a_i$, $i = k+1, \ldots, m$.

The Householder method We process columns of A in sequence, in each stage premultiplying a current A by the requisite Householder transformation. The end result is an upper triangular matrix R in its standard form.

Example

$$A = \begin{bmatrix} 2 & 4 & 7 \\ 0 & 3 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & -2 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{u} = \begin{bmatrix} 0 \\ 0 \\ 5 \\ 1 \\ -2 \end{bmatrix} \quad \Rightarrow \quad \left(I - 2\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^2}\right) A = \begin{bmatrix} 2 & 4 & 7 \\ 0 & 3 & -1 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Calculation of Q If the matrix Q is required in an explicit form, set $\Omega = I$ initially and, for each successive reflection, replace Ω by

$$\left(I - 2\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^2}\right) \Omega = \Omega - \frac{2}{\|\boldsymbol{u}\|^2}\boldsymbol{u}(\boldsymbol{u}^\top \Omega).$$

As in the case of Givens rotations, by the end of the computation, $Q = \Omega^\top$. However, if we require just the vector $\boldsymbol{c} = Q^\top \boldsymbol{b}$, say, rather than the matrix Q, then we set initially $\boldsymbol{c} = \boldsymbol{b}$ and in each stage replace $\boldsymbol{c}$ by

$$\left(I - 2\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\|\boldsymbol{u}\|^2}\right) \boldsymbol{c} = \boldsymbol{c} - 2\frac{\boldsymbol{u}^\top \boldsymbol{c}}{\|\boldsymbol{u}\|^2}\boldsymbol{u}.$$

Deciding between Givens and Householder transformations If A is dense, it is in general more convenient to use Householder reflections. Givens rotations come into their own, however, when A has many leading zeros in its rows. In an extreme case, if an $n \times n$ matrix A consists of zeros underneath the first subdiagonal, they can be 'rotated away' in just $n-1$ Givens rotations, at the cost of $\mathcal{O}(n^2)$ operations!

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 7[1]

5 Linear least squares

5.1 Statement of the problem

Suppose that an $m \times n$ matrix A and a vector $\boldsymbol{b} \in \mathbb{R}^m$ are given. The equation $A\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{x} \in \mathbb{R}^n$ is unknown, has in general no solution (if $m > n$) or an infinity of solutions (if $m < n$). Problems of this form occur frequently when we collect m observations (which, typically, are prone to measurement error) and wish to exploit them to form an n-variable linear model, where $n \ll m$. (In statistics, this is known as *linear regression.*) Bearing in mind the likely presence of errors in A and $\boldsymbol{b}$, we seek $\boldsymbol{x} \in \mathbb{R}^n$ that minimises the Euclidean length $\|A\boldsymbol{x} - \boldsymbol{b}\|$. This is the *least squares problem.*

Theorem $\boldsymbol{x} \in \mathbb{R}^n$ is a solution of the least squares problem iff $A^\top(A\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{0}$.

Proof. If $\boldsymbol{x}$ is a solution then it minimises

$$f(\boldsymbol{x}) := \|A\boldsymbol{x} - \boldsymbol{b}\|^2 = \langle A\boldsymbol{x} - \boldsymbol{b}, A\boldsymbol{x} - \boldsymbol{b}\rangle = \boldsymbol{x}^\top A^\top A\boldsymbol{x} - 2\boldsymbol{x}^\top A^\top \boldsymbol{b} + \boldsymbol{b}^\top \boldsymbol{b}.$$

Hence $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$. But $\frac{1}{2}\nabla f(\boldsymbol{x}) = A^\top A\boldsymbol{x} - A^\top \boldsymbol{b}$, hence $A^\top(A\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{0}$.
Conversely, suppose that $A^\top(A\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{0}$ and let $\boldsymbol{u} \in \mathbb{R}^n$. Hence, letting $\boldsymbol{y} = \boldsymbol{u} - \boldsymbol{x}$,

$$\begin{aligned}\|A\boldsymbol{u} - \boldsymbol{b}\|^2 &= \langle A\boldsymbol{x} + A\boldsymbol{y} - \boldsymbol{b}, A\boldsymbol{x} + A\boldsymbol{y} - \boldsymbol{b}\rangle = \langle A\boldsymbol{x} - \boldsymbol{b}, A\boldsymbol{x} - \boldsymbol{b}\rangle + 2\boldsymbol{y}^\top A^\top(A\boldsymbol{x} - \boldsymbol{b})\\ &\quad + \langle A\boldsymbol{y}, A\boldsymbol{y}\rangle = \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \|A\boldsymbol{y}\|^2 \geq \|A\boldsymbol{x} - \boldsymbol{b}\|^2\end{aligned}$$

and $\boldsymbol{x}$ is indeed optimal. □

Corollary Optimality of $\boldsymbol{x} \Leftrightarrow$ the vector $A\boldsymbol{x} - \boldsymbol{b}$ is orthogonal to all columns of A.

5.2 Normal equations

One way of finding optimal $\boldsymbol{x}$ is by solving the $n \times n$ linear system $A^\top A\boldsymbol{x} = A^\top \boldsymbol{b}$ – the method of *normal equations.* This approach is popular in many applications. However, there are three disadvantages. Firstly, $A^\top A$ might be singular, secondly sparse A might be replaced by a dense $A^\top A$ and, finally, forming $A^\top A$ might lead to loss of accuracy. Thus, suppose that our computer works in the IEEE arithmetic standard (≈ 15 significant digits) and let

$$A = \begin{bmatrix} 10^8 & -10^8 \\ 1 & 1 \end{bmatrix} \qquad \Longrightarrow \qquad A^\top A = \begin{bmatrix} 10^{16}+1 & -10^{16}+1 \\ -10^{16}+1 & 10^{16}+1 \end{bmatrix} \approx 10^{16} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Given $\boldsymbol{b} = [0, 2]^\top$ the solution of $A\boldsymbol{x} = \boldsymbol{b}$ is $[1, 1]^\top$, as can be easily found by Gaussian elimination. However, our computer 'believes' that $A^\top A$ is singular!

5.3 QR and least squares

Lemma Let A be any $m \times n$ matrix and let $\boldsymbol{b} \in \mathbb{R}^m$. The vector $\boldsymbol{x} \in \mathbb{R}^n$ minimises $\|A\boldsymbol{x} - \boldsymbol{b}\|$ iff it minimises $\|\Omega A\boldsymbol{x} - \Omega\boldsymbol{b}\|$ for an arbitrary $m \times m$ orthogonal matrix Ω.

Proof. Given an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^m$, we have

$$\|\Omega\boldsymbol{v}\|^2 = \boldsymbol{v}^\top \Omega^\top \Omega\boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{v} = \|\boldsymbol{v}\|^2.$$

[1] Corrections and suggestions to these notes should be emailed to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/`.

In particular, $\|\Omega A\boldsymbol{x} - \Omega\boldsymbol{b}\| = \|A\boldsymbol{x} - \boldsymbol{b}\|$. □

An irrelevant, yet important remark The property that orthogonal matrices leave the Euclidean distance intact is called *isometry* and it has many important ramifications throughout mathematics and mathematical physics.

Method of solution Suppose that $A = QR$, a QR factorization with R in a *standard form.* Because of the lemma, letting $\Omega := Q^\top$,

$$\|A\boldsymbol{x} - \boldsymbol{b}\| = \|Q^\top(A\boldsymbol{x} - \boldsymbol{b})\| = \|R\boldsymbol{x} - Q^\top\boldsymbol{b}\|,$$

therefore we seek $\boldsymbol{x} \in \mathbb{R}^n$ that minimises $\|R\boldsymbol{x} - Q^\top\boldsymbol{b}\|$.
In general ($m > n$) many rows of R consist of zeros. Suppose for simplicity that $\operatorname{rank} R = \operatorname{rank} A = n$. Then the bottom $m - n$ rows of R are zero. Therefore we find $\boldsymbol{x}$ by solving the (nonsingular) linear system given by the first n equations of $R\boldsymbol{x} = Q^\top\boldsymbol{b}$. Similar (although more complicated) algorithm applies when $\operatorname{rank} R \leq n - 1$. Note, recalling our former remark, that we don't require Q explicitly, just to evaluate $Q^\top\boldsymbol{b}$.

6 Polynomial interpolation

6.1 The interpolation problem

Given $n + 1$ distinct real points $x_0, x_1, \ldots, x_n$ and real numbers $f_0, f_1, \ldots, f_n$, we seek a function $p : \mathbb{R} \to \mathbb{R}$ such that $p(x_i) = f_i$, $i = 0, 1, \ldots, n$. Such a function is called an *interpolant.*
We denote by $\mathbb{P}_n[x]$ the set of all real polynomials of degree at most n and observe that each $p \in \mathbb{P}_n[x]$ is uniquely defined by its $n + 1$ coefficients. In other words, we have $n + 1$ degrees of freedom, while interpolation at $x_0, x_1, \ldots, x_n$ constitutes $n + 1$ conditions. This, intuitively, justifies seeking an interpolant from $\mathbb{P}_n[x]$.

6.2 The Lagrange formula

Although, in principle, we may solve a linear problem with $n + 1$ unknowns to determine a polynomial interpolant, this can be accomplished more easily by using the explicit *Lagrange formula.* We claim that

$$p(x) = \sum_{k=0}^{n} f_k \prod_{\substack{\ell=0 \\ \ell \neq k}}^{n} \frac{x - x_\ell}{x_k - x_\ell}, \qquad x \in \mathbb{R}.$$

Note that $p \in \mathbb{P}_n[x]$, as required. We wish to show that it interpolates the data. Define

$$L_k(x) := \prod_{\substack{\ell=0 \\ \ell \neq k}}^{n} \frac{x - x_\ell}{x_k - x_\ell}, \qquad j = 0, 1, \ldots, n$$

(*Lagrange cardinal polynomials*). It is trivial to verify that $L_j(x_j) = 1$ and $L_j(x_k) = 0$ for $k \neq j$, hence

$$p(x_j) = \sum_{k=0}^{n} f_k L_k(x_j) = f_j, \qquad j = 0, 1, \ldots, n,$$

and p is an interpolant,

Uniqueness Suppose that both $p \in \mathbb{P}_n[x]$ and $q \in \mathbb{P}_n[x]$ interpolate to the same $n + 1$ data. Then the nth degree polynomial $p - q$ vanishes at $n + 1$ distinct points. But the only nth-degree polynomial with $\geq n + 1$ zeros is the zero polynomial. Therefore $p - q \equiv 0$ and the interpolating polynomial is unique.

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 8[1]

6.3 The error of polynomial interpolation

Let $[a, b]$ be a closed interval of $\mathbb{R}$. We denote by $C[a, b]$ the space of all continuous functions from $[a, b]$ to $\mathbb{R}$ and let $C^s[a, b]$, where s is a positive integer, stand for the linear space of all functions in $C[a, b]$ that possess s continuous derivatives.

Theorem Given $f \in C^{n+1}[a, b]$, let $p \in \mathbb{P}_n[x]$ interpolate the values $f(x_i)$, $i = 0, 1, \ldots, n$, where $x_0, \ldots, x_n \in [a, b]$ are pairwise distinct. Then for every $x \in [a, b]$ there exists $\xi \in [a, b]$ such that

$$f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) \prod_{i=0}^{n} (x - x_i). \tag{6.1}$$

Proof. The formula (6.1) is true when $x = x_j$ for $j \in \{0, 1, \ldots, n\}$, since both sides of the equation vanish. Let $x \in [a, b]$ be any other point and define

$$\phi(t) := [f(t) - p(t)] \prod_{i=0}^{n} (x - x_i) - [f(x) - p(x)] \prod_{i=0}^{n} (t - x_i), \qquad t \in [a, b].$$

[Note: The variable in ϕ is t, whereas x is a fixed parameter.] Note that $\phi(x_j) = 0$, $j = 0, 1, \ldots, n$, and $\phi(x) = 0$. Hence, ϕ has at least $n + 2$ distinct zeros in $[a, b]$. Moreover, $\phi \in C^{n+1}[a, b]$.
We now apply the *Rolle theorem:* if the function $g \in C^1[a, b]$ vanishes at two distinct points in $[a, b]$ then its derivative vanishes at an intermediate point. We deduce that ϕ' vanishes at (at least) $n + 1$ distinct points in $[a, b]$. Next, applying Rolle to ϕ', we conclude that ϕ'' vanishes at n points in $[a, b]$. In general, we prove by induction that $\phi^{(s)}$ vanishes at $n + 2 - s$ distinct points of $[a, b]$ for $s = 0, 1, \ldots, n + 1$. Letting $s = n + 1$, we have $\phi^{(n+1)}(\xi) = 0$ for some $\xi \in [a, b]$. Hence

$$0 = \phi^{(n+1)}(\xi) = [f^{(n+1)}(\xi) - p^{(n+1)}(\xi)] \prod_{i=0}^{n} (x - x_i) - [f(x) - p(x)] \frac{\mathrm{d}^{n+1}}{\mathrm{d}t^{n+1}} \prod_{i=0}^{n} (\xi - x_i).$$

Since $p^{(n+1)} \equiv 0$ and $\mathrm{d}^{n+1} \prod_{i=0}^{n} (t - x_i)/\mathrm{d}t^{n+1} \equiv (n + 1)!$, we obtain (6.1). □

Runge's example We interpolate $f(x) = 1/(1 + x^2)$, $x \in [-5, 5]$, at the equally-spaced points $x_j = -5 + 10\frac{j}{n}$, $j = 0, 1, \ldots, n$. Some of the errors are displayed below

x	$f(x) - p(x)$	$\prod_{i=0}^{n}(x - x_i)$
0.75	3.2×10^{-3}	-2.5×10^{6}
1.75	7.7×10^{-3}	-6.6×10^{6}
2.75	3.6×10^{-2}	-4.1×10^{7}
3.75	5.1×10^{-1}	-7.6×10^{8}
4.75	$4.0 \times 10^{+2}$	-7.3×10^{10}

Table: Errors for $n = 20$

Figure: Errors for $n = 15$

The growth in the error is explained by the product term in (6.1) (the rightmost column of the table). Adding more interpolation points makes the largest error even worse. A remedy to this

[1] Corrections and suggestions to these notes should be emailed to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/`.

state of affairs is to cluster points toward the end of the range. A considerably smaller error is attained for $x_j = 5\cos\frac{(n-j)\pi}{n}$, $j = 0, 1, \ldots, n$ (so-called *Chebyshev points*). It is possible to prove that this choice of points minimizes the magnitude of $\max_{x\in[-5,5]} |\prod_{i=0}^{n}(x - x_i)|$.

6.4 Divided differences: a definition

Given pairwise-distinct points $x_0, x_1, \ldots, x_n \in [a, b]$, we let $p \in \mathbb{P}_n[x]$ interpolate $f \in C[a, b]$ there. The coefficient of x^n in p is called the *divided difference* and denoted by $f[x_0, x_1, \ldots, x_n]$. We say that this divided difference is of *degree n*.
We can derive $f[x_0, \ldots, x_n]$ from the Lagrange formula,

$$f[x_0, x_1, \ldots, x_n] = \sum_{k=0}^{n} f(x_k) \prod_{\substack{\ell=0\\ \ell\neq k}}^{n} \frac{1}{x_k - x_\ell}. \tag{6.2}$$

Theorem Let $[\bar{a}, \bar{b}]$ be the shortest interval that contains $x_0, x_1, \ldots, x_n$ and let $f \in C^n[\bar{a}, \bar{b}]$. Then there exists $\xi \in [\bar{a}, \bar{b}]$ such that

$$f[x_0, x_1, \ldots, x_n] = \tfrac{1}{n!} f^{(n)}(\xi). \tag{6.3}$$

Proof. Let p be the interpolating polynomial. The error function $f - p$ has at least $n+1$ zeros in $[\bar{a}, \bar{b}]$ and, applying Rolle's theorem n times, it follows that $f^{(n)} - p^{(n)}$ vanishes at some $\xi \in [\bar{a}, \bar{b}]$. But $p(x) = \frac{1}{n!} p^{(n)}(\zeta)x^n +$ lower order terms (for any $\zeta \in \mathbb{R}$), therefore, letting $\zeta = \xi$,

$$f[x_0, x_1, \ldots, x_n] = \tfrac{1}{n!} p^{(n)}(\xi) = \tfrac{1}{n!} f^{(n)}(\xi)$$

and we deduce (6.3). □

Application It is a consequence of the theorem that divided differences can be used to approximate derivatives.

6.5 Recurrence relations for divided differences

Our next topic is a useful way to calculate divided differences (and, ultimately, to derive yet another means to construct an interpolating polynomial). We commence with the remark that $f[x_i]$ is the coefficient of x^0 in the polynomial of degree 0 (i.e., a constant) that interpolates $f(x_i)$, hence $f[x_i] = f(x_i)$.

Theorem Suppose that $x_0, x_1, \ldots, x_{k+1}$ are pairwise distinct, where $k \geq 0$. Then

$$f[x_0, x_1, \ldots, x_{k+1}] = \frac{f[x_1, x_2, \ldots, x_{k+1}] - f[x_0, x_1, \ldots, x_k]}{x_{k+1} - x_0}. \tag{6.4}$$

Proof. Let $p, q \in \mathbb{P}_k[x]$ be the polynomials that interpolate f at

$$\{x_0, x_1, \ldots, x_k\} \qquad \text{and} \qquad \{x_1, x_2, \ldots, x_{k+1}\}$$

respectively and define

$$r(x) := \frac{(x - x_0)q(x) + (x_{k+1} - x)p(x)}{x_{k+1} - x_0} \in \mathbb{P}_{k+1}[x].$$

We readily verify that $r(x_i) = f(x_i)$, $i = 0, 1, \ldots, k+1$. Hence r is the $(k+1)$-degree interpolating polynomial and $f[x_0, \ldots, x_{k+1}]$ is the coefficient of x^{k+1} therein. The recurrence (6.4) follows from the definition of divided differences. □

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 9[1]

6.6 The Newton interpolation formula

Recalling that $f[x_i] = f(x_i)$, the recursive formula allows for fast evaluation of the *divided difference table,* in the following manner:

$$\begin{tikzcd}
f[x_0] \arrow[r] & f[x_0,x_1] \arrow[r] & f[x_0,x_1,x_2] \arrow[r] & f[x_0,x_1,x_2,x_3] \arrow[r] & \cdots \\
f[x_1] \arrow[r] \arrow[ur] & f[x_1,x_2] \arrow[r] \arrow[ur] & f[x_1,x_2,x_3] \arrow[r] \arrow[ur] & \cdots & \\
\vdots & & & & \\
f[x_n] & & & &
\end{tikzcd}$$

This can be done in $\mathcal{O}\big(n^2\big)$ operations and the outcome are the numbers $\{f[x_0,x_1,\ldots,x_l]\}_{l=0}^{k}$. We now provide an alternative representation of the interpolating polynomial. Again, $f(x_i)$, $i = 0,1,\ldots,k$, are given and we seek $p \in \mathbb{P}_k[x]$ such that $p(x_i) = f(x_i)$, $i = 0,\ldots,k$.

Theorem Suppose that $x_0, x_1, \ldots, x_k$ are pairwise distinct. The polynomial

$$p_k(x) := f[x_0] + f[x_0,x_1](x - x_0) + \cdots + f[x_0,x_1,\ldots,x_k]\prod_{i=0}^{k-1}(x - x_i) \in \mathbb{P}_k[x]$$

obeys $p_k(x_i) = f(x_i)$, $i = 0,1,\ldots,k$.

Proof. By induction on k. The statement is obvious for $k = 0$ and we suppose that it is true for k. We now prove that $p_{k+1}(x) - p_k(x) = f[x_0,x_1,\ldots,x_{k+1}]\prod_{i=0}^{k}(x - x_i)$. Clearly, $p_{k+1} - p_k \in \mathbb{P}_{k+1}[x]$ and the coefficient of x^{k+1} therein is, by definition, $f[x_0,\ldots,x_{k+1}]$. Moreover, $p_{k+1}(x_i) - p_k(x_i) = 0$, $i = 0,1,\ldots,k$, hence it is a multiple of $\prod_{i=0}^{k}(x - x_i)$, and this proves the asserted form of $p_{k+1} - p_k$. The explicit form of p_{k+1} follows by adding $p_{k+1} - p_k$ to p_k. □

We have derived the *Newton interpolation formula,* which requires only the top row of the divided difference table. It has several advantages over Lagrange's. In particular, its evaluation at a given point x (provided that divided differences are known) requires just $\mathcal{O}(k)$ operations, as long as we do it by the *Horner scheme*

$$\begin{aligned} p_k(x) = \{\{\{f[x_0,\ldots,x_k](x - x_{k-1}) + f[x_0,\ldots,x_{k-1}]\} \times (x - x_{k-2}) + f[x_0,\ldots,x_{k-2}]\} \\ \times (x - x_3) + \cdots\} + f[x_0]. \end{aligned}$$

On the other hand, the Lagrange formula is often better when we wish to manipulate the interpolation polynomial as part of a larger mathematical expression. We'll see an example in the section on *Gaussian quadrature.*

7 Orthogonal polynomials

7.1 Orthogonality in general linear spaces

We have already seen the scalar product $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{i=1}^{n} x_i y_i$, acting on $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Likewise, given arbitrary *weights* $w_1, w_2, \ldots, w_n > 0$, we may define $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{i=1}^{n} w_i x_i y_i$. In general, a *scalar* (or

[1] Corrections and suggestions to these notes should be emailed to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/`.

inner) *product* is any function $\mathbb{V} \times \mathbb{V} \to \mathbb{R}$, where $\mathbb{V}$ is a vector space over the reals, subject to the following three axioms:
Symmetry: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle \; \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{V}$;
Nonnegativity: $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0 \; \forall \boldsymbol{x} \in \mathbb{V}$ and $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ iff $\boldsymbol{x} = \boldsymbol{0}$; and
Linearity: $\langle a\boldsymbol{x} + b\boldsymbol{y}, \boldsymbol{z} \rangle = a\langle \boldsymbol{x}, \boldsymbol{z} \rangle + b\langle \boldsymbol{y}, \boldsymbol{z} \rangle \; \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{V}$, $a, b \in \mathbb{R}$.
Given a scalar product, we may define *orthogonality:* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{V}$ are orthogonal if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.
Let $\mathbb{V} = C[a, b]$, $w \in \mathbb{V}$ be a fixed *positive* function and define $\langle f, g \rangle := \int_a^b w(x) f(x) g(x) \, \mathrm{d}x$ for all $f, g \in \mathbb{V}$. It is easy to verify all three axioms of the scalar product.

7.2 Orthogonal polynomials – definition, existence, uniqueness

Given a scalar product in $\mathbb{V} = \mathbb{P}_n[x]$, we say that $p_n \in \mathbb{P}_n[x]$ is the nth *orthogonal polynomial* if $\langle p_n, p \rangle = 0$ for all $p \in \mathbb{P}_{n-1}[x]$. *[Note: different inner products lead to different orthogonal polynomials.]* A polynomial in $\mathbb{P}_n[x]$ is *monic* if the coefficient of x^n therein equals one.

Theorem For every $n \geq 0$ there exists a unique monic orthogonal polynomial of degree n. Moreover, any $p \in \mathbb{P}_n[x]$ can be expanded as a linear combination of $p_0, p_1, \ldots, p_n$,

Proof. We let $p_0(x) \equiv 1$ and prove the theorem by induction on n. Thus, suppose that $p_0, p_1, \ldots, p_n$ have been already derived consistently with both assertions of the theorem and let $q(x) := x^{n+1} \in \mathbb{P}_{n+1}[x]$. Motivated by the *Gram–Schmidt algorithm,* we choose

$$p_{n+1}(x) = q(x) - \sum_{k=0}^{n} \frac{\langle q, p_k \rangle}{\langle p_k, p_k \rangle} p_k(x), \qquad x \in \mathbb{R}. \tag{7.1}$$

Clearly, $p_{n+1} \in \mathbb{P}_{n+1}[x]$ and it is monic (since all the terms in the sum are of degree $\leq n$).
Let $m \in \{0, 1, \ldots, n\}$. It follows from (7.1) and the induction hypothesis that

$$\langle p_{n+1}, p_m \rangle = \langle q, p_m \rangle - \sum_{k=0}^{n} \frac{\langle q, p_k \rangle}{\langle p_k, p_k \rangle} \langle p_k, p_m \rangle = \langle q, p_m \rangle - \frac{\langle q, p_m \rangle}{\langle p_m, p_m \rangle} \langle p_m, p_m \rangle = 0.$$

Hence, p_{n+1} is orthogonal to $p_0, \ldots, p_n$. Consequently, according to the second inductive assertion, it is orthogonal to all $p \in \mathbb{P}_n[x]$.
To prove uniqueness, we suppose the existence of two monic orthogonal polynomials $p_{n+1}, \tilde{p}_{n+1} \in \mathbb{P}_{n+1}[x]$. Let $p := p_{n+1} - \tilde{p}_{n+1} \in \mathbb{P}_n[x]$, hence $\langle p_{n+1}, p \rangle = \langle \tilde{p}_{n+1}, p \rangle = 0$, and this implies

$$0 = \langle p_{n+1}, p \rangle - \langle \tilde{p}_{n+1}, p \rangle = \langle p_{n+1} - \tilde{p}_{n+1}, p \rangle = \langle p, p \rangle,$$

and we deduce $p \equiv 0$.
Finally, in order to prove that each $p \in \mathbb{P}_{n+1}[x]$ is a linear combination of $p_0, \ldots, p_{n+1}$, we note that we can always write it in the form $p = c p_{n+1} + q$, where c is the coefficient of x^{n+1} in p and where $q \in \mathbb{P}_n[x]$. According to the induction hypothesis, q can be expanded as a linear combination of $p_0, p_1, \ldots, p_n$, hence our assertion is true. □

Well-known examples of orthogonal polynomials include

Name	Notation	Interval	Weight function
Legendre	P_n	$[-1, 1]$	$w(x) \equiv 1$
Chebyshev	T_n	$[-1, 1]$	$w(x) = (1 - x^2)^{-1/2}$
Laguerre	L_n	$[0, \infty)$	$w(x) = \mathrm{e}^{-x}$
Hermite	H_n	$(-\infty, \infty)$	$w(x) = \mathrm{e}^{-x^2}$

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 10[1]

7.3 The three-term recurrence relation

How to construct orthogonal polynomials? (7.1) might help, but it suffers from the same problem as the Gram–Schmidt algorithm in Euclidean spaces: loss of accuracy due to imprecisions in the calculation of scalar products. A considerably better procedure follows from our next theorem.

Theorem Monic orthogonal polynomials are given by the formula

$$\begin{aligned} &p_{-1}(x) \equiv 0, \qquad\qquad p_0(x) \equiv 1, \\ &p_{n+1}(x) = (x - \alpha_n)p_n(x) - \beta_n p_{n-1}(x), \qquad n = 0, 1, \ldots, \end{aligned} \tag{7.2}$$

where

$$\alpha_n := \frac{\langle p_n, xp_n \rangle}{\langle p_n, p_n \rangle}, \qquad \beta_n = \frac{\langle p_n, p_n \rangle}{\langle p_{n-1}, p_{n-1} \rangle} > 0.$$

Proof. Pick $n \geq 0$ and let

$$\psi(x) := p_{n+1}(x) - (x - \alpha_n)p_n(x) + \beta_n p_{n-1}(x).$$

Since p_n and p_{n+1} are monic, it follows that $\psi \in \mathbb{P}_n[x]$. Moreover, because of orthogonality of p_{n-1}, p_n, p_{n+1},

$$\langle \psi, p_\ell \rangle = \langle p_{n+1}, p_\ell \rangle - \langle p_n, (x - \alpha_n)p_\ell \rangle + \beta_n \langle p_{n-1}, p_\ell \rangle = 0, \qquad \ell = 0, 1, \ldots, n-2.$$

Because of monicity, $xp_{n-1} = p_n + q$, where $q \in \mathbb{P}_{n-1}[x]$. Thus, from the definition of α_n, β_n,

$$\begin{aligned} \langle \psi, p_{n-1} \rangle &= -\langle p_n, xp_{n-1} \rangle + \beta_n \langle p_{n-1}, p_{n-1} \rangle = -\langle p_n, p_n \rangle + \beta_n \langle p_{n-1}, p_{n-1} \rangle = 0, \\ \langle \psi, p_n \rangle &= -\langle xp_n, p_n \rangle + \alpha_n \langle p_n, p_n \rangle = 0. \end{aligned}$$

Every $p \in \mathbb{P}_n[x]$ that obeys $\langle p, p_\ell \rangle = 0$, $\ell = 0, 1, \ldots, n$, must necessarily be the zero polynomial. For suppose that it is not so and let x^s be the highest coefficient of x in p. Then $\langle p, p_s \rangle \neq 0$, which is impossible. We deduce that $\psi \equiv 0$, hence (7.2) is true. □

Example *Chebyshev polynomials* We choose the scalar product

$$\langle f, g \rangle := \int_{-1}^{1} f(x)g(x) \frac{\mathrm{d}x}{\sqrt{1 - x^2}}, \qquad f, g \in C[-1, 1]$$

and define $T_n \in \mathbb{P}_n[x]$ by the relation $T_n(\cos\theta) = \cos(n\theta)$. Hence $T_0(x) \equiv 1$, $T_1(x) = x$, $T_2(x) = 2x^2 - 1$ etc. Changing the integration variable,

$$\begin{aligned} \langle T_n, T_m \rangle &= \int_{-1}^{1} T_n(x)T_m(x) \frac{\mathrm{d}x}{\sqrt{1 - x^2}} = \int_0^\pi \cos n\theta \cos m\theta \, \mathrm{d}\theta \\ &= \tfrac{1}{2} \int_0^\pi [\cos(n+m)\theta + \cos(n-m)\theta] \, \mathrm{d}\theta = 0 \end{aligned}$$

whenever $n \neq m$. The recurrence relation for Chebyshev polynomials is particularly simple,

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),$$

as can be verified at once from the identity

$$\cos[(n+1)\theta] + \cos[(n-1)\theta] = 2\cos(\theta)\cos(n\theta).$$

Note that the T_ns aren't monic, hence the inconsistency with (7.2). To obtain monic polynomials take $T_n(x)/2^{n-1}$, $n \geq 1$.

[1] Corrections and suggestions to these notes should be emailed to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/`.

7.4 Least-squares polynomial fitting

Given $f \in C[a,b]$ and a scalar product $\langle g,h\rangle = \int_a^b w(x)g(x)h(x)\,\mathrm{d}x$, we wish to pick $p \in \mathbb{P}_n[x]$ so as to minimise $\langle f-p, f-p\rangle$. Again, we stipulate that $w(x) > 0$ for $x \in (a,b)$. Intuitively speaking, p approximates f and is an alternative to an interpolating polynomial. (The situation is similar to the one that we have already encountered in numerical linear algebra, least-squares' fitting *vs* solving linear equations.)
Let $p_0, p_1, \ldots, p_n$ be orthogonal polynomials w.r.t. the underlying inner product, $p_\ell \in \mathbb{P}_\ell[x]$. They form a basis of $\mathbb{P}_n[n]$, therefore for every $p \in \mathbb{P}_n$ there exist $c_0, c_1, \ldots, c_n \in \mathbb{R}$ such that $p = \sum_{k=0}^n c_k p_k$. Because of orthogonality,

$$\langle f-p, f-p\rangle = \left\langle f - \sum_{k=0}^n c_k p_k, f - \sum_{k=0}^n c_k p_k \right\rangle = \langle f,f\rangle - 2\sum_{k=0}^n c_k\langle p_k, f\rangle + \sum_{k=0}^n c_k^2 \langle p_k, p_k\rangle.$$

To derive optimal $c_0, c_1, \ldots, c_n$ we seek to minimise the last expression. (Note that it is a quadratic function in the c_is.) Since

$$\tfrac{1}{2}\frac{\partial}{\partial c_k}\langle f-p, f-p\rangle = -\langle p_k, f\rangle + c_k\langle p_k, p_k\rangle, \qquad k = 0,1,\ldots,n,$$

setting the gradient to zero yields

$$p(x) = \sum_{k=0}^n \frac{\langle p_k, f\rangle}{\langle p_k, p_k\rangle} p_k(x). \tag{7.3}$$

Note that

$$\langle f-p, f-p\rangle = \langle f,f\rangle - \sum_{k=0}^n \{2c_k\langle p_k, f\rangle - c_k^2\langle p_k, p_k\rangle\} = \langle f,f\rangle - \sum_{k=0}^n \frac{\langle p_k, f\rangle^2}{\langle p_k, p_k\rangle}. \tag{7.4}$$

This identity can be rewritten as $\langle f-p, f-p\rangle + \langle p,p\rangle = \langle f,f\rangle$, reminiscent of the Pythagoras theorem.

How to choose n**?** Note that $c_k = \langle p_k, f\rangle / \langle p_k, p_k\rangle$ is independent of n. Thus, we can continue to add terms to (7.3) until $\langle f-p, f-p\rangle$ is below specified *tolerance* ε. Because of (7.4), we need to pick n so that $\langle f,f\rangle - \varepsilon < \sum_{k=0}^n \langle p_k, f\rangle^2/\langle p_k, p_k\rangle$.

Theorem *(The Parseval identity)* Let $[a,b]$ be finite. Then

$$\sum_{k=0}^\infty \frac{\langle p_k, f\rangle^2}{\langle p_k, p_k\rangle} = \langle f,f\rangle. \tag{7.5}$$

Incomplete proof. Let

$$\sigma_n := \sum_{k=0}^n \frac{\langle p_k, f\rangle^2}{\langle p_k, p_k\rangle}, \qquad n = 0,1,\ldots,$$

hence $\langle f-p, f-p\rangle = \langle f,f\rangle - \sigma_n \geq 0$. The sequence $\{\sigma\}_{n=0}^\infty$ increases monotonically and $\sigma_n \leq \langle f,f\rangle$ implies that $\lim_{n\to\infty}\sigma_n$ exists. According to the *Weierstrass theorem*, any function in $C[a,b]$ can be approximated arbitrarily close by a polynomial, hence $\lim_{n\to\infty}\langle f-p, f-p\rangle = 0$ and we deduce that $\sigma_n \overset{n\to\infty}{\longrightarrow} \langle f,f\rangle$ and (7.5) is true. □

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 11[1]

7.5 Least-squares fitting to discrete function values

Suppose that $m \ge n+1$. We are given m function values $f(x_1), f(x_2), \ldots, f(x_m)$, where the x_ks are pairwise distinct, and seek $p \in \mathbb{P}_n[x]$ that minimises $\langle f-p, f-p\rangle$, where

$$\langle g, h\rangle := \sum_{k=1}^{m} g(x_k)h(x_k). \tag{7.6}$$

One alternative is to express p as $\sum_{\ell=0}^{n} c_\ell x^\ell$ and find optimal $c_0, \ldots, c_n$ as a solution of a linear least squares problem similarly to Section 5, using QR factorization. An alternative is to construct orthogonal polynomials w.r.t. the scalar product (7.6). The theory is identical to that of subsections 7.1–4, except that we have enough data to evaluate only $p_0, p_1, \ldots, p_{m-1}$. However, we need just $p_0, p_1, \ldots, p_n$ and $n \le m-1$, and we have enough information to implement the algorithm. Thus

1. Employ the three-term recurrence (7.2) to calculate $p_0, p_1, \ldots, p_n$ (of course, using the scalar product (7.6));
2. Form $p(x) = \displaystyle\sum_{k=0}^{n} \frac{\langle p_k, f\rangle}{\langle p_k, p_k\rangle} p_k(x)$.

Since the work for each k is bounded by a constant multiple of m, the complete cost is $\mathcal{O}(mn)$, as compared with $\mathcal{O}\big(n^2 m\big)$ if QR is used.

7.6 Gaussian quadrature

We are again in $C[a,b]$ and a scalar product is defined as in subsection 7.1, namely $\langle f, g\rangle = \int_a^b w(x)f(x)g(x)\,\mathrm{d}x$, where $w(x) > 0$ for $x \in (a,b)$. Our goal is to approximate integrals by finite sums,

$$\int_a^b w(x)f(x)\,\mathrm{d}x \approx \sum_{k=1}^{\nu} b_k f(c_k), \qquad f \in C[a,b].$$

The above is known as a *quadrature formula.* Here ν is given, whereas the points $b_1, \ldots, b_\nu$ (the *weights*) and $c_1, \ldots, c_\nu$ (the *nodes*) are independent of the choice of f.

A reasonable approach to achieving high accuracy is to require that the approximant is exact for all $f \in \mathbb{P}_m[x]$, where m is as large as possible – this results in *Gaussian quadrature* and we will demonstrate that $m = 2\nu - 1$ can be attained.

Firstly, we claim that $m = 2\nu$ is impossible. To prove this, choose arbitrary nodes $c_1, \ldots, c_\nu$ and note that $p(x) := \prod_{k=1}^{\nu}(x-c_k)^2$ lives in $\mathbb{P}_{2\nu}[x]$. But $\int_a^b w(x)p(x)\,\mathrm{d}x > 0$, while $\sum_{k=1}^{\nu} b_k p(c_k) = 0$ for any choice of weights $b_1, \ldots, b_\nu$. Hence the integral and the quadrature do not match.

Let $p_0, p_1, p_2, \ldots$ denote, as before, the monic polynomials which are orthogonal w.r.t. the underlying scalar product.

Theorem Given $n \ge 1$, all the zeros of p_n are real, distinct and lie in the interval (a,b).

Proof. Recall that $p_0 \equiv 1$. Thus, by orthogonality,

$$\int_a^b w(x)p_n(x)\,\mathrm{d}x = \int_a^b w(x)p_0(x)p_n(x)\,\mathrm{d}x = \langle p_0, p_n\rangle = 0$$

[1] Corrections and suggestions to these notes should be emailed to A.Iserles@damtp.cam.ac.uk. All handouts are available on the WWW at the URL http://www.damtp.cam.ac.uk/user/na/PartIB/.

and we deduce that p_n changes sign at least once in (a,b).
Denote by $m \geq 1$ the number of the sign changes of p_n in (a,b) and assume that $m \leq n-1$. Denoting the points where a sign change occurs by $\xi_1, \xi_2, \ldots, \xi_m$, we let $q(x) := \prod_{j=1}^{m}(x-\xi_j)$. Since $q \in \mathbb{P}_m[x]$, $m \leq n-1$, it follows that $\langle q, p_n \rangle = 0$. On the other hand, it follows from our construction that $q(x)p_n(x)$ does not change sign throughout $[a,b]$ and vanishes at a finite number of points, hence

$$|\langle q, p_n \rangle| = \left| \int_a^b w(x)q(x)p_n(x)\,\mathrm{d}x \right| = \int_a^b w(x)|q(x)p_n(x)|\,\mathrm{d}x > 0,$$

a contradiction. It follows that $m = n$ and the proof is complete. □

We commence our construction of *Gaussian quadrature* by choosing pairwise-distinct nodes $c_1, c_2, \ldots, c_\nu \in [a,b]$ and define the *interpolatory weights*

$$b_k := \int_a^b w(x) \prod_{\substack{j=1 \\ j \neq k}}^{\nu} \frac{x-c_j}{c_k-c_j}\,\mathrm{d}x, \qquad k = 1, 2, \ldots, \nu.$$

Theorem The quadrature formula with the above choice is exact for all $f \in \mathbb{P}_{\nu-1}[x]$. Moreover, if $c_1, c_2, \ldots, c_\nu$ are the zeros of p_ν then it is exact for all $f \in \mathbb{P}_{2\nu-1}[x]$.

Proof. Every $f \in \mathbb{P}_{\nu-1}[x]$ is its own interpolating polynomial, hence by Lagrange's formula

$$f(x) = \sum_{k=0}^{\nu} f(c_k) \prod_{\substack{j=1 \\ j \neq k}}^{\nu} \frac{x-c_j}{c_k-c_j}. \tag{7.7}$$

The quadrature is exact for all $f \in \mathbb{P}_{\nu-1}[x]$ if $\int_a^b w(x)f(x)\,\mathrm{d}x = \sum_{k=1}^{\nu} b_k f(c_k)$, and this, in tandem with the interpolating-polynomial representation, yields the stipulated form of $b_1, \ldots, b_\nu$.
Let $c_1, \ldots, c_\nu$ be the zeros of p_ν. Given any $f \in \mathbb{P}_{2\nu-1}[x]$, we can represent it uniquely as $f = qp_\nu + r$, where $q, r \in \mathbb{P}_{\nu-1}[x]$. Thus, by orthogonality,

$$\begin{aligned} \int_a^b w(x)f(x)\,\mathrm{d}x &= \int_a^b w(x)[q(x)p_\nu(x) + r(x)]\,\mathrm{d}x = \langle q, p_\nu \rangle + \int_a^b w(x)r(x)\,\mathrm{d}x \\ &= \int_a^b w(x)r(x)\,\mathrm{d}x. \end{aligned}$$

On the other hand, the choice of quadrature knots gives

$$\sum_{k=1}^{\nu} b_k f(c_k) = \sum_{k=1}^{\nu} b_k [q(c_k)p_\nu(c_k) + r(c_k)] = \sum_{k=1}^{\nu} b_k r(c_k).$$

Hence the integral and its approximant coincide, because $r \in \mathbb{P}_{\nu-1}[x]$ and the quadrature is exact for all polynomials in $\mathbb{P}_{\nu-1}[x]$. □

Example Let $[a,b] = [-1,1]$, $w(x) \equiv 1$. Then the underlying orthogonal polynomials are the *Legendre polynomials:* $P_0 \equiv 1$, $P_1(x) = x$, $P_2(x) = \frac{3}{2}x^2 - \frac{1}{2}$, $P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x$, $P_4(x) = \frac{35}{8}x^4 - \frac{15}{4}x^2 + \frac{3}{8}$ (it is customary to use this, non-monic, normalisation). The nodes of Gaussian quadrature are
$n = 1$: $c_1 = 0$;
$n = 2$: $c_1 = -\frac{\sqrt{3}}{3}$, $c_2 = \frac{\sqrt{3}}{3}$;
$n = 3$: $c_1 = -\frac{\sqrt{15}}{5}$, $c_2 = 0$, $c_3 = \frac{\sqrt{15}}{5}$;
$n = 4$: $c_1 = -\sqrt{\frac{3}{7} + \frac{2}{35}\sqrt{30}}$, $c_2 = -\sqrt{\frac{3}{7} - \frac{2}{35}\sqrt{30}}$, $c_3 = \sqrt{\frac{3}{7} - \frac{2}{35}\sqrt{30}}$, $c_4 = \sqrt{\frac{3}{7} + \frac{2}{35}\sqrt{30}}$.

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Lecture 12[1]

8 The Peano kernel theorem

8.1 The theorem

Our point of departure is the *Taylor formula with an integral remainder term,*

$$f(x) = f(a)+(x-a)f'(a)+\frac{(x-a)^2}{2!}f''(a)+\cdots+\frac{(x-a)^k}{k!}f^{(k)}(a)+\frac{1}{k!}\int_a^x (x-\theta)^k f^{(k+1)}(\theta)\,\mathrm{d}\theta, \quad (8.1)$$

which can be verified by integration by parts for functions $f \in \mathrm{C}^{k+1}[a,b]$, $a < b$. Suppose that we are given an approximant (e.g. to a function, a derivative at a given point, an integral etc.) which is *exact* for all $f \in \mathbb{P}_k[x]$. The Taylor formula produces an expression for the error that depends on $f^{(k+1)}$. This is the basis for the *Peano kernel theorem.*

Formally, let $L(f)$ be an error of an approximant. Thus, L maps $\mathrm{C}^{k+1}[a,b]$, say, to $\mathbb{R}$. We assume that it is *linear*, i.e. $L(\alpha f+\beta g) = \alpha L(f)+\beta L(g)\ \forall \alpha,\beta \in \mathbb{R}$, and that $L(f) = 0$ for all $f \in \mathbb{P}_k[x]$. In general, a linear mapping from a function space (e.g. $\mathrm{C}^{k+1}[a,b]$) to $\mathbb{R}$ is called a *linear functional.* The formula (8.1) implies

$$L(f) = \frac{1}{k!}L\left\{\int_a^x (x-\theta)^k f^{(k+1)}(\theta)\,\mathrm{d}\theta\right\}, \qquad a \le x \le b.$$

To make the range of integration independent of x, we introduce the notation

$$(x-\theta)_+^k := \begin{cases} (x-\theta)^k, & x \ge \theta, \\ 0, & x \le \theta, \end{cases} \qquad \text{whence} \qquad L(f) = \frac{1}{k!}L\left\{\int_a^b (x-\theta)_+^k f^{(k+1)}(\theta)\,\mathrm{d}\theta\right\}.$$

Let $K(\theta) := L[(x-\theta)_+^k]$ for $x \in [a,b]$. *[Note: K is independent of f.]* The function K is called the *Peano kernel* of L. ***Suppose that it is allowed to exchange the order of action of $\int$ and L.*** Because of the linearity of L, we then have

$$L(f) = \frac{1}{k!}\int_a^b K(\theta) f^{(k+1)}(\theta)\,\mathrm{d}\theta. \quad (8.2)$$

The Peano kernel theorem Let L be a linear functional such that $L(f) = 0$ for all $f \in \mathbb{P}_k[x]$. Provided that $f \in \mathrm{C}^{k+1}[a,b]$ and the above exchange of L with the integration sign is valid, the formula (8.2) is true. □

8.2 An example and few useful formulae

We approximate a derivative by a linear combination of function values:

$$f'(0) \approx -\frac{3}{2}f(0) + 2f(1) - \frac{1}{2}f(2).$$

Therefore, $L(f) := f'(0) - [-\frac{3}{2}f(0) + 2f(1) - \frac{1}{2}f(2)]$ and it is easy to check that $L(f) = 0$ for $f \in \mathbb{P}_2[x]$. (Verify by trying $f(x) = 1, x, x^2$ and using linearity of L.) Thus, for $f \in \mathrm{C}^3[0,2]$ we have

$$L(f) = \tfrac{1}{2}\int_0^2 K(\theta) f'''(\theta)\,\mathrm{d}\theta.$$

[1] Corrections and suggestions to these notes should be emailed to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/`.

To evaluate the Peano kernel K, we fix θ. Letting $g(x) := (x-\theta)_+^2$, we have

$$\begin{aligned} K(\theta) = L(g) &= g'(0) - \left[-\tfrac{3}{2}g(0) + 2g(1) - \tfrac{1}{2}g(2)\right] \\ &= 2(0-\theta)_+ - \left[-\tfrac{3}{2}(0-\theta)_+^2 + 2(1-\theta)_+^2 - \tfrac{1}{2}(2-\theta)_+^2\right] \\ &= \begin{cases} -2\theta + \frac{3}{2}\theta^2 + (2\theta - \frac{3}{2}\theta^2) \equiv 0, & \theta \le 0, \\ -2(1-\theta)^2 + \frac{1}{2}(2-\theta)^2 = 2\theta - \frac{3}{2}\theta^2, & 0 \le \theta \le 1, \\ \frac{1}{2}(2-\theta)^2, & 1 \le \theta \le 2, \\ 0, & \theta \ge 2. \end{cases} \end{aligned}$$

[Note: It is obvious that $K(\theta) = 0$ for $\theta \notin [0,2]$, since then L acts on a quadratic polynomial.] This gives the form of the Peano kernel for our example.

Back to the general case. . . Typically, forming L involves differentiation, integration and linear combination of function values. Since

$$\frac{\mathrm{d}}{\mathrm{d}x}(x-\theta)_+^k = k(x-\theta)_+^{k-1}, \qquad \int_0^x (t-\theta)_+^k \,\mathrm{d}t = \frac{1}{k+1}[(x-\theta)_+^{k+1} - (a-\theta)_+^{k+1}],$$

the exchange of L with integration is justified in these cases. Similarly for differentiation and, trivially, for linear combinations.

Theorem Suppose that K doesn't change sign in (a,b) and that $f \in \mathrm{C}^{k+1}[a,b]$. Then

$$L(f) = \frac{1}{k!}\left[\int_a^b K(\theta)\,\mathrm{d}\theta\right] f^{(k+1)}(\xi) \quad \text{for some} \quad \xi \in (a,b).$$

Proof. Let $K \ge 0$. Then

$$L(f) \ge \frac{1}{k!}\int_a^b K(\theta) \min_{x\in[a,b]} f^{(k+1)}(x)\,\mathrm{d}\theta = \frac{1}{k!}\left(\int_a^b K(\theta)\,\mathrm{d}\theta\right) \min_{x\in[a,b]} f^{(k+1)}(x).$$

Likewise $L(f) \le \frac{1}{k!}\left[\int_a^b K(\theta)\,\mathrm{d}\theta\right] \max_{x\in[a,b]} f^{(k+1)}(x)$, consequently

$$\min_{x\in[a,b]} f^{(k+1)}(x) \le \frac{L[f]}{\frac{1}{k!}\int_a^b K(\theta)\,\mathrm{d}\theta} \le \max_{x\in[a,b]} f^{(k+1)}(x)$$

and the required result follows from the mean value theorem. Similar analysis is true in the case $K \le 0$. □

Function norms: We can measure the 'size' of function g in various manners. Particular importance is afforded to the *1-norm* $\|g\|_1 = \int_a^b |f(x)|\,\mathrm{d}x$, the *2-norm* $\|g\|_2 = \left\{\int_a^b [g(x)]^2\,\mathrm{d}x\right\}^{1/2}$ and the ∞*-norm* $\|g\|_\infty = \max_{x\in[a,b]} |g(x)|$.

Back to our example We have $K \ge 0$ and $\int_0^2 K(\theta)\,\mathrm{d}\theta = \frac{2}{3}$. Consequently $L(f) = \frac{1}{2!} \times \frac{2}{3} f'''(\xi) = \frac{1}{3}f'''(\xi)$ for some $\xi \in (0,2)$. We deduce in particular that $|L(f)| \le \frac{1}{3}\|f'''\|_\infty$.

Likewise we can easily deduce from $\left|\int_a^b f(x)g(x)\,\mathrm{d}x\right| \le \|g\|_\infty \|f\|_1$ that

$$|L(f)| \le \frac{1}{k!}\|K\|_1 \|f^{(k+1)}\|_\infty \qquad \text{and} \qquad |L(f)| \le \frac{1}{k!}\|K\|_\infty \|f^{(k+1)}\|_1.$$

This is valid also when K changes sign. Moreover, the *Cauchy–Schwarz inequality*

$$\left|\int_a^b f(x)g(x)\,\mathrm{d}x\right| \le \|f\|_2 \|g\|_2$$

implies the inequality

$$|L(f)| \le \frac{1}{k!}\|K\|_2 \|f^{(k+1)}\|_2.$$

All these provide a very powerful means to bound the size of the error in our approximation procedures and verify how well 'polynomial assumptions' translate to arbitrary functions in $\mathrm{C}^{k+1}[a,b]$.

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Exercise Sheet 1

1. Calculate *all* LU factorizations of the matrix

$$A = \begin{bmatrix} 10 & 6 & -2 & 1 \\ 10 & 10 & -5 & 0 \\ -2 & 2 & -2 & 1 \\ 1 & 3 & -2 & 3 \end{bmatrix},$$

where all diagonal elements of L are one. By using one of these factorizations, find *all* solutions of the equation $A\boldsymbol{x} = \boldsymbol{b}$ where $\boldsymbol{b}^\top = [-2, 0, 2, 1]$.

2. By using column pivoting if necessary to exchange rows of A, an LU factorization of a real $n \times n$ matrix A is calculated, where L has ones on its diagonal, and where the moduli of the off-diagonal elements of L do not exceed one. Let α be the largest of the moduli of the elements of A. Prove by induction on i that elements of U satisfy the condition $|u_{ij}| \le 2^{i-1}\alpha$. Then construct 2×2 and 3×3 nonzero matrices A that yield $|u_{22}| = 2\alpha$ and $|u_{33}| = 4\alpha$ respectively.

3. Let A be a real $n \times n$ matrix that has the factorization $A = LU$, where L is lower triangular with ones on its diagonal and U is upper triangular. Prove that, for every integer $k \in \{1, 2, \ldots, n\}$, the first k rows of U span the same space as the first k rows of A. Prove also that the first k columns of A are in the k-dimensional subspace that is spanned by the first k columns of L. Hence deduce that no LU factorization of the given form exists if we have $\operatorname{rank} H_k < \operatorname{rank} B_k$, where H_k is the leading $k \times k$ submatrix of A and where B_k is the $n \times k$ matrix whose columns are the first k columns of A.

4. Calculate the Cholesky factorization of the matrix

$$\begin{bmatrix} 1 & 1 & & & & \\ 1 & 2 & 1 & & & \\ & 1 & 3 & 1 & & \\ & & 1 & 4 & 1 & \\ & & & 1 & 5 & 1 \\ & & & & 1 & \lambda \end{bmatrix}.$$

Deduce from the factorization the value of λ that makes the matrix singular. Also find this value of λ by seeking the vector in the null-space of the matrix whose first component is one.

5. Let A be an $n \times n$ nonsingular band matrix that satisfies the condition $a_{ij} = 0$ if $|i-j| > r$, where r is small, and let Gaussian elimination *with column pivoting* be used to solve $Ax = b$. Identify all the coefficients of the intermediate equations that can become nonzero. Hence deduce that the total number of additions and multiplications of the complete calculation can be bounded by a constant multiple of nr^2.

6. The iteration $\boldsymbol{x}^{k+1} = H\boldsymbol{x}^k + b$ is applied for $k = 0, 1, \ldots$, where H is the real 2×2 matrix

$$H = \begin{bmatrix} \alpha & \gamma \\ 0 & \beta \end{bmatrix},$$

with γ large and $|\alpha| < 1$, $|\beta| < 1$. Calculate the elements of H^k and show that they tend to zero as $k \to \infty$. Further, establish the equation $\boldsymbol{x}^k - \boldsymbol{x}^* = H^k(\boldsymbol{x}^0 - \boldsymbol{x}^*)$, where $\boldsymbol{x}^*$ is defined by $\boldsymbol{x}^* = H\boldsymbol{x}^* + \boldsymbol{b}$. Thus deduce that the sequence $(\boldsymbol{x}^k)_{k=0}^{\infty}$ converges to $\boldsymbol{x}^*$.

7. For some choice of $\boldsymbol{x}^0$ the iterative method

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{x}^{k+1} + \begin{bmatrix} 0 & 0 & 0 \\ \xi & 0 & 0 \\ \eta & \zeta & 0 \end{bmatrix} \boldsymbol{x}^k = \boldsymbol{b}$$

is applied for $k = 0, 1, \ldots$, in order to solve the linear system

$$\begin{bmatrix} 1 & 1 & 1 \\ \xi & 1 & 1 \\ \eta & \zeta & 1 \end{bmatrix} \boldsymbol{x} = \boldsymbol{b},$$

where ξ, η and ζ are constants. Find all values of the constants such that the sequence $(x^k)_{k=0}^{\infty}$ converges for every x^0 and b. Give an example of nonconvergence when $\xi = \eta = \zeta = -1$. Is the solution always found in at most two iterations when $\xi = \zeta = 0$?

8. Let a_1, a_2 and a_3 denote the columns of the matrix

$$A = \begin{bmatrix} 6 & 6 & 1 \\ 3 & 6 & 1 \\ 2 & 1 & 1 \end{bmatrix}.$$

Apply the Gram–Schmidt procedure to A, which generates orthonormal vectors $\boldsymbol{q}_1, \boldsymbol{q}_2$ and $\boldsymbol{q}_3$. Note that this calculation provides real numbers r_{jk} such that $\boldsymbol{a}_k = \sum_{j=1}^{k} r_{jk} \boldsymbol{q}_j$, $k = 1, 2, 3$. Hence express A as the product $A = QR$, where Q and R are orthogonal and upper-triangular matrices respectively.

9. Calculate the QR factorization of the matrix of Exercise 8 by using three Givens rotations. Explain why the initial rotation can be any one of the three types $\Omega^{(1,2)}$, $\Omega^{(1,3)}$ and $\Omega^{(2,3)}$. Prove that the final factorization is independent of this initial choice in exact arithmetic, provided that we satisfy the condition that in each row of R the leading nonzero element is positive.

10. Let A be an $n \times n$ matrix, and for $i = 1, 2, \ldots, n$ let $k(i)$ be the number of zero elements in the i-th row of A that come before all nonzero elements in this row and before the diagonal element a_{ii}. Show that the QR factorization of A can be calculated by using at most $\frac{1}{2}n(n-1) - \sum k(i)$ Givens rotations. Hence show that, if A is an upper triangular matrix except that there are nonzero elements in its first column, i.e. $a_{ij} = 0$ when $2 \leq j < i \leq n$, then its QR factorization can be calculated by using only $2n - 3$ Givens rotations. [*Hint*: Your should find the order of the first $(n-2)$ rotations that brings your matrix to the form considered above.]

11. Calculate the QR factorization of the matrix of Exercise 8 by using two Householder reflections. Show that, if this technique is used to generate the QR factorization of a general $n \times n$ matrix A, then the computation can be organised so that the total number of additions and multiplications is bounded above by a constant multiple of n^3.

12. Let

$$A = \begin{bmatrix} 3 & 4 & 7 & -2 \\ 5 & 4 & 9 & 3 \\ 1 & -1 & 0 & 3 \\ 1 & -1 & 0 & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} 11 \\ 29 \\ 16 \\ 10 \end{bmatrix}.$$

Calculate the QR factorization of A by using Householder reflections. In this case A is singular and you should choose Q so that the last row of R is zero. Hence identify all the least squares solutions of the inconsistent system $A\boldsymbol{x} = \boldsymbol{b}$, where we require $\boldsymbol{x}$ to minimize $\|A\boldsymbol{x} - \boldsymbol{b}\|_2$. Verify that all the solutions give the same vector of residuals $A\boldsymbol{x} - \boldsymbol{b}$, and that this vector is orthogonal to the columns of A. There is no need to calculate the elements of Q explicitly.

Mathematical Tripos Part IB: Easter 2006

Numerical Analysis – Exercise Sheet 2[1]

13. Suppose that the function values $f(0), f(1), f(2)$ and $f(3)$ are given and that we wish to estimate

$$f(6), \quad f'(0) \quad \text{and} \quad \int_0^3 f(x)\,\mathrm{d}x.$$

One method is to let p be the cubic polynomial that interpolates these function values, and then to employ the approximants

$$p(6), \quad p'(0) \quad \text{and} \quad \int_0^3 p(x)\,\mathrm{d}x$$

respectively. Deduce from the Lagrange formula for p that each approximant is a linear combination of the four data with constant coefficients. Calculate the numerical values of these constants. Verify your work by showing that the approximants are exact when f is an arbitrary cubic polynomial.

14. Let f be a function in $C^4[0,1]$ and let p be a cubic polynomial that interpolates $f(0), f'(0), f(1)$ and $f'(1)$. Deduce from the Rolle theorem that for every $x \in [0,1]$ there exists $\xi \in [0,1]$ such that the equation

$$f(x) - p(x) = \tfrac{1}{24} x^2 (x-1)^2 f^{(4)}(\xi)$$

is satisfied.

15. Let a, b and c be distinct real numbers (not necessarily in ascending order), and let $f(a), f(b), f'(a), f'(b)$ and $f'(c)$ be given. Because there are five data, one might try to approximate f by a polynomial of degree at most four that interpolates the data. Prove by a general argument that this interpolation problem has a solution and the solution is unique if and only if there is no nonzero polynomial $p \in \mathbb{P}_4[x]$ that satisfies $p(a) = p(b) = p'(a) = p'(b) = p'(c) = 0$. Hence, given a and b, show that there exists a unique value of $c \neq a, b$ such that there is no unique solution.

16. Let $f : \mathbb{R} \to \mathbb{R}$ be a given function and let p be the polynomial of degree at most n that interpolates f at the pairwise distinct points $x_0, x_1, \ldots, x_n$. Further, let x be any real number that is not an interpolation point. Deduce the identity

$$f(x) - p(x) = f[x_0, x_1, \ldots, x_n, x] \prod_{j=0}^{n} (x - x_j)$$

from the definition of the divided difference $f[x_0, x_1, \ldots, x_n, x]$.

[1] Corrections and suggestions to these notes should be emailed to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/`.

17. Simulating a computer that works to only four decimal places, form the table of divided differences of the values $f(0) = 0$, $f(0.1) = 0.0998$, $f(0.4) = 0.3894$ and $f(0.7) = 0.6442$ of $\sin x$. Hence identify the polynomial that is given by Newton's interpolation method. Due to rounding errors, this polynomial should differ from the one that would be given by exact arithmetic. Take the view, however, that the *computed* values of $f[0.0, 0.1]$, $f[0.0, 0.1, 0.4]$ and $f[0.0, 0.1, 0.4, 0.7]$ and the function value $f(0)$ are correct. Then, by working backwards through the difference table, identify the values of $f(0), f(0.1), f(0.4)$ and $f(0.7)$ that would give these divided differences in exact arithmetic.

18. Set $f(x) = 2x - 1$, $x \in [0, 1]$. We require a function of form

$$p(x) = \sum_{k=0}^{n} a_k \cos(k\pi x), \quad 0 \le x \le 1,$$

that satisfies the condition

$$\int_0^1 [f(x) - p(x)]^2 \, \mathrm{d}x < 10^{-4}.$$

Explain why it is sufficient if the value of $a_0^2 + \frac{1}{2}\sum_{k=1}^{n} a_k^2$ exceeds $\frac{1}{3} - 10^{-4}$, where the coefficients $\{a_k\}_{k=0}^{n}$ are calculated to minimize this integral. Hence find the smallest acceptable value of n.

19. The polynomials $\{p_n\}_{n \in \mathbb{Z}^+}$ are defined by the three-term recurrence formula

$$\begin{aligned} p_0(x) &\equiv 1, \\ p_1(x) &= 2x, \\ p_{n+1}(x) &= 2xp_n(x) - p_{n-1}(x), \qquad n = 1, 2, \ldots. \end{aligned}$$

Prove that they are orthogonal with respect to the inner product

$$\langle f, g \rangle = \int_{-1}^{1} f(x)g(x)\sqrt{1 - x^2} \, \mathrm{d}x$$

and evaluate $\langle p_n, p_n \rangle$ for $n \in \mathbb{Z}^+$. *[Hint: Prove that* $p_n(x) = \sin(n+1)\theta / \sin\theta$*, where* $x = \cos\theta$*.]*

20. Calculate the coefficients b_1, b_2, c_1 and c_2 so that the approximant

$$\int_0^1 f(x) \, \mathrm{d}x \approx b_1 f(c_1) + b_2 f(c_2)$$

is exact when f is a cubic polynomial. You may exploit the fact that c_1 and c_2 are the zeros of a quadratic polynomial that is orthogonal to all linear polynomials. Verify your calculation by testing the formula when $f(x) = 1, x, x^2$ and x^3.

21. The functions $p_0, p_1, p_2, \ldots$ are generated by the Rodrigues formula

$$p_n(x) = \mathrm{e}^x \frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^n \mathrm{e}^{-x}), \qquad 0 \le x < \infty.$$

Show that these functions are polynomials and prove by integration by parts that for every $p \in \mathbb{P}_{n-1}[x]$ we have the orthogonality condition $\langle p_n, p \rangle = 0$ with respect to the scalar product

$$\langle f, g \rangle := \int_0^\infty \mathrm{e}^{-x} f(x) g(x) \, \mathrm{d}x.$$

Derive the coefficients of p_3, p_4 and p_5 from the Rodrigues formula. Verify that these coefficients are compatible with a three term recurrence relation of the form

$$p_5(x) = (\gamma x - \alpha) p_4(x) - \beta p_3(x), \qquad x \in \mathbb{R},$$

where α, β and γ are constants.

22. Let $p(\frac{1}{2}) = \frac{1}{2}(f(0) + f(1))$, where f is a function in $C^2[0,1]$. Find the least constants c_0, c_1 and c_2 such that the error bounds

$$|f(\tfrac{1}{2}) - p(\tfrac{1}{2})| \le c_k \|f^{(k)}\|_\infty \,, \qquad k = 0, 1, 2,$$

are valid. *[Note: The cases $k = 0$ and $k = 1$ are easy if one works from first principles, and the Peano kernel theorem is suitable when $k = 2$. Also try the Peano kernel theorem when $k = 1$.]*

23. Express the divided difference $f[0,1,2,4]$ in the form

$$f[0,1,2,4] = \int_0^4 K(\theta) f'''(\theta) \, \mathrm{d}\theta,$$

assuming that f''' exists and is continuous. Sketch the kernel function $K(\theta)$ for $0 \le \theta \le 4$. By integrating $K(\theta)$ analytically and using the mean value theorem prove that

$$f[0,1,2,4] = \tfrac{1}{6} f'''(\xi)$$

for some point $\xi \in [0,4]$. Note that another proof of this result was given in the lecture on divided differences.

24. Let f be a function in $C^4[0,1]$ and let ξ be any fixed point in $[0,1]$. Calculate the coefficients α, β, γ and δ such that the approximant

$$f'''(\xi) \approx \alpha f(0) + \beta f(1) + \gamma f'(0) + \delta f'(1)$$

is exact for all cubic polynomials. Prove that the inequality

$$|f'''(\xi) - \alpha f(0) - \beta f(1) - \gamma f'(0) - \delta f'(1)| \le \left\{ \tfrac{1}{2} - \xi + 2\xi^3 - \xi^4 \right\} \|f^{(4)}\|_\infty$$

is satisfied. Show that this inequality holds as an equation if we allow f to be the function

$$f(x) = \begin{cases} -(x-\xi)^4, & 0 \leq x \leq \xi, \\ (x-\xi)^4, & \xi \leq x \leq 1. \end{cases}$$

25. *[Not easy!]* Given f and g in $C[a,b]$, let $h := fg$. Prove by induction that the divided differences of h satisfy the equation

$$h[x_0, x_1, \ldots, x_n] = \sum_{j=0}^{n} f[x_0, x_1, \ldots, x_j] g[x_j, x_{j+1}, \ldots, x_n].$$

By expressing the differences in terms of derivatives and by letting the points $x_0, x_1, \ldots, x_n$ become coincident, deduce the Leibniz formula for the nth derivative of a product of two functions.