A. C. Faul

Solutions to Odd-Numbered Exercises for A Concise Introduction to Numerical Analysis

Contents

CHAPTER 1

Fundamentals Exercises

1.1 EXERCISE 1.1

Write a C*-routine which implements the calculation of* $\sqrt{x^2+y^2}$ *in a way which avoids overflow. Consider the cases where* x *and* y *differ largely in magnitude.*

Solution

This calculation is finding the Euclidean norm of a vector (x, y). For a vector with n components the best procedure to calculate the norm is to first find the component with largest absolute value and divide each component by this value. Then we calculate the sum of squares in ascending order starting with the smallest component. This way it is less likely that small components are lost. Taking then the square root and multiplying with the maximum absolute value of the components gives the norm. If the components differ too much in magnitude the smaller will not contribute.

1.2 EXERCISE 1.3

Let A *be an* $n \times n$ *nonsingular band matrix that satisfies the condition* $A_{ij} = 0$ *for* $|i - j| > r$*, where* r *is small, and let Gaussian elimination (introduced in Linear Systems 2.2) be used to solve* $Ax = b$*. Deduce that the total number of additions and multiplications of the complete calculation can be bounded by a constant multiple of* nr^2*.*

Solution

Gaussian elimination is equivalent to LU-factorization where L is lower triangular and U is upper triangular. Let $\mathbf{l}_1, \ldots, \mathbf{l}_n$ be the columns of L and

$\mathbf{u}_1^T, \ldots, \mathbf{u}_n^T$ the rows of U. Then

$$A = LU = (\mathbf{l}_1 \cdots \mathbf{l}_n) \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} = \sum_{k=1}^{n} \mathbf{l}_k \mathbf{u}_k^T.$$

Since the first $k-1$ components of $\mathbf{l}_k$ and $\mathbf{u}_k^T$ are all zero, each of the matrices $\mathbf{l}_k \mathbf{u}_k^T$ has zeros in its first k rows and columns. L is normalized so that all its diagonal elements are 1.

The algorithm starts with letting $\mathbf{u}_1^T$ be the first row of $A = A_0$ and $\mathbf{l}_1$ the first column of A_0 divided by A_{11}. We then form the matrix

$$A_1 = A_0 - \mathbf{l}_1 \mathbf{u}_1^T.$$

The first row and column of A_1 are zero and hence $\mathbf{u}_2^T$ is the second row of A_1, while $\mathbf{l}_2$ is its second column scaled such that $L_{22} = 1$ and so forth.

For general k, $\mathbf{u}_k^T$ is the k-th row of A_{k-1} and $\mathbf{l}_k$ is the k-th column of A_{k-1} scaled such that $L_{kk} = 1$. Next we calculate $A_k = A_{k-1} - \mathbf{l}_k \mathbf{u}_k^T$.

Now if A is a banded matrix then so are U and L which can be see from the construction above. Moreover $\mathbf{u}_k^T$ is the k-th row of $A - \sum_{i=1}^{k-1} \mathbf{l}_i \mathbf{u}_i^T$ which written in components is

$$U_{kj} = A_{kj} - \sum_{i=1}^{k-1} L_{ki} U_{ij}$$

Now L_{ki} is zero for $i < k - r$, which means that the sum only runs from $\max(1, k-r)$ to $k-1$ and thus there are only at most r multiplications and additions. Similarly, U_{kj} is zero for $j > k + r$ and thus we calculate U_{kj} only for $j \leq \min(n, k+r)$ which is at most r values. Since k runs from 1 to n we have at most nr^2 operations to calculate U. Similarly for L. Solving the two banded triangular systems involves $2nr$ operations.

1.3 EXERCISE 1.5

Examine the condition of the evaluating $\cos x$.

Solution

The condition number K is defined as

$$K(x) = \left| \frac{x f'(x)}{f(x)} \right|$$

For $f(x) = \cos x$ this becomes

$$K(x) = \left| \frac{x \sin x}{\cos x} \right| = |x \tan x|.$$

This becomes arbitrarily large for multiples of $\frac{\pi}{2}$ which means that the relative accuracy of $\cos x$ is arbitrarily inaccurate whenever its value is close to zero.

1.4 EXERCISE 1.7

Let

$$A = \begin{pmatrix} 1000 & 999 \\ 999 & 998 \end{pmatrix}$$

Calculate A^{-1}*, the eigenvalues and eigenvectors of* A*,* $K_2(A)$ *and* $K_\infty(A)$*. What is special about the vectors* $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ *and* $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$*?*

Solution

The inverse of the matrix is

$$A^{-1} = \begin{pmatrix} -998 & 999 \\ 999 & -1000 \end{pmatrix}$$

The characteristic polynomial is

$$(1000 - \lambda)(998 - \lambda) - 999^2 = \lambda^2 - 1998\lambda - 1$$

with roots $999 \pm \sqrt{999^2 + 1} = 999 \pm 999\sqrt{1 + \frac{1}{999^2}}$. Thus one eigenvalue is close to zero, while the other is close to 1998. The corresponding eigenvectors are

$$\begin{pmatrix} \frac{1}{999} \pm \sqrt{1 + \frac{1}{999^2}} \\ 1 \end{pmatrix} \approx \begin{pmatrix} -0.9989995 \\ 1 \end{pmatrix}, \begin{pmatrix} 1.0010015 \\ 1 \end{pmatrix}.$$

The condition number relating to the 2-norm is

$$K_2(A) = \frac{1 + \sqrt{1 + \frac{1}{999^2}}}{1 - \sqrt{1 + \frac{1}{999^2}}} \approx 3992006.$$

The ∞-norm of A and A^{-1} is

$$\|A\|_\infty = \|A^{-1}\|_\infty = 1999$$

and thus $K_\infty(A) = 1999^2 = 3996001$. Using the ∞-norm, the magnification factor is $\|Ax\|_\infty / \|x\|_\infty = \|A\|_\infty$ and

$$\begin{pmatrix} 1000 & 999 \\ 999 & 998 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1999 \\ 1997 \end{pmatrix}.$$

Thus $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is a direction of maximum magnification. Equivalently, $\begin{pmatrix} 1999 \\ 1997 \end{pmatrix}$ is a direction of minimum magnification by A^{-1}. On the other hand,

$$\begin{pmatrix} -998 & 999 \\ 999 & -1000 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1997 \\ -1999 \end{pmatrix}.$$

Thus $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ is a direction of maximum magnification by A^{-1} and $\begin{pmatrix} 1997 \\ -1999 \end{pmatrix}$ is a direction of minimum magnification by A.

1.5 EXERCISE 1.9

(a) Define absolute error, relative error and state their relationship.

(b) Show how the relative error builds up in multiplication and division.

(c) Explain forward and backward error analysis using the example of approximating

$$\cos x \approx f(x) = 1 - x^2/2.$$

(d) Considering the binary floating point representation of numbers, explain the concept of the hidden bit.

(e) Explain the biased representation of the exponent in binary floating point representation.

(f) How are 0, ∞ *and NaN represented?*

(g) How are the numbers 2^k *for positive and negative* k *represented?*

Solution

(a) Let x be a real numbers not dangerously close to overflow or underflow and let x^* denote the floating-point representation of x. The absolute error ϵ is

$$x^* = x + \epsilon$$

and the relative error δ is

$$x^* = x(1 + \delta) = x + x\delta.$$

Thus

$$\epsilon = x\delta \quad \text{or, if} \quad x \neq 0, \quad \delta = \frac{\epsilon}{x}.$$

(b) Let

$$\begin{aligned} x_1^* &= x_1(1 + \delta_1) \\ x_2^* &= x_2(1 + \delta_2) \end{aligned}$$

Then

$$\begin{aligned} x_1^* \times x_2^* &= x_1 x_2(1 + \delta_1)(1 + \delta_2) \\ &= x_1 x_2(1 + \delta_1 + \delta_2 + \delta_1\delta_2) \end{aligned}$$

The term $\delta_1\delta_2$ can be neglected, since it is small. The worst case is, when δ_1 and δ_2 have the same sign, i.e. the relative error in $x_1^* \times x_2^*$ is no worse than $|\delta_1| + |\delta_2|$.

Division can be easily analyzed in the same way by using the binomial expansion to write

$$\frac{1}{x_2^*} = \frac{1}{x_2}(1+\delta_2)^{-1} = \frac{1}{x_2}(1-\delta_2+\ldots).$$

Then

$$\begin{aligned} x_1^*/x_2^* &= x_1/x_2(1+\delta_1)(1-\delta_2+\ldots) \\ &= x_1/x_2(1+\delta_1-\delta_2-\delta_1\delta_2+\ldots) \end{aligned}$$

The worst case is, when δ_1 and δ_2 have the opposite sign. Again, the relative error in x_1^*/x_2^* is no worse than $|\delta_1|+|\delta_2|$.

(c) Forward error analysis examines how perturbations of the input propagate. Backward error analysis examines the question: How much error in input would be required to explain all output error? It assumes that an approximate solution to a problem is good if it is the exact solution to a nearby problem. For the example $\cos x \approx f(x) = 1 - x^2/2$, the forward error is simply $f(x) - \cos x$. For the backward error we need to find x^* such that $\cos x^* = f(x)$. In particular,

$$x^* = \arccos f(x).$$

(d) If the floating point number is normalized, that is the leading digit is nonzero, then it has to be 1, when binary representation is used. It therefore does not need to be stored, as long as a special representation for 0 is used.

(e) The representation of the exponent uses a biased representation. In the case of single precision, where the exponent is stored in 8 bits, the bias is 127 (for double precision it is 1023). What this means is if k is the value of the exponent bits interpreted as an unsigned integer, then the exponent of the floating point number is $k-127$. This is called the unbiased exponent to distinguish form the biased exponent k.

(f) The unbiased exponents range between $e_{\min}-1 = -127$ and $e_{\max}+1 = 128$ in single precision. The numbers $e_{\min} - 1 = -127$ and $e_{\max} + 1 = 128$ are used to encode special quantities. More precisely 0 is encoded with the exponent being $e_{\min} - 1$ and the significand being entirely zero. ∞ is encoded with the exponent being $e_{\max} + 1$ and the significand being entirely zero. NaN is encoded with the exponent being $e_{\max} + 1$ and the significand being nonzero.

(g) Since the numbers 2^k are powers of two, the significand has to have the value 1. However, this is the hidden bit and thus the stored significand is entirely zero. The unbiased exponent k has to be stored in biased form which is $k + 127$ in single precision.

1.6 EXERCISE 1.11

(a) Define absolute error, relative error and state their relationship.

(b) Explain absolute error test and relative error test and give examples of circumstances when they are unsuitable. What is a mixed error test?

(c) Explain loss of significance.

(d) Let $x_1 = 3.0001$ be the true value approximated by $x_1^* = 3.0001 + 10^{-5}$ and $x_2 = -3.0000$ be the true value approximated by $x_2^* = -3.0000 + 10^{-5}$. State the absolute and relative errors in x_1^* and x_2^*. Calculate the absolute error and relative error in approximating $x_1 + x_2$ by $x_1^* + x_2^*$. How many times bigger is the relative error in the sum compared to the relative error in x_1^* and x_2^*?

(e) Let

$$f(x) = x - \sqrt{x^2 + 1}, \qquad x \geq 0. \tag{1.1}$$

Explain when and why loss of significance occurs in the evaluation of f.

(f) Derive an alternative formula for evaluating f which avoids loss of significance.

(g) Test your alternative by considering a decimal precision $p = 16$ and $x = 10^8$. What answer does your alternative formula give compared to the original formula?

(h) Explain condition and condition number in general terms.

(i) Derive the condition number for evaluating a differentiable function f at a point x, i.e. calculating $f(x)$.

(j) Considering $f(x)$ as defined in (1.1), find the smallest interval in which the condition number lies. Is the problem well-conditioned or ill-conditioned?

Solution

(a) Let x be a real numbers not dangerously close to overflow or underflow and let x^* denote the floating-point representation of x. The absolute error ϵ is

$$x^* = x + \epsilon$$

and the relative error δ is

$$x^* = x(1 + \delta) = x + x\delta.$$

Thus

$$\epsilon = x\delta \quad \text{or, if} \quad x \neq 0, \quad \delta = \frac{\epsilon}{x}.$$

(b) There are two forms of error testing, one using a target absolute accuracy ϵ_t, the other using a target relative error δ_t. In the first case the calculation is terminated when

$$|\epsilon_n| \leq \epsilon_t, \tag{1.2}$$

where ϵ_n denotes the absolute error in the n-th approximation. In the second case the calculation is terminated when

$$|\epsilon_n| \leq \delta_t |x_n|. \tag{1.3}$$

Both methods are flawed under certain circumstances. If x is large, say 10^{20}, and $u = 10^{-16}$, then ϵ_n is never likely to be much less than 10^4, so condition (1.2) is unlikely to be satisfied if ϵ_t is chosen too small even when the process converges. On the other hand, if $|x_n|$ is very small, then $\delta_t|x_n|$ may underflow and test (1.3) may never be satisfied (unless the error becomes exactly zero).

As (1.2) is useful when (1.3) is not, and vice versa, so-called mixed error tests have been developed. In the simplest form of such a test, a target error η_t is prescribed and the calculation is terminated when the condition

$$|\epsilon_n| \leq \eta_t(1 + |x_n|)$$

is satisfied. If $|x_n|$ is small η_t may be thought of as target absolute error, or if $|x_n|$ is large η_t may be thought of as target relative error.

(c) Loss of significance occurs whenever two similar numbers of equal sign are subtracted (or two similar numbers of opposite sign are added), and is a major cause of inaccuracy in floating-point algorithms. The relative error increases, since the result of the calculation is small.

(d) For $x_1 = 3.0001$ with $x_1^* = 3.0001 + 10^{-5}$ the absolute error is $\epsilon_1 = 10^{-5}$ and the relative error is $\delta_1 = 10^{-5}/3.0001$. For $x_2 = -3.0000$ with $x_2^* = -3.0000 + 10^{-5}$ the absolute error is $\epsilon_2 = 10^{-5}$ and the relative error is $\delta_2 = -10^{-5}/3.0000$. For the sum we have

$$\begin{aligned} x_1 + x_2 &= 0.0001 \\ x_1^* + x_2^* &= 0.0001 + 2 * 10^{-5}. \end{aligned}$$

Thus the absolute error in approximating $x_1 + x_2$ by $x_1^* + x_2^*$ is $x_1 + x_2 - (x_1^* + x_2^*) = -2 * 10^{-5}$ and the relative error is $-2 * 10^{-5}/0.0001 = -2 * 10^{-1}$ which is bigger than the relative error in x_1 and x_2 by a factor of 10^4.

(e) For

$$f(x) = x - \sqrt{x^2 + 1}$$

loss of significance occurs when x becomes large enough such that the representation of $x^2 + 1$ equals x^2, that is $(x^2 + 1)^* = x^2$. In this case $f(x)$ evaluates to 0.

(f) An alternative way of evaluating f is

$$\begin{aligned} f(x) &= x - \sqrt{x^2+1} \times \frac{x+\sqrt{x^2+1}}{x+\sqrt{x^2+1}} \\ &= \frac{x^2-(x^2+1)}{x+\sqrt{x^2+1}} = \frac{-1}{x+\sqrt{x^2+1}} \end{aligned}$$

(g) For decimal precision $p = 16$ $(10^8)^2 + 1$ is represented as 10^{16}. Thus in its original formulation $f(10^8)$ is zero. In the alternative formula we have

$$f(10^8) = \frac{-1}{10^8 + \sqrt{(10^8)^2+1}} \approx -0.5 * 10^{-8}.$$

(h) The condition of a problem is a qualitative or quantitative statement about how easy it is to solve irrespective of the algorithm used to solve it. The condition number of a numerical problem measures the asymptotically worst case of how much the outcome can change in proportion to small perturbations in the input data. A problem with a low condition number is said to be well-conditioned, while a problem with a high condition number is said to be ill-conditioned.

(i) For the problem of evaluating a differentiable function f at a point x, i.e. calculating $f(x)$, let $\hat{x}$ be a point close to x. The condition number K is defined as the relative change in $f(x)$ caused by a unit relative change in x.

$$\begin{aligned} K(x) &= \lim_{\hat{x}\to x} \left| \frac{f(x)-f(\hat{x})}{f(x)} \right| \left| \frac{x}{x-\hat{x}} \right| \\ &= \left| \frac{x}{f(x)} \right| \lim_{\hat{x}\to x} \left| \frac{f(x)-f(\hat{x})}{x-\hat{x}} \right| \\ &= \left| \frac{xf'(x)}{f(x)} \right| \end{aligned}$$

(j) For $f(x) = x - \sqrt{x^2+1}$ we have

$$\begin{aligned} K(x) &= \left| \frac{xf'(x)}{f(x)} \right| = \left| \frac{x(1-x(x^2+1)^{-1/2})}{x-\sqrt{x^2+1}} \right| \\ &= \left| \frac{\frac{x}{\sqrt{x^2+1}} \left(\sqrt{x^2+1}-x\right)}{x-\sqrt{x^2+1}} \right| = \frac{x}{\sqrt{x^2+1}}, \end{aligned}$$

since $x \geq 0$. Since the denominator $\sqrt{x^2+1} \geq x$, $K(x)$ lies in the interval $[0,1]$ for all x. Thus the problem is well-conditioned.

CHAPTER 2

Linear Systems Exercises

2.1 EXERCISE 2.1

Implement backward substitution.

Solution

```
function [x]=Backward(A,b)
% Solves the upper triangular system of equations Ax = b
% A input argument, square upper triangular matrix
% b input argument
% x solution

[n,m]=size(A); % finding the size of A
if n≠ m
    disp('input is not a square matrix');
    return;
end
if size(b,1) ≠ n
    disp('input dimensions do not match');
    return;
end
x = zeros(n,1); % initialise x to the same dimension
if abs(A(n,n)) > 1e—12 % not comparing to zero because of possible
                       % rounding errors
    x(n) = b(n)/A(n,n); % solve for the last element of x
else
    disp('input singular'); % A is singular if any of the diagonal
                            % elements are zero
    return;
end
for k=n:—1:1 % the loop considers one row after the other backwards
    if abs(A(k,k))>1e—12 % not comparing to zero because of possible
                         % rounding errors
```

```
        temp = 0;
        for j=n:-1:k+1
            temp = temp + A(k,j) * x(j); % Multiply the elements of
                                         % the k-th row of A after the
                                         % diagonal by theelements of x
                                         % already calculated
        end
        x(k) = (b(k)-temp)/A(k,k); % solve for the k-th element of x
    else
        disp('input singular'); % A is singular if any of the diagonal
                                % elements are zero
        return;
    end
end
```

2.2 EXERCISE 2.3

By using pivoting if necessary an LU factorization is calculated of an $n \times n$ matrix A, where L has ones on the diagonal and the moduli of all off-diagonal elements do not exceed 1. Let α be the largest moduli of the elements of A. Prove by induction that elements of U satisfy $|U_{i,j}| \leq 2^{i-1}\alpha$. Construct 2×2 and 3×3 nonzero matrices A that give $|U_{2,2}| = 2\alpha$ and $|U_{3,3}| = 4\alpha$.

Solution

For $i = 1$ we have $|U_{1,j}| = |A_{1,j}| \leq \alpha$. Assume the assertion is true for all $i \leq k-1$. Now $\mathbf{u}_k^T$ is the k-th row of $A - \sum_{i=1}^{k-1} \mathbf{l}_i \mathbf{u}_i^T$ or written as vectors

$$\mathbf{u}_k = \begin{pmatrix} A_{k,1} \\ \vdots \\ A_{k,n} \end{pmatrix} - \sum_{i=1}^{k-1} L_{k,i}\mathbf{u}_i.$$

For each element of $\mathbf{u}_k$ we can write

$$|U_{k,j}| = |A_{k,j} - \sum_{i=1}^{k-1} L_{k,i}U_{i,j}| \leq |A_{k,j}| + \sum_{i=1}^{k-1} 2^{i-1}\alpha \leq (1 + \frac{2^{k-1}-1}{2-1})\alpha = 2^{k-1}\alpha$$

Examples for $n = 2$ and $n = 3$ are

$$\begin{pmatrix} \alpha & \alpha \\ -\alpha & \alpha \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \alpha & \alpha \\ 0 & 2\alpha \end{pmatrix}$$

and

$$\begin{pmatrix} \alpha & \frac{1}{2}\alpha & \alpha \\ \alpha & 0 & -\alpha \\ \alpha & \alpha & -\alpha \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \alpha & \frac{1}{2}\alpha & \alpha \\ 0 & -\frac{1}{2}\alpha & -2\alpha \\ 0 & 0 & -4\alpha \end{pmatrix}$$

2.3 EXERCISE 2.5

Let A be a real nonsingular $n \times n$ matrix that has the factorization $A = LU$, where L is lower triangular with ones on its diagonal and U is upper triangular. Show that for $k = 1, \ldots, n$ the first k rows of U span the same subspace as the first k rows of A. Show also that the first k columns of A are in the k-dimensional subspace spanned by the first k columns of L.

Solution

Since $A_0 = A$ and $\mathbf{u}_1^T$ is the first row of A_0, the first row of U and A are identical and span the same space. We continue with induction on k. Since $\mathbf{u}_k^T$ is the k-th row of $A_{k-1} = A - \sum_{i=1}^{k-1} \mathbf{l}_i \mathbf{u}_i^T$, it is the k-th row of A minus $\sum_{i=1}^{k-1} L_{k,i} \mathbf{u}_i^T$. We see immediately that the k-th row of A is a linear combination of the first k rows of U. Furthermore, the sum lies in the space spanned by the first $k-1$ rows of U which is the same as the space spanned by the first $k-1$ rows of A. Thus $\mathbf{u}_k^T$ is a linear combination of the first k rows of A.

The argument for the columns of L and A is very similar apart from the scaling. We do not have the case where there happens to be a zero at the (k, k) entry, since A is nonsingular.

2.4 EXERCISE 2.7

Let $\mathbf{a}_1, \mathbf{a}_2$ and $\mathbf{a}_3$ denote the columns of the matrix

$$A = \begin{pmatrix} 3 & 6 & -1 \\ -6 & -6 & 1 \\ 2 & 1 & -1 \end{pmatrix}$$

Using the Gram-Schmidt procedure generate orthonormal vectors $\mathbf{q}_1, \mathbf{q}_2$ and $\mathbf{q}_3$ and real numbers $R_{i,j}$ such that $\mathbf{a}_i = \sum_{j=1}^{i} R_{i,j} \mathbf{q}_j$, $i = 1, 2, 3$. Thus express A as the product $A = QR$, where Q is orthogonal and R is upper triangular.

Solution

Firstly,

$$\mathbf{a}_1 = \begin{pmatrix} 3 \\ -6 \\ 2 \end{pmatrix}, \qquad \mathbf{a}_2 = \begin{pmatrix} 6 \\ -6 \\ 1 \end{pmatrix}, \qquad \mathbf{a}_3 = \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}.$$

The length of $\mathbf{a}_1$ is $\|\mathbf{a}_1\| = \sqrt{9 + 36 + 4} = 7$ and thus $\mathbf{q}_1 = \frac{1}{7}\mathbf{a}_1$ and $R_{1,1} = 7$. The scalar product of $\mathbf{q}_1$ with $\mathbf{a}_2$ is $\frac{1}{7}(18 + 36 + 2) = 8 = R_{1,2}$. Hence

$$\mathbf{w} = \mathbf{a}_2 - \langle \mathbf{q}_1, \mathbf{a}_2 \rangle \mathbf{q}_1 = \begin{pmatrix} 6 \\ -6 \\ 1 \end{pmatrix} - \frac{8}{7} \begin{pmatrix} 3 \\ -6 \\ 2 \end{pmatrix} = \frac{3}{7} \begin{pmatrix} 6 \\ 2 \\ -3 \end{pmatrix}$$

The length of $\mathbf{w}$ is $\|\mathbf{w}\| = \frac{3}{7}\sqrt{36+4+9} = 3 = R_{2,2}$ and

$$\mathbf{q}_2 = \frac{1}{7}\begin{pmatrix} 6 \\ 2 \\ -3 \end{pmatrix}$$

We have

$$\begin{aligned} \langle \mathbf{q}_1, \mathbf{a}_3 \rangle &= \tfrac{1}{7}(-3-6-2) = -\tfrac{11}{7} = R_{1,3} \\ \langle \mathbf{q}_2, \mathbf{a}_3 \rangle &= \tfrac{1}{7}(-6+2+3) = -\tfrac{1}{7} = R_{2,3} \end{aligned}$$

Now

$$\begin{aligned} \mathbf{w} &= \mathbf{a}_3 - \langle \mathbf{q}_1, \mathbf{a}_3 \rangle \mathbf{q}_1 - \langle \mathbf{q}_2, \mathbf{a}_3 \rangle \mathbf{q}_2 = \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix} + \frac{11}{49}\begin{pmatrix} 3 \\ -6 \\ 2 \end{pmatrix} + \frac{1}{49}\begin{pmatrix} 6 \\ 2 \\ -3 \end{pmatrix} \\ &= -\frac{5}{49}\begin{pmatrix} 2 \\ 3 \\ 6 \end{pmatrix} \end{aligned}$$

The length of $\mathbf{w}$ is $\|\mathbf{w}\| = \frac{5}{49}\sqrt{4+9+36} = \frac{5}{7} = R_{3,3}$ and

$$\mathbf{q}_3 = -\frac{1}{7}\begin{pmatrix} 2 \\ 3 \\ 6 \end{pmatrix}$$

Hence the factorization is given by

$$A = QR = \frac{1}{7}\begin{pmatrix} 3 & 6 & -2 \\ -6 & 2 & -3 \\ 2 & -3 & -6 \end{pmatrix}\begin{pmatrix} 7 & 8 & -\frac{11}{7} \\ 0 & 3 & -\frac{1}{7} \\ 0 & 0 & \frac{5}{7} \end{pmatrix}.$$

2.5 EXERCISE 2.9

Calculate the QR factorization of the matrix in exercise 2.7 by using two Householder rotations. Show that for a general $m \times n$ matrix A the computational cost is $O(mn^2)$.

Solution

Since there is a level of choice within Householder rotations, several solutions are listed here. The first choice of $\mathbf{u}$ is $\mathbf{a}_1 + \|\mathbf{a}_1\|\mathbf{e}_1$ where $\mathbf{e}_1$ denotes the first unit vector.

$$\mathbf{u} = \begin{pmatrix} 3 \\ -6 \\ 2 \end{pmatrix} + \sqrt{9+36+4}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 10 \\ -6 \\ 2 \end{pmatrix}$$

We have $\|\mathbf{u}\|^2 = 100 + 36 + 4 = 140$ and

$$\begin{pmatrix} 10 \\ -6 \\ 2 \end{pmatrix} \begin{pmatrix} 10 & -6 & 2 \end{pmatrix} = \begin{pmatrix} 100 & -60 & 20 \\ -60 & 36 & -12 \\ 20 & -12 & 4 \end{pmatrix}$$

The first Householder transformation is

$$I - 2\frac{\mathbf{u}\mathbf{u}^T}{\|\mathbf{u}\|^2} = I - \frac{2}{140}\begin{pmatrix} 100 & -60 & 20 \\ -60 & 36 & -12 \\ 20 & -12 & 4 \end{pmatrix} = \frac{1}{35}\begin{pmatrix} -15 & -30 & -10 \\ 30 & 17 & 6 \\ -10 & 6 & 33 \end{pmatrix}$$

Multiplying this with A we arrive at

$$B = \frac{1}{35}\begin{pmatrix} -15 & -30 & -10 \\ 30 & 17 & 6 \\ -10 & 6 & 33 \end{pmatrix}\begin{pmatrix} 3 & 6 & -1 \\ -6 & -6 & 1 \\ 2 & 1 & -1 \end{pmatrix} = \begin{pmatrix} -7 & -8 & -\frac{11}{7} \\ 0 & \frac{12}{5} & -\frac{19}{35} \\ 0 & -\frac{9}{5} & -\frac{17}{35} \end{pmatrix}$$

The next choice for $\mathbf{u}$ is

$$\mathbf{u} = \begin{pmatrix} 0 \\ \frac{12}{5} \\ -\frac{9}{5} \end{pmatrix} + \frac{1}{5}\sqrt{144+81}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{27}{5} \\ -\frac{9}{5} \end{pmatrix}$$

we have $\|\mathbf{u}\|^2 = (729 + 81)/25 = 810/25 = 162/5$ and

$$\frac{1}{25}\begin{pmatrix} 0 \\ 27 \\ -9 \end{pmatrix}\begin{pmatrix} 0 & 27 & -9 \end{pmatrix} = \frac{1}{25}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 729 & -243 \\ 0 & -243 & 81 \end{pmatrix}$$

The second Householder transformation is

$$I - 2\frac{\mathbf{u}\mathbf{u}^T}{\|\mathbf{u}\|} = I - \frac{2*5}{25*162}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 729 & -243 \\ 0 & -243 & 81 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 5 & 0 & 0 \\ 0 & -4 & 3 \\ 0 & 3 & 4 \end{pmatrix}$$

Multiplying this with B we arrive at

$$\begin{aligned} R &= \tfrac{1}{5}\begin{pmatrix} 5 & 0 & 0 \\ 0 & -4 & 3 \\ 0 & 3 & 4 \end{pmatrix}\begin{pmatrix} -7 & -8 & -\frac{11}{7} \\ 0 & \frac{12}{5} & -\frac{19}{35} \\ 0 & -\frac{9}{5} & -\frac{17}{35} \end{pmatrix} \\ &= \begin{pmatrix} -7 & -8 & \frac{11}{7} \\ 0 & -3 & -\frac{1}{7} \\ 0 & 0 & -\frac{5}{7} \end{pmatrix} \end{aligned}$$

If instead we choose

$$\mathbf{u} = \begin{pmatrix} 0 \\ \frac{12}{5} \\ -\frac{9}{5} \end{pmatrix} - \frac{1}{5}\sqrt{144+81}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{3}{5} \\ -\frac{9}{5} \end{pmatrix},$$

we have $\|\mathbf{u}\|^2 = (9+81)/25 = 90/25 = 18/5$ and

$$\frac{1}{25}\begin{pmatrix} 0 \\ -3 \\ -9 \end{pmatrix}\begin{pmatrix} 0 & -3 & -9 \end{pmatrix} = \frac{1}{25}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 9 & 27 \\ 0 & 27 & 81 \end{pmatrix}$$

The second Householder transformation is then

$$I - 2\frac{\mathbf{u}\mathbf{u}^T}{\|\mathbf{u}\|} = I - \frac{2*5}{25*18}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 9 & 27 \\ 0 & 27 & 81 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 5 & 0 & 0 \\ 0 & 4 & -3 \\ 0 & -3 & -4 \end{pmatrix}$$

Multiplying this with B we arrive at

$$\begin{aligned} R &= \tfrac{1}{5}\begin{pmatrix} 5 & 0 & 0 \\ 0 & 4 & -3 \\ 0 & -3 & -4 \end{pmatrix}\begin{pmatrix} -7 & -8 & -\frac{11}{7} \\ 0 & \frac{12}{5} & -\frac{19}{35} \\ 0 & -\frac{9}{5} & -\frac{17}{35} \end{pmatrix} \\ &= \begin{pmatrix} -7 & -8 & \frac{11}{7} \\ 0 & 3 & -\frac{1}{7} \\ 0 & 0 & -\frac{5}{7} \end{pmatrix} \end{aligned}$$

If we choose in the first place

$$\mathbf{u} = \begin{pmatrix} 3 \\ -6 \\ 2 \end{pmatrix} - \sqrt{9+36+4}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -4 \\ -6 \\ 2 \end{pmatrix}$$

We have $\|\mathbf{u}\|^2 = 16+36+4 = 56$ and

$$\begin{pmatrix} -4 \\ -6 \\ 2 \end{pmatrix}\begin{pmatrix} -4 & -6 & 2 \end{pmatrix} = \begin{pmatrix} 16 & 24 & -8 \\ 24 & 36 & -12 \\ -8 & -12 & 4 \end{pmatrix}$$

The first Householder transformation is

$$I - 2\frac{\mathbf{u}\mathbf{u}^T}{\|\mathbf{u}\|^2} = I - \frac{2}{56}\begin{pmatrix} 16 & 24 & -8 \\ 24 & 36 & -12 \\ -8 & -12 & 4 \end{pmatrix} = \frac{1}{7}\begin{pmatrix} 3 & -6 & 2 \\ -6 & -2 & 3 \\ 2 & 3 & 6 \end{pmatrix}$$

Multiplying this with A we arrive at

$$B = \frac{1}{7}\begin{pmatrix} 3 & -6 & 2 \\ -6 & -2 & 3 \\ 2 & 3 & 6 \end{pmatrix}\begin{pmatrix} 3 & 6 & -1 \\ -6 & -6 & 1 \\ 2 & 1 & -1 \end{pmatrix} = \begin{pmatrix} 7 & 8 & -\frac{11}{7} \\ 0 & -3 & \frac{1}{7} \\ 0 & 0 & \frac{5}{7} \end{pmatrix}$$

and only one Householder transformation is necessary.

It requires $O(nm)$ operations to form $\mathbf{w}^T := \mathbf{u}^T A$ and $A - \frac{2}{\|\mathbf{u}\|^2}\mathbf{u}\mathbf{w}$. This is done $n-1$ times. Thus the overall number of operations is $O(mn^2)$.

2.6 EXERCISE 2.11

Starting with an arbitrary $\mathbf{x}^{(0)}$ *the sequence* $\mathbf{x}^{(k)}$, $k = 1, 2, \ldots$, *is calculated by*

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{x}^{(k+1)} + \begin{pmatrix} 0 & 0 & 0 \\ \alpha & 0 & 0 \\ \gamma & \beta & 0 \end{pmatrix} \mathbf{x}^{(k)} = \mathbf{b}$$

in order to solve the linear system

$$\begin{pmatrix} 1 & 1 & 1 \\ \alpha & 1 & 1 \\ \gamma & \beta & 1 \end{pmatrix} \mathbf{x} = \mathbf{b},$$

where $\alpha, \beta, \gamma \in \mathbb{R}$ *are constant. Find all values for* α, β, γ *such that the sequence converges for every* $\mathbf{x}^{(0)}$ *and* $\mathbf{b}$. *What happens when* $\alpha = \beta = \gamma = -1$ *and* $\alpha = \beta = 0$*?*

Solution

The iteration matrix is given by

$$\begin{aligned} H &= -\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 & 0 \\ \alpha & 0 & 0 \\ \gamma & \beta & 0 \end{pmatrix} \\ &= -\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ \alpha & 0 & 0 \\ \gamma & \beta & 0 \end{pmatrix} \\ &= \begin{pmatrix} \alpha & 0 & 0 \\ -\alpha + \gamma & \beta & 0 \\ -\gamma & -\beta & 0 \end{pmatrix} \end{aligned}$$

The eigenvalues of H are given by

$$(\alpha - \lambda)(\beta - \lambda)(-\lambda) = 0.$$

Hence the eigenvalues are α, β and zero. Therefore we need for convergence $|\alpha| < 1$ and $|\beta| < 1$.

For $\alpha = \beta = \gamma = -1$ the iteration matrix becomes

$$H = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

In this case

$$H^2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix} = -H.$$

Hence $H^k = (-1)^{k-1}H$. The sequence then alternates for any starting vector $\mathbf{x}^{(0)}$ between $H\mathbf{x}^{(0)}$ and $-H\mathbf{x}^{(0)}$.

For $\alpha = \beta = 0$ the iteration matrix becomes

$$H = \begin{pmatrix} 0 & 0 & 0 \\ \gamma & 0 & 0 \\ -\gamma & 0 & 0 \end{pmatrix}.$$

We have $H^2 = 0$ and thus the sequence has to converge after only two iterations.

2.7 EXERCISE 2.13

The Gauss–Seidel method is used to solve $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{pmatrix} 100 & 11 \\ 9 & 1 \end{pmatrix}.$$

Find the eigenvalues of the iteration matrix. Then show that with relaxation the spectral radius can be reduced by nearly a factor of 3. In addition show that after one iterations with the relaxed method the error $\|\mathbf{x}^{(k)} - \mathbf{x}^*\|$ is reduced by more than a factor of 3. Estimate the number of iterations the original Gauss–Seidel would need to achieve a similar decrease in the error.

Solution

The Gauss-Seidel iteration matrix is

$$H = \begin{pmatrix} 100 & 0 \\ 9 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 11 \\ 0 & 0 \end{pmatrix} = \frac{1}{100} \begin{pmatrix} 0 & 11 \\ 0 & -99 \end{pmatrix}.$$

The eigenvalues are 0 and $-\frac{99}{100}$ and thus the spectral radius is 0.99. To find out how many iterations are needed to reduce the error by a factor of 3, we need to find k such that $(0.99)^k < 1/3$. This means $k > \frac{\log 1/3}{\log 0.99} \approx 109.3$. Thus 110 iterations are needed.

The iteration matrix of the relaxation method is given by $H_\omega = (1-\omega)I + \omega H$. The optimal ω minimizes $\{|1-\omega+\omega\lambda| : \lambda = 0, -\frac{99}{100}\}$. Setting $-(1-\omega-\frac{99}{100}\omega) = 1-\omega$ gives $\omega = \frac{200}{299}$. The eigenvalues of the relaxation method are then $-\frac{99}{299}$ and $\frac{99}{299}$. The relaxed spectral radius is $\frac{99}{299} = \frac{100}{299} \times \frac{99}{100} = \frac{1}{2.99}\frac{99}{100}$ and thus the spectral radius is reduced by a factor of 2.99, nearly 3.

One iteration reduces the error by $\frac{99}{299} \approx 0.331$, which is a factor of approximately 3. Thus relaxation is very effective in this case.

2.8 EXERCISE 2.15

Use the standard form of the conjugate gradient method to solve

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

starting with $\mathbf{x}^{(0)} = 0$. *Show that the residuals* $\mathbf{r}^{(0)}, \mathbf{r}^{(1)}$ *and* $\mathbf{r}^{(2)}$ *are mutually orthogonal and that the search directions* $\mathbf{d}^{(0)}, \mathbf{d}^{(1)}$ *and* $\mathbf{d}^{(2)}$ *are mutually conjugate and that* $\mathbf{x}^{(3)}$ *is the solution.*

Solution

The intermediate values are as follows:

$$\omega^{(0)} = \frac{1}{2}, \qquad \mathbf{d}^{(0)} = \mathbf{r}^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

$$\mathbf{x}^{(1)} = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{g}^{(1)} = \begin{pmatrix} -\frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}, \quad \beta^{(1)} = \frac{1}{6},$$

$$\mathbf{d}^{(1)} = \frac{1}{6}\begin{pmatrix} 4 \\ 1 \\ -2 \end{pmatrix}, \quad \omega^{(1)} = \frac{3}{5},$$

$$\mathbf{x}^{(2)} = \frac{1}{10}\begin{pmatrix} 9 \\ 6 \\ 3 \end{pmatrix}, \quad \mathbf{g}^{(2)} = \frac{1}{10}\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}, \quad \beta^{(2)} = \frac{3}{25},$$

$$\mathbf{d}^{(2)} = \frac{1}{50}\begin{pmatrix} 9 \\ -9 \\ 3 \end{pmatrix}, \quad \omega^{(2)} = \frac{5}{9}.$$

The method converges with

$$\mathbf{x}^{(3)} = \begin{pmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{3} \end{pmatrix}.$$

It is easy to check the orthogonality and conjugacy properties.

2.9 EXERCISE 2.17

Let A *be the bidiagonal* $n \times n$ *matrix*

$$A = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \lambda & 1 \\ & & & \lambda \end{pmatrix}.$$

Find an explicit expression for A^k. Letting $n = 3$, the sequence $\mathbf{x}^{(k+1)}$, $k = 0, 1, 2, \ldots$, is generated by the power method $\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)}/\|\mathbf{x}^{(k)}\|$, starting with some $\mathbf{x}^{(0)} \in \mathbb{R}^3$. From the expression for A^k deduce that the second and third component of $\mathbf{x}^{(k)}$ tend to zero as k tend to infinity. Further show that this implies $A\mathbf{x}^{(k)} - \lambda\mathbf{x}^{(k)}$ tends to zero.

Solution

The matrix A is in the *Jordan canonical form*. Let

$$B = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{pmatrix}.$$

We can express A^k as

$$\begin{aligned} A^k &= [\lambda I + B]^k \\ &= \lambda^k I + \binom{k}{1} \lambda^{k-1} B + \binom{k}{2} \lambda^{k-2} B^2 + \ldots + \binom{k}{k} B^k. \end{aligned}$$

With every multiplication by B the superdiagonal consisting of 1 is moving up by 1. Thus B^j is the zero matrix except for the j-th superdiagonal where all entries are 1. Hence for $k \geq n$

$$A^k = \begin{pmatrix} \lambda^k & \binom{k}{1} \lambda^{k-1} & \cdots & \binom{k}{n-1} \lambda^{k-n+1} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \binom{k}{1} \lambda^{k-1} \\ 0 & \cdots & 0 & \lambda^k \end{pmatrix}.$$

For $n = 3$ this becomes

$$A^k = \begin{pmatrix} \lambda^k & k\lambda^{k-1} & k(k-1)\lambda^{k-2}/2 \\ 0 & \lambda^k & k\lambda^{k-1} \\ 0 & 0 & \lambda^k \end{pmatrix}.$$

Now

$$\mathbf{x}^{(k)} = \frac{A\mathbf{x}^{(k-1)}}{\|A\mathbf{x}^{(k-1)}\|} = \frac{A\frac{A\mathbf{x}^{(k-2)}}{\|A\mathbf{x}^{(k-2)}\|}}{\|A\frac{A\mathbf{x}^{(k-2)}}{\|A\mathbf{x}^{(k-2)}\|}\|} = \frac{A^2\mathbf{x}^{(k-2)}}{\|A^2\mathbf{x}^{(k-2)}\|} = \ldots = \frac{A^k\mathbf{x}^{(0)}}{\|A^k\mathbf{x}^{(0)}\|}.$$

The second and third component of $A^k\mathbf{x}^{(0)}$ are of magnitude $k\lambda^{k-1}$ and λ^k while $\|A^k\mathbf{x}^{(0)}\|$ is of magnitude $k^2\lambda^k$. Therefore the second and third component of $A^k\mathbf{x}^{(0)}$ tend to zero as k tends to infinity.

For the last part of the question

$$A\mathbf{x}^{(k)} - \lambda\mathbf{x}^{(k)} = B\mathbf{x}^{(k)}$$

and the resultant vector has the second component of $\mathbf{x}^{(k)}$ as its first component, the third component of $\mathbf{x}^{(k)}$ as its second component and zero as its third component. Thus this tends to zero as k tends to infinity.

2.10 EXERCISE 2.19

The symmetric matrix

$$A = \begin{pmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{pmatrix}$$

has the eigenvector $\mathbf{v} = (2, 1, 2)^T$. *Use a Householder reflection to find an orthogonal matrix* S *such that* $S\mathbf{v}$ *is a multiple of the first standard unit vector* $\mathbf{e}_1$. *Calculate* SAS. *The resultant matrix is suitable for deflation and hence identify the remaining eigenvalues and eigenvectors.*

Solution

Since

$$\begin{pmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 16 \\ 8 \\ 16 \end{pmatrix} = 8 \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix},$$

the given eigenvector has the eigenvalue $\lambda_1 = 8$. $\|\mathbf{v}\| = \sqrt{4+1+4} = 3$. One choice $\mathbf{u}$ for the Householder reflection is then

$$\mathbf{u} = \mathbf{v} + 3\mathbf{e}_1 = \begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix},$$

where the sign was chosen to avoid loss of significance. We have $\|\mathbf{u}\|^2 = 25 + 1 + 4 = 30$ and

$$\begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix} \begin{pmatrix} 5 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 25 & 5 & 10 \\ 5 & 1 & 2 \\ 10 & 2 & 4 \end{pmatrix}.$$

The Householder reflection is then given by

$$S = I - \frac{2}{30} \begin{pmatrix} 25 & 5 & 10 \\ 5 & 1 & 2 \\ 10 & 2 & 4 \end{pmatrix} = \frac{1}{15} \begin{pmatrix} -10 & -5 & -10 \\ -5 & 14 & -2 \\ -10 & -2 & 11 \end{pmatrix}.$$

The deflated matrix is

$$\begin{aligned} SAS &= \frac{1}{15^2}\begin{pmatrix} -10 & -5 & -10 \\ -5 & 14 & -2 \\ -10 & -2 & 11 \end{pmatrix}\begin{pmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{pmatrix}\begin{pmatrix} -10 & -5 & -10 \\ -5 & 14 & -2 \\ -10 & -2 & 11 \end{pmatrix} \\ &= \begin{pmatrix} 8 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \end{aligned}$$

The other eigenvalues are therefore a double eigenvalue -1. The corresponding eigenvectors of SAS are $\mathbf{e}_2$ and $\mathbf{e}_3$, but these equal also $S\mathbf{v}_i$ for $i = 2, 3$. Thus $\mathbf{v}_2$ and $\mathbf{v}_3$ are found by multiplying by S. The eigenvectors are

$$\mathbf{v}_2 = \frac{1}{15}\begin{pmatrix} -5 \\ 14 \\ -2 \end{pmatrix} \text{ and } \mathbf{v}_3 = \frac{1}{15}\begin{pmatrix} -10 \\ -2 \\ 11 \end{pmatrix}.$$

Note that $\mathbf{v}_2$ and $\mathbf{v}_3$ span an eigenspace and thus any multiple or linear combination of $\mathbf{v}_2$ and $\mathbf{v}_3$ is also an eigenvector.

If we choose the other sign when forming $\mathbf{u}$, the solution is as follows.

$$\mathbf{u} = \mathbf{v} - 3\mathbf{e}_1 = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}.$$

We have $\|\mathbf{u}\|^2 = 1 + 1 + 4 = 6$ and

$$\begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}\begin{pmatrix} -1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & -1 & -2 \\ -1 & 1 & 2 \\ -2 & 2 & 4 \end{pmatrix}.$$

The Householder reflection is then given by

$$S = I - \frac{2}{6}\begin{pmatrix} 1 & -1 & -2 \\ -1 & 1 & 2 \\ -2 & 2 & 4 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 2 & 1 & 2 \\ 1 & 2 & -2 \\ 2 & -2 & -1 \end{pmatrix}.$$

The deflated matrix is

$$\begin{aligned} SAS &= \frac{1}{3^2}\begin{pmatrix} 2 & 1 & 2 \\ 1 & 2 & -2 \\ 2 & -2 & -1 \end{pmatrix}\begin{pmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{pmatrix}\begin{pmatrix} 2 & 1 & 2 \\ 1 & 2 & -2 \\ 2 & -2 & -1 \end{pmatrix} \\ &= \begin{pmatrix} 8 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \end{aligned}$$

As before the eigenvectors of SAS are $\mathbf{e}_2$ and $\mathbf{e}_3$, but these equal also $S\mathbf{v}_i$ for $i = 2, 3$. Thus $\mathbf{v}_2$ and $\mathbf{v}_3$ are found by multiplying by S which is different to before. The eigenvectors are

$$\mathbf{v}_2 = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} \text{ and } \mathbf{v}_3 = \begin{pmatrix} 2 \\ -2 \\ -1 \end{pmatrix}.$$

2.11 EXERCISE 2.21

(a) Explain the technique of splitting for solving the linear system $A\mathbf{x} = \mathbf{b}$ *iteratively where* A *is an* $n \times n$*, non-singular matrix. Define the iteration matrix* H *and state the property it has to satisfy to ensure convergence.*

(b) Define the Gauss-Seidel and Jacobi iterations and state their iteration matrices respectively.

(c) Describe relaxation and consider briefly the cases when the relaxation parameter ω *equals 0 and 1.*

(d) Show how the the iteration matrix H_ω *of the relaxed method is related to the iteration matrix* H *of the original method and thus how the eigenvalues are related. How should* ω *be chosen?*

(e) We now consider the tridiagonal matrix A *with diagonal elements* $A_{i,i} = 1$ *and off-diagonal elements* $A_{i,i-1} = A_{i,i+1} = 1/4$*. Calculate the iteration matrices* H *of the Jacobi method and* H_ω *of the relaxed Jacobi method.*

(f) The eigenvectors of both H *and* H_ω *are* $\mathbf{v}_1, \ldots, \mathbf{v}_n$ *where the* i*-th component of* $\mathbf{v}_k$ *is given by* $(\mathbf{v}_k)_i = \sin \frac{\pi i k}{n+1}$*. Calculate the eigenvalues of* H *by evaluating* $H\mathbf{v}_k$ *(Hint:* $\sin(x \pm y) = \sin x \cos y \pm \cos x \sin y$*).*

(g) Using the formula for the eigenvalues of H_ω *derived earlier state the eigenvalues of* H_ω *and show that the relaxed method converges for* $0 < \omega \leq 4/3$.

Solution

(a) We can rewrite $A\mathbf{x} = \mathbf{b}$ in the form

$$(A - B)\mathbf{x} = -B\mathbf{x} + \mathbf{b},$$

where the matrix B is chosen in such a way that $A - B$ is non-singular and the system $(A-B)\mathbf{x} = \mathbf{y}$ is easily solved for any right hand side $\mathbf{y}$. A simple iterative scheme starts with an estimate $\mathbf{x}^{(0)} \in \mathbb{R}^n$ of the solution and generates the sequence $\mathbf{x}^{(k)}$, $k = 1, 2, \ldots$, by solving

$$(A - B)\mathbf{x}^{(k+1)} = -B\mathbf{x}^{(k)} + \mathbf{b}.$$

Let $\mathbf{x}^*$ solve $A\mathbf{x} = \mathbf{b}$. It also satisfies $(A-B)\mathbf{x}^* = -B\mathbf{x}^* + \mathbf{b}$. Subtracting this equation from the above equation gives

$$(A - B)(\mathbf{x}^{(k+1)} - \mathbf{x}^*) = -B(\mathbf{x}^{(k)} - \mathbf{x}^*).$$

We denote $\mathbf{x}^{(k)} - \mathbf{x}^*$ by $\mathbf{e}^{(k)}$. It is the error in the k-th iteration. Since $A - B$ is non-singular, we can write

$$\mathbf{e}^{(k+1)} = -(A - B)^{-1}B\mathbf{e}^{(k)}.$$

The matrix $H := -(A-B)^{-1}B$ is known as the *iteration matrix*. We have $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$ for all $\mathbf{x}^{(0)} \in \mathbb{R}^n$ if and only if $\rho(H) < 1$. That is all eigenvalues of H must have modulus less than 1.

(b) **Jacobi method** We choose $A - B = D$, the diagonal part of A, or in other words we let $B = L + U$. The iteration step is given by

$$D\mathbf{x}^{(k+1)} = -(L+U)\mathbf{x}^{(k)} + \mathbf{b}.$$

The iteration matrix is $-D^{-1}(L+U) = -D^{-1}(A-D)$.

Gauss–Seidel method We set $A - B = L + D$, the lower triangular portion of A, or in other words $B = U$. The sequence $\mathbf{x}^{(k)}, k = 1, \ldots,$ is generated by

$$(L+D)\mathbf{x}^{(k+1)} = -U\mathbf{x}^{(k)} + \mathbf{b}.$$

The iteration matrix is $-(L+D)^{-1}U$.

(c) For relaxation, we first calculate $(A-B)\tilde{\mathbf{x}}^{(k+1)} = -B\mathbf{x}^{(k)} + \mathbf{b}$ as an intermediate value and then let

$$\mathbf{x}^{(k+1)} = \omega\tilde{\mathbf{x}}^{(k+1)} + (1-\omega)\mathbf{x}^{(k)}$$

for $k = 0, 1, \ldots$, where $\omega \in \mathbb{R}$ is called the *relaxation parameter*. Of course $\omega = 1$ corresponds to the original method without relaxation. The choice $\omega = 0$ does not make sense, since in this case $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}$.

(d) Since $\tilde{\mathbf{x}}^{(k+1)} = -(A-B)^{-1}B\mathbf{x}^{(k)} + (A-B)^{-1}\mathbf{b}$, let $\mathbf{c} = (A-B)^{-1}\mathbf{b}$, the relaxation iteration matrix H_ω can then be deduced from

$$\mathbf{x}^{(k+1)} = \omega\tilde{\mathbf{x}}^{(k+1)} + (1-\omega)\mathbf{x}^{(k)} = \omega H\mathbf{x}^{(k)} + (1-\omega)\mathbf{x}^{(k)} + \omega\mathbf{c}$$

as

$$H_\omega = \omega H + (1-\omega)I.$$

It follows that an eigenvalue λ of H is related to an eigenvalue λ_ω of H_ω by $\lambda_\omega = \omega\lambda + (1-\omega)$. The best choice for ω would be to minimize $\max\{|\omega\lambda_i + (1-\omega)|, i = 1, \ldots, n\}$ where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of H.

(e) The iteration matrix of the Jacobi method is $H = -D^{-1}(A-D)$ and has entries $H_{i,i} = 0$ and off-diagonal elements $H_{i,i-1} = H_{i,i+1} = -1/4$. All the other entries are zero. The iteration matrix of the relaxed Jacobi method is $H_\omega = \omega H + (1-\omega)I$ and has entries $H_{i,i} = (1-\omega)$ and off-diagonal elements $H_{i,i-1} = H_{i,i+1} = -\omega/4$. All the other entries are zero.

(f) The i-th component of $H\mathbf{v}_k$ is given by

$$\begin{aligned}(H\mathbf{v}_k)_i &= -\frac{1}{4}\sin\frac{\pi(i-1)k}{n+1} - \frac{1}{4}\sin\frac{\pi(i+1)k}{n+1} \\ &= -\frac{1}{4}(\sin\frac{\pi ik}{n+1}\cos\frac{\pi k}{n+1} - \cos\frac{\pi ik}{n+1}\sin\frac{\pi k}{n+1}) \\ &\quad -\frac{1}{4}(\sin\frac{\pi ik}{n+1}\cos\frac{\pi k}{n+1} + \cos\frac{\pi ik}{n+1}\sin\frac{\pi k}{n+1}) \\ &= -\frac{1}{2}\cos\frac{\pi k}{n+1}\sin\frac{\pi ik}{n+1},\end{aligned}$$

where we used $\sin(x \pm y) = \sin x\cos y \pm \cos x\sin y$ with $x = \frac{\pi ik}{n+1}$ and $y = \frac{\pi k}{n+1}$. Thus the eigenvalues are $\lambda_k = -\frac{1}{2}\cos\frac{\pi k}{n+1}$, $k = 1, \ldots, n$.

(g) The eigenvalues of the relaxed method are

$$\lambda_{\omega k} = -\omega\frac{1}{2}\cos\frac{\pi k}{n+1} + 1 - \omega = 1 - \omega(1 + \frac{1}{2}\cos\frac{\pi k}{n+1}).$$

For convergence these have to lie in the interval $(-1, 1)$. That means $\omega(1 + \frac{1}{2}\cos\frac{\pi k}{n+1})$ has to lie in the interval $(0, 2)$. Since $\cos\frac{\pi k}{n+1} \in (-1, 1)$ for $k = 1, \ldots, n$, we have $1 + \frac{1}{2}\cos\frac{\pi k}{n+1} \in (1/2, 3/2)$. From this we can deduce $0 < \omega \leq 3/4$.

2.12 EXERCISE 2.23

(a) Given an $n \times n$ matrix A, define the concept of LU factorization and how it can be used to solve the system of equations $A\mathbf{x} = \mathbf{b}$.

(b) State two other applications of the LU factorization.

(c) Describe the algorithm to obtain an LU factorization. How many operations does this generally require?

(d) Describe the concept of pivoting in the context of solving the system of equations $A\mathbf{x} = \mathbf{b}$ by LU factorization.

(e) How does the algorithm need to be adjusted if in the process we encounter a column with all entries equal to zero? What does it mean if there is a column consisting entirely of zeros in the process?

(f) How can sparsity be exploited in the LU factorization?

(g) Calculate the LU factorization with pivoting *of the matrix*

$$A = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix}.$$

Solution

(a) The LU factorization factorizes A into a lower triangular matrix L (i.e. $L_{i,j} = 0$ for $i < j$) and a upper triangular matrix U (i.e. $U_{i,j} = 0$ for $i > j$) such that is $A = LU$. The linear system then becomes $L(U\mathbf{x}) = \mathbf{b}$, which we decompose into $L\mathbf{y} = \mathbf{b}$ and $U\mathbf{x} = \mathbf{y}$. Both these systems can be solved easily by back substitution.

(b) Other applications of the LU factorization are

- Calculation of determinant:

$$\det A = (\det L)(\det U) = (\prod_{k=1}^{n} L_{k,k})(\prod_{k=1}^{n} U_{k,k}).$$

- Non-singularity testing: $A = LU$ is non-singular if and only if all the diagonal elements of L and U are nonzero.
- Calculating the inverse: The inverse of triangular matrices can be easily calculated directly. Subsequently $A^{-1} = U^{-1}L^{-1}$.

(c) The algorithm of the LU factorization is as follows:

(1) Set $A_0 := A$ and $k = 1$.

(2) Set $\mathbf{u}_k^T$ to the k-th row of A_{k-1} and $\mathbf{l}_k$ to the k-th column of A_{k-1}, scaled so that $L_{k,k} = 1$.

(3) Calculate $A_k := A_{k-1} - \mathbf{l}_k\mathbf{u}_k^T$ before incrementing k by 1 and returning to step (2) if $k \leq n$.

The full LU accumulation requires $O(n^3)$ operations.

(d) Pivoting means having obtained A_{k-1}, we exchange two rows of A_{k-1} so that the element of largest magnitude in the k-th column is in the pivotal position (k, k), i.e.

$$|(A_{k-1})_{k,k}| \geq |(A_{k-1})_{j,k}|, \qquad j = 1, \ldots, n.$$

Since the exchange of rows can be regarded as the pre-multiplication of the relevant matrix by a permutation matrix, we need to do the same exchange in the portion of L that has been formed already (i.e. the first $k - 1$ columns):

$$A_{k-1}^{\text{new}} = PA_{k-1} = PA - P\sum_{j=1}^{k-1} \mathbf{l}_j\mathbf{u}_j^T = PA - \sum_{j=1}^{k-1} P\mathbf{l}_j\mathbf{u}_j^T.$$

We also need to record the permutations of rows to solve for $\mathbf{b}$.

(e) If the entire k-th column of A_{k-1} is zero, we let $\mathbf{l}_k$ be the k-th unit vector while $\mathbf{u}_k^T$ is the k-th row of A_{k-1} as before. With this choice we retain that the matrix $\mathbf{l}_k\mathbf{u}_k^T$ has the same k-th row and column as A_{k-1}. If a column consisting entirely of zeros is encountered in the process, it means that the matrix A is singular and the solution is not unique.

(f) If A is 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. Therefore we should use the freedom to exchange rows and columns in a preliminary calculation so that many of the zero elements are leading zero elements in rows and columns.

(g) Starting from

$$A = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix}$$

we swap the first and third row to obtain

$$A_0 = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 4 & 3 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{pmatrix}.$$

The first row of U is the first row of A_0, $\mathbf{u}_1^T = (8, 7, 9, 5)$, and the first column of L is the first column of A_0 scaled by $1/8$, $\mathbf{l}_1^T = (1, 1/2, 1/4, 3/4)$. Then

$$\mathbf{l}_1\mathbf{u}_1^T = \begin{pmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{4} \\ \frac{3}{4} \end{pmatrix} \begin{pmatrix} 8 & 7 & 9 & 5 \end{pmatrix} = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 4 & \frac{7}{2} & \frac{9}{2} & \frac{5}{2} \\ 2 & \frac{7}{4} & \frac{9}{4} & \frac{5}{4} \\ 6 & \frac{21}{4} & \frac{27}{4} & \frac{5}{4} \end{pmatrix}.$$

We then calculate

$$A_1 = A_0 - \mathbf{l}_1\mathbf{u}_1^T = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} \\ 0 & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \end{pmatrix}.$$

Since the leading nonzero coefficient in the fourth row is the largest in the second column we swap the second and fourth row:

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ 0 & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ 0 & -\frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} \end{pmatrix}.$$

We have to do the same swap on the portion of L which has already been calculated, $\mathbf{l}_1^T = (1, 3/4, 1/4, 1/2)$.

The second row of U is the second row of A_1, $\mathbf{u}_2^T = (0, \frac{7}{4}, \frac{9}{4}, \frac{17}{4})$, and the second column of L is the second column of A_1 scaled by $\frac{4}{7}$, $\mathbf{l}_2^T = (0, 1, -\frac{3}{7}, -\frac{2}{7})$. Then

$$\mathbf{l}_2\mathbf{u}_2^T = \begin{pmatrix} 0 \\ 1 \\ -\frac{3}{7} \\ -\frac{2}{7} \end{pmatrix} \begin{pmatrix} 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ 0 & -\frac{3}{4} & -\frac{27}{28} & -\frac{51}{28} \\ 0 & -\frac{1}{2} & -\frac{9}{14} & -\frac{17}{14} \end{pmatrix}.$$

We then calculate

$$\begin{aligned} A_2 &= A_1 - \mathbf{l}_2\mathbf{u}_2^T = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{35}{28} + \frac{27}{28} & -\frac{35}{28} + \frac{51}{28} \\ 0 & 0 & -\frac{21}{14} + \frac{9}{14} & -\frac{21}{14} + \frac{17}{14} \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{2}{7} & \frac{4}{7} \\ 0 & 0 & -\frac{6}{7} & -\frac{2}{7} \end{pmatrix}. \end{aligned}$$

Since the leading nonzero coefficient in the fourth row is the largest in the third column we swap the third and fourth row:

$$A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{6}{7} & -\frac{2}{7} \\ 0 & 0 & -\frac{2}{7} & \frac{4}{7} \end{pmatrix}.$$

We have to do the same swap on the portions of L which have already been calculated, $\mathbf{l}_1^T = (1, 3/4, 1/2, 1/4)$ and $\mathbf{l}_2^T = (0, 1, -\frac{2}{7}, -\frac{3}{7})$.

The third row of U is the third row of A_2, $\mathbf{u}_3^T = (0, 0, -\frac{6}{7}, -\frac{2}{7})$, and the third column of L is the third column of A_2 scaled by $-\frac{7}{6}$, $\mathbf{l}_3^T = (0, 0, 1, \frac{1}{3})$. Then

$$\mathbf{l}_3\mathbf{u}_3^T = \begin{pmatrix} 0 \\ 0 \\ 1 \\ \frac{1}{3} \end{pmatrix} \begin{pmatrix} 0 & 0 & -\frac{6}{7} & -\frac{2}{7} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{6}{7} & -\frac{2}{7} \\ 0 & 0 & -\frac{2}{7} & -\frac{2}{21} \end{pmatrix}.$$

We then calculate

$$A_3 = A_2 - \mathbf{l}_3\mathbf{u}_3^T = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{12}{21} + \frac{2}{21} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{2}{3} \end{pmatrix}.$$

The fourth row of U is the fourth row of A_3, $\mathbf{u}_4^T = (0, 0, 0, \frac{2}{3})$, and the fourth column of L is $\mathbf{l}_4^T = (0, 0, 0, 1)$. Summarizing we have

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{3}{4} & 1 & 0 & 0 \\ \frac{1}{2} & -\frac{2}{7} & 1 & 0 \\ \frac{1}{4} & -\frac{3}{7} & \frac{1}{3} & 1 \end{pmatrix} \qquad \text{and} \qquad U = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ 0 & 0 & -\frac{6}{7} & -\frac{2}{7} \\ 0 & 0 & 0 & \frac{2}{3} \end{pmatrix}.$$

To check (but this is not required) we may multiply

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{3}{4} & 1 & 0 & 0 \\ \frac{1}{2} & -\frac{2}{7} & 1 & 0 \\ \frac{1}{4} & -\frac{3}{7} & \frac{1}{3} & 1 \end{pmatrix} \begin{pmatrix} 8 & 7 & 9 & 5 \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ 0 & 0 & -\frac{6}{7} & -\frac{2}{7} \\ 0 & 0 & 0 & \frac{2}{3} \end{pmatrix} = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \\ 4 & 3 & 3 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix},$$

which is the original matrix after swapping first and third row,the second and fourth row and then third and fourth row.

2.13 EXERCISE 2.25

(a) Explain the technique of splitting for solving the linear system $A\mathbf{x} = \mathbf{b}$ iteratively where A is an $n \times n$, non-singular matrix. Define the iteration matrix H and state the property it has to satisfy to ensure convergence.

(b) Define what it means for a matrix to be positive definite. Show that all diagonal elements of a positive definite matrix are positive.

(c) State the Householder-John theorem and explain how it can be used to design iterative methods for solving $A\mathbf{x} = \mathbf{b}$.

(d) Let the iteration matrix H have a real *eigenvector $\mathbf{v}$ with* real *eigenvalue λ. Show that the condition of the Householder-John theorem implies that $|\lambda| < 1$.*

(e) We write A in the form $A = L + D + U$, where L is the subdiagonal (or strictly lower triangular), D is the diagonal and U is the superdiagonal (or strictly upper triangular) portion of A. The following iterative scheme is suggested

$$(L + \omega D)\mathbf{x}^{(k+1)} = -[(1 - \omega)D + U]\mathbf{x}^{(k)} + \mathbf{b}.$$

Using the Householder-John theorem, for which values of ω does the scheme converge in the case when A is symmetric and positive definite.

Solution

(a) We can rewrite $A\mathbf{x} = \mathbf{b}$ in the form

$$(A - B)\mathbf{x} = -B\mathbf{x} + \mathbf{b},$$

where the matrix B is chosen in such a way that $A - B$ is non-singular and the system $(A-B)\mathbf{x} = \mathbf{y}$ is easily solved for any right hand side $\mathbf{y}$. A simple iterative scheme starts with an estimate $\mathbf{x}^{(0)} \in \mathbb{R}^n$ of the solution and generates the sequence $\mathbf{x}^{(k)}$, $k = 1, 2, \ldots$, by solving

$$(A - B)\mathbf{x}^{(k+1)} = -B\mathbf{x}^{(k)} + \mathbf{b}.$$

Let $\mathbf{x}^*$ solve $A\mathbf{x} = \mathbf{b}$. It also satisfies $(A-B)\mathbf{x}^* = -B\mathbf{x}^* + \mathbf{b}$. Subtracting this equation from the above equation gives

$$(A - B)(\mathbf{x}^{(k+1)} - \mathbf{x}^*) = -B(\mathbf{x}^{(k)} - \mathbf{x}^*).$$

We denote $\mathbf{x}^{(k)} - \mathbf{x}^*$ by $\mathbf{e}^{(k)}$. It is the error in the k-th iteration. Since $A - B$ is non-singular, we can write

$$\mathbf{e}^{(k+1)} = -(A - B)^{-1}B\mathbf{e}^{(k)}.$$

The matrix $H := -(A-B)^{-1}B$ is known as the *iteration matrix*. We have $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$ for all $\mathbf{x}^{(0)} \in \mathbb{R}^n$ if and only if $\rho(H) < 1$. That is all eigenvalues of H must have modulus less than 1.

(b) A matrix A is said to be positive definite if for all $\mathbf{x} \neq 0$ we have $\mathbf{x}^T A\mathbf{x} > 0$. Denoting the i-th standard vector by $\mathbf{e}_i$, we have $A_{i,i} = \mathbf{e}_i^T A\mathbf{e}_i > 0$. Thus we can deduce that each diagonal element of A is positive.

(c) We have

Theorem (Householder–John theorem)**.** *If A and B are real matrices such that both A and $A - B - B^T$ are symmetric and positive definite, then the spectral radius of $H = -(A - B)^{-1}B$ is strictly less than one.*

Thus if A is symmetric and positive definite B can be chosen in such a way that $A - B - B^T$ is also positive definite.

(d) From $H\mathbf{v} = \lambda\mathbf{v}$ we deduce $-B\mathbf{v} = \lambda(A - B)\mathbf{v}$. λ cannot equal one since otherwise A would map $\mathbf{v}$ to zero and be singular. Multiplying by $\mathbf{v}^T$ from then left, we deduce $-\mathbf{v}^T B\mathbf{v} = \lambda\mathbf{v}^T(A - B)\mathbf{v}$ or in other words

$$\mathbf{v}^T B\mathbf{v} = \frac{\lambda}{\lambda - 1}\mathbf{v}^T A\mathbf{v}.$$

Since the positive definiteness of A and $A - B - B^T$ implies $\mathbf{v}^T A\mathbf{v} > 0$ and $\mathbf{v}^T(A - B - B^T)\mathbf{v} > 0$, we can insert the above result in the latter inequality and deduce

$$\begin{aligned} 0 < \mathbf{v}^T A\mathbf{v} - \mathbf{v}^T B\mathbf{v} - \mathbf{v}^T B^T\mathbf{v} &= \left(1 - \frac{\lambda}{\lambda - 1} - \frac{\lambda}{\lambda - 1}\right)\mathbf{v}^T A\mathbf{v} \\ &= \frac{-1 - \lambda}{\lambda - 1}\mathbf{v}^T A\mathbf{v}. \end{aligned}$$

Since $\mathbf{v}^T A\mathbf{v} > 0$, for the equality to hold the fraction has to be greater than zero. Multiplying the denominator and numerator of the fraction by $\lambda + 1$ shows that the fraction equals $-(\lambda+1)^2/(\lambda^2-1)$. Thus we must have $|\lambda| < 1$.

(e) For the scheme

$$(L + \omega D)\mathbf{x}^{(k+1)} = -[(1-\omega)D + U]\mathbf{x}^{(k)} + \mathbf{b}.$$

we have $A - B = L + \omega D$ and $B = (1-\omega)D + U$. Now since A is symmetric, $U = L^T$ and thus $B = (1-\omega)D + L^T$. The matrix A is symmetric and positive definite, hence we have to consider the matrix $A - B - B^T = L + \omega D - (1-\omega)D^T - L = (2\omega - 1)D$, since $D^T = D$. It has been shown above that all diagonal elements of D are positive, since it is the diagonal portion of the positive definite matrix A. Therefore the matrix $(2\omega - 1)D$ is positive definite if and only if $2\omega > 1$.

2.14 EXERCISE 2.27

(a) Use Gaussian Elimination with backwards substitution to solve the linear system:

$$\begin{array}{rrrcr} 5x_1+ & 10x_2+ & 9x_3 & = & 4 \\ 10x_1+ & 26x_2+ & 26x_3 & = & 10 \\ 15x_1+ & 54x_2+ & 66x_3 & = & 27 \end{array}$$

(b) How is the LU factorization defined, if A is an $n \times n$ square matrix and how can it be used to solve the system of equations $A\mathbf{x} = \mathbf{b}$?

(c) Describe the algorithm to obtain an LU factorization.

(d) By which factor does the number of operations increase to obtain an LU factorization if n is increased by a factor of 10?

(e) What needs to be done if during Gaussian Elimination or LU factorization a zero entry is encountered on the diagonal? Distinguish two different cases.

(f) Describe scaled and total pivoting. Explain why it is necessary under certain circumstances.

(g) Perform an LU factorization on the matrix arising from the system of equations given in (a).

Solution

(a) For the Gaussian Elimination we write the system of equations in the form

$$\left(\begin{array}{ccc|c} 5 & 10 & 9 & 4 \\ 10 & 26 & 26 & 10 \\ 15 & 54 & 66 & 27 \end{array}\right)$$

Subtracting two times the first row from the second row and three times the first row from the third row gives

$$\left(\begin{array}{ccc|c} 5 & 10 & 9 & 4 \\ 0 & 6 & 8 & 2 \\ 0 & 24 & 39 & 15 \end{array}\right)$$

Now subtracting four times the third row from the second row gives

$$\left(\begin{array}{ccc|c} 5 & 10 & 9 & 4 \\ 0 & 6 & 8 & 2 \\ 0 & 0 & 7 & 7 \end{array}\right)$$

Considering the last equation when back-substituting immediately gives $x_3 = 1$. Inserting this in the second equation yields

$$6x_2 + 8 = 2$$

and therefore $x_2 = -1$. Finally inserting x_2 and x_3 in the first equation

$$5x_1 - 10 + 9 = 4$$

and hence $x_1 = 1$.

(b) The LU factorization factorizes A into a lower triangular matrix L (i.e. $L_{i,j} = 0$ for $i < j$) and a upper triangular matrix U (i.e. $U_{i,j} = 0$ for $i > j$) such that is $A = LU$. The linear system then becomes $L(U\mathbf{x}) = \mathbf{b}$, which we decompose into $L\mathbf{y} = \mathbf{b}$ and $U\mathbf{x} = \mathbf{y}$. Both these systems can be solved easily by back substitution.

(c) The algorithm of the LU factorization is as follows:

(1) Set $A_0 := A$ and $k = 1$.

(2) Set the k-th row of U, $\mathbf{u}_k^T$, to the k-th row of A_{k-1} and the k-th column of L, $\mathbf{l}_k$, to the k-th column of A_{k-1}, scaled so that $L_{k,k} = 1$.

(3) Calculate $A_k := A_{k-1} - \mathbf{l}_k\mathbf{u}_k^T$ before incrementing k by 1 and returning to step (2) if $k \leq n$.

(d) The full LU factorization requires $O(n^3)$ operations. Thus if n increases by a factor of 10, The number of operations increases by a factor of 1000.

(e) If during Gaussian Elimination or LU factorization a zero entry is encountered on the diagonal (e.g. the (j, j) position), the row cannot be used to introduce zeros in that column below the diagonal. If there are non-zero entries in this column below the diagonal, a row with a nonzero entry is exchanged with the j-th row and this row is used to introduce zeros and as j-th row of U, $\mathbf{u}_j^T$. If there are only zeros on the diagonal and below the diagonal, then the system of equations has no unique solution. The variable x_j can be freely chosen. The LU factorization algorithm lets $\mathbf{l}_j$ be the j-th unit vector and $\mathbf{u}_j^T$ the current j-th row.

(f) Scaled pivoting considers the size of a coefficient relative to the other coefficients in the same equation. We can calculate the relative size by dividing each coefficient by the largest absolute value in that row. The row with the largest scaled coefficient is moved into the pivotal position.

In total (or complete or maximal) pivoting the pivotal equation and pivotal variable are selected by choosing the largest (unscaled) coefficient of any of the remaining variables. This is moved into the (k, k) position. This can involve exchange of columns as well as rows.

A small (compared to the other values) non-zero number as pivotal value is not suitable, since the row gets scaled by the reciprocal of this number. This can make small errors bigger which can accumulate introducing a large error in the end result.

(g) The matrix arising from the system of equations is

$$A = \begin{pmatrix} 5 & 10 & 9 \\ 10 & 26 & 26 \\ 15 & 54 & 66 \end{pmatrix}.$$

Thus

$$\mathbf{l}_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \mathbf{u}_1^T = \begin{pmatrix} 5 & 10 & 9 \end{pmatrix}, \mathbf{l}_1\mathbf{u}_1^T = \begin{pmatrix} 5 & 10 & 9 \\ 10 & 20 & 18 \\ 15 & 30 & 27 \end{pmatrix}$$

The next matrix to consider is

$$A - \mathbf{l}_1\mathbf{u}_1^T = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 6 & 8 \\ 0 & 24 & 39 \end{pmatrix}$$

Hence

$$\mathbf{l}_2 = \begin{pmatrix} 0 \\ 1 \\ 4 \end{pmatrix}, \mathbf{u}_2^T = \begin{pmatrix} 0 & 6 & 8 \end{pmatrix}, \mathbf{l}_2\mathbf{u}_2^T = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 6 & 8 \\ 0 & 24 & 32 \end{pmatrix}.$$

The last columns of L and U are

$$\mathbf{l}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \mathbf{u}_3^T = \begin{pmatrix} 0 & 0 & 7 \end{pmatrix}.$$

Summarizing

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix}, U = \begin{pmatrix} 5 & 10 & 9 \\ 0 & 6 & 8 \\ 0 & 0 & 7 \end{pmatrix}.$$

Note that we arrive at the same upper triangular matrix as when performing Gaussian Elimination. (Calculating L from the already found U is also an acceptable solution.)

2.15 EXERCISE 2.29

(a) Explain the technique of splitting for solving the linear system $A\mathbf{x} = \mathbf{b}$ iteratively where A is an $n \times n$, non-singular matrix. Define the iteration matrix H and state the property it has to satisfy to ensure convergence.

(b) Define the Gauss-Seidel and Jacobi iterations and state their iteration matrices respectively.

(c) Let

$$A = \begin{pmatrix} 2 & \frac{\sqrt{3}}{2} & \frac{1}{2} \\ \frac{\sqrt{3}}{2} & 2 & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} & 2 \end{pmatrix}$$

Derive the iteration matrix for the Jacobi *iterations and state the eigenvalue equation. Check that the numbers $-3/4, 1/4, 1/2$ satisfy the eigenvalue equation and thus are the eigenvalues of the iteration matrix.*

(d) The matrix given in (c) is positive definite. State the Householder-John theorem and apply it to show that the Gauss-Seidel *iterations for this matrix converge.*

(e) Describe relaxation and show how the the iteration matrix H_ω of the relaxed method is related to the iteration matrix H of the original method and thus how the eigenvalues are related. How should ω be chosen?

(f) For the eigenvalues given in (c) calculate the best choice of ω and the eigenvalues of the relaxed method.

Solution

(a) We can rewrite $A\mathbf{x} = \mathbf{b}$ in the form

$$(A - B)\mathbf{x} = -B\mathbf{x} + \mathbf{b},$$

where the matrix B is chosen in such a way that $A - B$ is non-singular and the system $(A-B)\mathbf{x} = \mathbf{y}$ is easily solved for any right hand side $\mathbf{y}$. A simple iterative scheme starts with an estimate $\mathbf{x}^{(0)} \in \mathbb{R}^n$ of the solution and generates the sequence $\mathbf{x}^{(k)}$, $k = 1, 2, \ldots$, by solving

$$(A - B)\mathbf{x}^{(k+1)} = -B\mathbf{x}^{(k)} + \mathbf{b}.$$

Let $\mathbf{x}^*$ solve $A\mathbf{x} = \mathbf{b}$. It also satisfies $(A-B)\mathbf{x}^* = -B\mathbf{x}^* + \mathbf{b}$. Subtracting this equation from the above equation gives

$$(A - B)(\mathbf{x}^{(k+1)} - \mathbf{x}^*) = -B(\mathbf{x}^{(k)} - \mathbf{x}^*).$$

We denote $\mathbf{x}^{(k)} - \mathbf{x}^*$ by $\mathbf{e}^{(k)}$. It is the error in the k-th iteration. Since $A - B$ is non-singular, we can write

$$\mathbf{e}^{(k+1)} = -(A - B)^{-1}B\mathbf{e}^{(k)}.$$

The matrix $H := -(A-B)^{-1}B$ is known as the *iteration matrix*. We have $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$ for all $\mathbf{x}^{(0)} \in \mathbb{R}^n$ if and only if $\rho(H) < 1$. That is all eigenvalues of H must have modulus less than 1.

(b) **Jacobi method** We choose $A - B = D$, the diagonal part of A, or in other words we let $B = L + U$. The iteration step is given by

$$D\mathbf{x}^{(k+1)} = -(L + U)\mathbf{x}^{(k)} + \mathbf{b}.$$

The iteration matrix is $-D^{-1}(L + U) = -D^{-1}(A - D)$.

Gauss–Seidel method We set $A - B = L + D$, the lower triangular portion of A, or in other words $B = U$. The sequence $\mathbf{x}^{(k)}, k = 1, \ldots,$ is generated by

$$(L + D)\mathbf{x}^{(k+1)} = -U\mathbf{x}^{(k)} + \mathbf{b}.$$

The iteration matrix is $-(L + D)^{-1}U$.

(c) For

$$A = \begin{pmatrix} 2 & \frac{\sqrt{3}}{2} & \frac{1}{2} \\ \frac{\sqrt{3}}{2} & 2 & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} & 2 \end{pmatrix}$$

the iteration matrix is given by

$$H = -\frac{1}{2}\begin{pmatrix} 0 & \frac{\sqrt{3}}{2} & \frac{1}{2} \\ \frac{\sqrt{3}}{2} & 0 & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 \end{pmatrix} = -\begin{pmatrix} 0 & \frac{\sqrt{3}}{4} & \frac{1}{4} \\ \frac{\sqrt{3}}{4} & 0 & \frac{\sqrt{3}}{4} \\ \frac{1}{4} & \frac{\sqrt{3}}{4} & 0 \end{pmatrix}.$$

The eigenvalue equation is

$$(-\lambda)^3 - \frac{\sqrt{3}}{4}\frac{\sqrt{3}}{4}\frac{1}{4} - \frac{1}{4}\frac{\sqrt{3}}{4}\frac{\sqrt{3}}{4} - \frac{1}{4}(-\lambda)\frac{1}{4} - \frac{\sqrt{3}}{4}\frac{\sqrt{3}}{4}(-\lambda) - \frac{\sqrt{3}}{4}\frac{\sqrt{3}}{4}(-\lambda)$$
$$= -\lambda^3 + \frac{7}{16}\lambda - \frac{3}{32} = 0.$$

For $-3/4, 1/4, 1/2$ we have

$$-(-\frac{3}{4})^3 + \frac{7}{16}(-\frac{3}{4}) - \frac{3}{32} = \frac{27}{4^3} - \frac{21}{4^3} - \frac{3}{32} = 0,$$
$$-(\frac{1}{4})^3 + \frac{7}{16}(\frac{1}{4}) - \frac{3}{32} = -\frac{1}{4^3} + \frac{7}{4^3} - \frac{3}{32} = 0,$$
$$-(\frac{1}{2})^3 + \frac{7}{16}(\frac{1}{2}) - \frac{3}{32} = -\frac{1}{8} + \frac{7}{32} - \frac{3}{32} = 0.$$

Thus $-3/4, -1/4, 1/2$ are the eigenvalues and the method converges.

(d)

Theorem (Householder–John theorem)**.** *If A and B are real matrices such that both A and $A - B - B^T$ are symmetric and positive definite, then the spectral radius of $H = -(A - B)^{-1}B$ is strictly less than one.*

The matrix A is symmetric and positive definite. For Gauss-Seidel $B = U$ and since A is symmetric, $B^T = U^T = L$. Thus $A - B - B^T = D$ and this is symmetric and positive definite, since the diagonal entries are positive.

(e) For relaxation, we first calculate $(A - B)\tilde{\mathbf{x}}^{(k+1)} = -B\mathbf{x}^{(k)} + \mathbf{b}$ as an intermediate value and then let

$$\mathbf{x}^{(k+1)} = \omega\tilde{\mathbf{x}}^{(k+1)} + (1 - \omega)\mathbf{x}^{(k)}$$

for $k = 0, 1, \ldots$, where $\omega \in \mathbb{R}$ is called the *relaxation parameter*.

Since $\tilde{\mathbf{x}}^{(k+1)} = -(A - B)^{-1}B\mathbf{x}^{(k)} + (A - B)^{-1}\mathbf{b}$, let $\mathbf{c} = (A - B)^{-1}\mathbf{b}$, the relaxation iteration matrix H_ω can then be deduced from

$$\mathbf{x}^{(k+1)} = \omega\tilde{\mathbf{x}}^{(k+1)} + (1 - \omega)\mathbf{x}^{(k)} = \omega H\mathbf{x}^{(k)} + (1 - \omega)\mathbf{x}^{(k)} + \omega\mathbf{c}$$

as

$$H_\omega = \omega H + (1 - \omega)I.$$

It follows that an eigenvalue λ of H is related to an eigenvalue λ_ω of H_ω by $\lambda_\omega = \omega\lambda + (1 - \omega)$. The best choice for ω would be to minimize $\max\{|\omega\lambda_i + (1 - \omega)|, i = 1, \ldots, n\}$ where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of H.

(f) With relaxation the eigenvalues become

$$-\tfrac{3}{4}\omega + 1 - \omega = 1 - \tfrac{7}{4}\omega$$

$$\tfrac{1}{4}\omega + 1 - \omega = 1 - \tfrac{3}{4}\omega$$

$$\tfrac{1}{2}\omega + 1 - \omega = 1 - \tfrac{1}{2}\omega$$

The best choice of ω is when $1 - \frac{1}{2}\omega = -(1 - \frac{7}{4}\omega)$ which gives $\omega = \frac{8}{9}$. In this case the eigenvalues are $-5/9, 1/3, 5/9$.

CHAPTER 3

Interpolation and Approximation Theory Exercises

3.1 EXERCISE 3.1

Let the function values $f(0), f(1), f(2)$ *and* $f(3)$ *be given. We want to estimate*

$$f(-1), f'(1) \text{ and } \int_0^3 f(x)dx.$$

To this end, we let p *be the cubic polynomial that interpolates these function values, and then approximate by*

$$p(-1), p'(1) \text{ and } \int_0^3 p(x)dx.$$

Using the Lagrange formula show that every approximation is a linear combination of the function values with constant coefficients and calculate these coefficients. Show that the approximations are exact if f *is any cubic polynomial.*

Solution

Let $f_0 = f(0), f_1 = f(1), f_2 = f(2)$ and $f_3 = f(3)$. Then

$$\begin{aligned} p(x) &= \sum_{k=0}^{3} \prod_{\substack{l=0 \\ l \neq k}}^{3} \frac{x - x_l}{x_k - x_l} f_k \\ &= f_0(1-x)\tfrac{1}{2}(2-x)\tfrac{1}{3}(3-x) \\ &\quad + f_1 x(2-x)\tfrac{1}{2}(3-x) \\ &\quad + f_2 \tfrac{1}{2} x(x-1)(3-x) \\ &\quad + f_3 \tfrac{1}{3} x \tfrac{1}{2}(x-1)(x-2) \\ &= \tfrac{1}{6} f_0(-x^3 + 6x^2 - 11x + 6) \\ &\quad + \tfrac{1}{2} f_1(x^3 - 5x^2 + 6x) \\ &\quad + \tfrac{1}{2} f_2(-x^3 + 4x^2 - 3x) \\ &\quad + \tfrac{1}{6} f_3(x^3 - 3x^2 + 2x). \end{aligned}$$

With these results we have

$$\begin{aligned} p(-1) &= \tfrac{1}{6} f_0(-(-1)^3 + 6(-1)^2 - 11(-1) + 6) \\ &\quad + \tfrac{1}{2} f_1((-1)^3 - 5(-1)^2 + 6(-1)) \\ &\quad + \tfrac{1}{2} f_2(-(-1)^3 + 4(-1)^2 - 3(-1)) \\ &\quad + \tfrac{1}{6} f_3((-1)^3 - 3(-1)^2 + 2(-1)) \\ &= 4f_0 - 6f_1 + 4f_2 - f_3. \end{aligned}$$

The derivative is

$$\begin{aligned} p'(x) &= \tfrac{1}{6} f_0(-3x^2 + 12x - 11) \\ &\quad + \tfrac{1}{2} f_1(3x^2 - 10x + 6) \\ &\quad + \tfrac{1}{2} f_2(-3x^2 + 8x - 3) \\ &\quad + \tfrac{1}{6} f_3(3x^2 - 6x + 2). \end{aligned}$$

with $p'(1) = -\frac{1}{3} f_0 - \frac{1}{2} f_1 + f_2 - \frac{1}{6} f_3$.

The integral is

$$\begin{aligned} \int_0^3 p(x) &= [\tfrac{1}{6} f_0(-\tfrac{1}{4}x^4 + 2x^3 - \tfrac{11}{2}x^2 + 6x) \\ &\quad + \tfrac{1}{2} f_1(\tfrac{1}{4}x^4 - \tfrac{5}{3}x^3 + 3x^2) \\ &\quad + \tfrac{1}{2} f_2(-\tfrac{1}{4}x^4 + \tfrac{4}{3}x^3 - \tfrac{3}{2}x^2) \\ &\quad + \tfrac{1}{6} f_3(\tfrac{1}{4}x^4 - x^3 + x^2)]_0^3 \\ &= \tfrac{3}{8} f_0 + \tfrac{9}{8} f_1 + \tfrac{9}{8} f_2 + \tfrac{3}{8} f_3. \end{aligned}$$

Note the symmetry in the last formula.

For the last part of the question let $f(x) = ax^3 + bx^2 + cx + d$ be a general

cubic polynomial. Then $f_0 = d$, $f_1 = a + b + c + d$, $f_2 = 8a + 4b + 2c + d$ and $f_3 = 27a + 9b + 3c + d$. Inserting these in the above formula we have

$$\begin{aligned} p(-1) &= 4d - 6(a+b+c+d) + 4(8a+4b+2c+d) - (27a+9b+3c+d) \\ &= -a + b - c + d = f(-1), \\ p'(1) &= -\tfrac{1}{3}d - \tfrac{1}{2}(a+b+c+d) + (8a+4b+2c+d) - \tfrac{1}{6}(27a+9b+3c+d) \\ &= 3a + 2b + c = f'(1) \end{aligned}$$

$$\begin{aligned} \int_0^3 p(x)dx &= \tfrac{3}{8}d + \tfrac{9}{8}(a+b+c+d) + \tfrac{9}{8}(8a+4b+2c+d) \\ &\quad + \tfrac{3}{8}(27a+9b+3c+d) \\ &= \tfrac{81}{4}a + 9b + \tfrac{9}{2}c + 3d = \textstyle\int_0^3 f(x)dx \end{aligned}$$

and thus the formulae are correct for any cubic polynomial.

3.2 EXERCISE 3.3

Let f be a real valued function and let p be the polynomial of degree at most n that interpolates f at the pairwise distinct points $x_0, x_1, \ldots, x_n$. Furthermore, let x be any real number that is not an interpolation point. Deduce for the error at x

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

(Hint: Use the definition for the divided difference $f[x_0, \ldots, x_n, x]$.)

Solution

By definition $f[x_0, \ldots, x_n, x]$ is the leading coefficient of the polynomial interpolating f at $x_0, x_1, \ldots, x_n$ and x. This coefficient is given by

$$f[x_0, \ldots, x_n, x] = \sum_{k=0}^{n} f(x_k) \left(\prod_{\substack{l=0 \\ l \neq k}}^{n} \frac{1}{x_k - x_l} \right) \frac{1}{x_k - x} + f(x) \prod_{l=0}^{n} \frac{1}{x - x_l}.$$

Multiplying both sides by $\prod_{l=0}^{n}(x - x_l)$ gives

$$f[x_0, \ldots, x_n, x] \prod_{l=0}^{n} (x - x_l) = \sum_{k=0}^{n} f(x_k) \left(- \prod_{\substack{l=0 \\ l \neq k}}^{n} \frac{x - x_l}{x_k - x_l} \right) + f(x) = -p(x) + f(x).$$

3.3 EXERCISE 3.5

Given $f, g \in C[a,b]$, let $h := fg$. Prove by induction that the divided differences of h satisfy the relation

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

By using the representation as derivatives of the differences and by letting the points $x_0,\ldots,x_n$ coincide, deduce the Leibniz formula for the n-th derivative of a product of two functions.

Solution

For $n = 0$ we have $h[x_0] = h(x_0) = f(x_0)g(x_0) = f[x_0]g[x_0]$. We may assume the assertion is true for any $n+1$ pairwise distinct points and hence for the $n+2$ points $x_0,\ldots,x_{n+1}$

$$\begin{aligned} h[x_0,\ldots,x_{n+1}] &= \frac{1}{x_{n+1}-x_0}\left(h[x_1,\ldots,x_{n+1}] - h[x_0,\ldots,x_n]\right) \\ &= \frac{1}{x_{n+1}-x_0}\left(\sum_{j=1}^{n+1} f[x_1,\ldots,x_j]g[x_j,\ldots,x_{n+1}]\right. \\ &\quad \left. -\sum_{j=0}^{n} f[x_0,\ldots,x_j]g[x_j,\ldots,x_n]\right) \\ &= \frac{1}{x_{n+1}-x_0}\sum_{j=0}^{n}\left(f[x_1,\ldots,x_{j+1}]g[x_{j+1},\ldots,x_{n+1}]\right. \\ &\quad \left. -f[x_0,\ldots,x_j]g[x_j,\ldots,x_n]\right) \end{aligned}$$

Now $f[x_1,\ldots,x_{j+1}] = (x_{j+1}-x_0)f[x_0,\ldots,x_{j+1}] + f[x_0,\ldots,x_j]$ and $g[x_j,\ldots,x_n] = g[x_{j+1},\ldots,x_{n+1}] - (x_{n+1}-x_j)g[x_j,\ldots,x_{n+1}]$. Inserting these into the sum, we obtain

$$\begin{aligned} h[x_0,\ldots,x_{n+1}] &= \frac{1}{x_{n+1}-x_0}\sum_{j=0}^{n}\left((x_{j+1}-x_0)f[x_0,\ldots,x_{j+1}]g[x_{j+1},\ldots,x_{n+1}]\right. \\ &\quad +f[x_0,\ldots,x_j]g[x_{j+1},\ldots,x_{n+1}] - f[x_0,\ldots,x_j]g[x_{j+1},\ldots,x_{n+1}] \\ &\quad \left.+(x_{n+1}-x_j)f[x_0,\ldots,x_j]g[x_j,\ldots,x_{n+1}]\right) \\ &= \sum_{j=0}^{n+1} f[x_0,\ldots,x_j]g[x_j,\ldots,x_{n+1}] \end{aligned}$$

If $x_0,\ldots,x_n$ coincide in ξ we have

$$h[x_0,\ldots,x_n] = \frac{1}{n!}h^{(n)}(\xi) = \sum_{j=0}^{n}\frac{1}{j!}f^{(j)}(\xi)\frac{1}{(n-j)!}g^{(n-j)}(\xi).$$

3.4 EXERCISE 3.7

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

$$p_n(x) = e^x \frac{d^n}{dx^n}(x^n e^{-x}), \qquad x \in \mathbb{R}^+.$$

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 given by

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

Thus these polynomials are the Laguerre polynomials. Calculate p_3, p_4 and p_5 from Rodrigues formula.

Solution

The k-th derivative, $k \le n$, of $x^n e^{-x}$ is given by

$$\frac{d^k}{dx^k}(x^n e^{-x}) = x^{n-k} q_k(x) e^{-x}$$

for some polynomial q_k of degree k. This is obviously true for $k = 0$ with $q_0(x) \equiv 1$. Assuming this is true for $k < n$, we can write for $k+1$

$$\begin{aligned}
\frac{d^{k+1}}{dx^{k+1}}(x^n e^{-x}) &= \frac{d}{dx}(x^{n-k} q_k(x) e^{-x}) \\
&= (n-k) x^{n-k-1} q_k(x) e^{-x} + x^{n-k} q_k'(x) e^{-x} - x^{n-k} q_k(x) e^{-x} \\
&= x^{n-k-1}((n-k) q_k(x) + x q_k'(x) - x q_k(x)) e^{-x}.
\end{aligned}$$

Thus $q_{k+1} = (n-k) q_k(x) + x q_k'(x) - x q_k(x)$. Letting $k = n$, it follows that

$$p_n(x) = e^x \frac{d^n}{dx^n}(x^n e^{-x}) = e^x x^{n-n} q_n(x) e^{-x} = q_n(x)$$

is a polynomial of degree n.

Lets look at the scalar product now:

$$\begin{aligned}
\langle p_n, p\rangle &= \int_0^\infty e^{-x} e^x \frac{d^n}{dx^n}(x^n e^{-x}) p(x) dx \\
&= \left[\frac{d^{n-1}}{dx^{n-1}}(x^n e^{-x}) p(x)\right]_0^\infty - \int_0^\infty \frac{d^{n-1}}{dx^{n-1}}(x^n e^{-x}) p'(x) dx \\
&= \left[x q_{n-1}(x) e^{-x} p(x)\right]_0^\infty - \int_0^\infty \frac{d^{n-1}}{dx^{n-1}}(x^n e^{-x}) p'(x) dx \\
&= -\int_0^\infty \frac{d^{n-1}}{dx^{n-1}}(x^n e^{-x}) p'(x) dx \\
&= \ldots \\
&= \pm\left[x^n q_0(x) e^{-x} p^{(n-1)}(x)\right]_0^\infty \mp \int_0^\infty x^n q_0(x) e^{-x} p^{(n)}(x) dx
\end{aligned}$$

Now $p \in \mathbb{P}_{n-1}[x]$ and thus its n-th derivative vanishes and hence the scalar product is zero.

The explicit polynomials are

$$\begin{aligned} p_3(x) &= -x^3 + 9x^2 - 18x + 6 \\ p_4(x) &= x^4 - 16x^3 + 72x^2 - 96x + 24 \\ p_5(x) &= -x^5 + 25x^4 - 200x^3 + 600x^2 - 600x + 120. \end{aligned}$$

3.5 EXERCISE 3.9

Express the divided difference $f[0,1,2,3]$ in the form

$$L(f) = f[0,1,2,3] = \frac{1}{2}\int_0^3 K(\theta) f'''(\theta) d\theta,$$

assuming that $f \in C^3[0,3]$. Sketch the kernel function $K(\theta)$ for $\theta \in [0,3]$. By integrating $K(\theta)$ and using the mean value theorem show that

$$f[0,1,2,3] = \frac{1}{6} f'''(\xi)$$

for some point $\xi \in [0,3]$.

Solution

Set

$$L(f) = f[0,1,2,3] = -\frac{1}{6}f(0) + \frac{1}{2}f(1) - \frac{1}{2}f(2) + \frac{1}{6}f(3).$$

The Peano kernel is then given by

$$\begin{aligned} K(\theta) &= L[(x-\theta)_+^2] \\ &= -\frac{1}{6}(0-\theta)_+^2 + \frac{1}{2}(1-\theta)_+^2 - \frac{1}{2}(2-\theta)_+^2 + \frac{1}{6}(3-\theta)_+^2 \\ &= \begin{cases} \frac{1}{6}\theta^2 & 0 \le \theta \le 1 \\ -\frac{1}{3}\theta^2 + \theta - \frac{1}{2} & 1 \le \theta \le 2 \\ \frac{1}{6}\theta^2 - \theta + \frac{3}{2} & 2 \le \theta \le 3 \\ 0 & \theta \le 0 \text{ or } \theta \ge 3. \end{cases} \end{aligned}$$

Thus $K(\theta)$ is piecewise quadratic, continuous and $K(\theta) \ge 0$. Thus we can

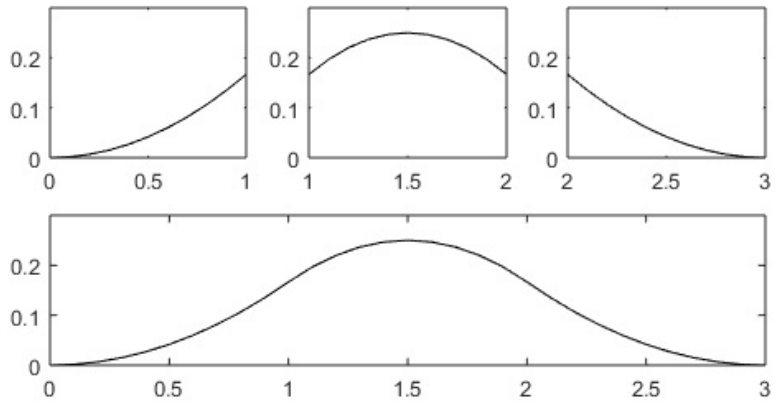


Figure 3.1 Plot of the piecewise quadratic Peano kernel

use the result of Theorem 3.11. The value of the integral of $K(\theta)$ over $[0,3]$ is

$$\begin{aligned}\int_0^3 K(\theta)d\theta &= \int_0^1 \frac{1}{6}\theta^2 d\theta + \int_1^2 -\frac{1}{3}\theta^2 + \theta - \frac{1}{2}d\theta + \int_2^3 \frac{1}{6}\theta^2 - \theta + \frac{3}{2}d\theta \\ &= \left[\frac{1}{18}\theta^3\right]_0^1 + \left[-\frac{1}{9}\theta^3 + \frac{1}{2}\theta^2 - \frac{1}{2}\theta\right]_1^2 + \left[\frac{1}{18}\theta^3 - \frac{1}{2}\theta^2 + \frac{3}{2}\theta\right]_2^3 \\ &= \frac{1}{18} + \left(-\frac{8}{9} + 2 - 1\right) - \left(-\frac{1}{9} + \frac{1}{2} - \frac{1}{2}\right) \\ &\quad + \left(\frac{3}{2} - \frac{9}{2} + \frac{9}{2}\right) - \left(\frac{4}{9} - 2 + 3\right) = \frac{6}{18} = \frac{1}{3}.\end{aligned}$$

Therefore $f[0,1,2,4] = \frac{1}{2}\frac{1}{3}f'''(\xi) = \frac{1}{6}f'''(\xi)$ for some $\xi \in [0,3]$.

3.6 EXERCISE 3.11

Let S be the set of cubic splines with knots $x_i = ih$ for $i = 0,\ldots,n$, where $h = 1/n$. An inexperienced user obtains an approximation to a twice differentiable function f by satisfying the conditions $s'(0) = f'(0)$, $s''(0) = f''(0)$ and $s(x_i) = f(x_i)$, $i = 0,\ldots,n$. Show how the changes in the first derivatives $s'(x_i)$ propagate if $s'(0)$ is increased by a small perturbation ϵ, i.e. $s'(0) = f'(0) + \epsilon$, but the remaining data remain the same.

Solution

The change is itself a cubic spline which satisfies $s(x_i) = 0$, $i = 0,\ldots,n$, $s''(0) = 0$ and $s'(0) = \epsilon$. A cubic spline which vanishes at x_i, $i = 0,\ldots,n$, has the derivatives

$$s'(x_i) = \alpha(-2 + \sqrt{3})^i + \beta(-2 - \sqrt{3})^i,$$

where α and β are determined by the remaining two conditions $s''(0) = 0$ and $s'(0) = \epsilon$. The latter yields $\alpha + \beta = s'(x_0) = s'(0) = \epsilon$. The second derivative at the point $x_0 = 0$ takes the value

$$\frac{4}{h}(s'(x_0) + s'(x_1)) = \frac{4}{h}(\alpha + \beta + \alpha(-2 + \sqrt{3}) + \beta(-2 - \sqrt{3})) = 0.$$

From this $\beta = (-1 + \sqrt{3})/(1 + \sqrt{3})\alpha$ follows. Using $\alpha + \beta = \epsilon$, it follows that

$$\alpha = \frac{\sqrt{3}+1}{2\sqrt{3}}\epsilon \text{ and } \beta = \frac{\sqrt{3}-1}{2\sqrt{3}}\epsilon.$$

Thus the change in the derivatives at $x_0, \ldots, x_n$ is

$$s'(x_i) = \frac{\sqrt{3}+1}{2\sqrt{3}}\epsilon(-2 + \sqrt{3})^i + \frac{\sqrt{3}-1}{2\sqrt{3}}\epsilon(-2 - \sqrt{3})^i,$$

Now $-2 + \sqrt{3} \approx -0.27$, so this component decreases with increasing i. However, $-2 - \sqrt{3} \approx -3.7$ and this component increases with increasing i. Moreover, because it is negative the first derivative changes sign at every knot. Thus the change begins to oscillate widely with increasing i.

3.7 EXERCISE 3.13

(a) Let Q_k, $k = 0, 1, \ldots$, be a set of polynomials orthogonal with respect to some inner product $\langle \cdot, \cdot \rangle$ in the interval $[a, b]$. Let f be a continuous function in $[a, b]$. Write explicitly the least-squares polynomial approximation to f by a polynomial of degree n in terms of the polynomials Q_k, $k = 0, 1, \ldots$.

(b) Let an inner product be defined by the formula

$$\langle g, h \rangle = \int_{-1}^{1} (1 - x^2)^{-1/2} g(x)h(x)dx.$$

The orthogonal polynomials are the Chebyshev polynomials of the first kind given by $Q_k(x) = \cos(k \arccos x), k \geq 0$. Using the substitution $x = \cos\theta$, calculate the inner products $\langle Q_k, Q_k \rangle$ for $k \geq 0$. (Hint: $2\cos^2 x = 1 + \cos 2x$.)

(c) For the inner product given above and the Chebyshev polynomials calculate the inner products $\langle Q_k, f \rangle$ for $k \geq 0$, $k \neq 1$, where f is given by $f(x) = (1 - x^2)^{1/2}$. (Hint: $\cos x \sin y = \frac{1}{2}[\sin(x + y) - \sin(x - y)]$.)

(d) Now for $k = 1$, calculate the inner product $\langle Q_1, f \rangle$.

(e) Thus for even n write the least squares polynomial approximation to f as linear combination of the Chebyshev polynomials with the correct coefficients.

Solution

(a) Let $p = \sum_{k=0}^{n} c_k Q_k$ be a general n-th degree polynomial expressed in the basis Q_k. The least-squares approximation minimizes the inner product $\langle f - p, f - p \rangle$. Because of orthogonality the inner product simplifies to

$$\begin{aligned} \langle f - p, f - p \rangle &= \left\langle f - \sum_{k=0}^{n} c_k Q_k, f - \sum_{j=0}^{n} c_j Q_j \right\rangle \\ &= \langle f, f \rangle - 2 \sum_{k=0}^{n} c_k \langle f, Q_k \rangle + \sum_{k=0}^{n} c_k^2 \langle Q_k, Q_k \rangle. \end{aligned}$$

This is a quadratic function in the c_ks and we can minimize it to find optimal values for the c_ks. Differentiating with respect to c_k gives

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

Setting the gradient to zero, we obtain

$$c_k = \frac{\langle Q_k, f \rangle}{\langle Q_k, Q_k \rangle}$$

and thus

$$p = \sum_{k=0}^{n} \frac{\langle Q_k, f \rangle}{\langle Q_k, Q_k \rangle} Q_k.$$

(b) The inner product $\langle Q_k, Q_k \rangle$ for $k \neq 0$ is given by

$$\langle Q_k, Q_k \rangle = \int_{-1}^{1} (1 - x^2)^{-1/2} Q_k^2(x) dx = \int_{-1}^{1} (1 - x^2)^{-1/2} \cos^2(k \arccos x) dx.$$

Using the substitution $x = \cos\theta$ with $\frac{dx}{d\theta} = \sin\theta$ the range of integration changes to $[0, \pi]$. Therefore

$$\begin{aligned} \langle Q_k, Q_k \rangle &= \int_0^{\pi} (1 - \cos^2\theta)^{-1/2} \cos^2 k\theta \sin\theta d\theta \\ &= \int_0^{\pi} \cos^2 k\theta d\theta \\ &= \int_0^{\pi} \frac{1}{2}(1 + \cos 2k\theta) d\theta = \frac{\pi}{2}, \end{aligned}$$

since the second term in the sum integrates to 0.

For $k = 0$ we have

$$\langle Q_0, Q_0 \rangle = \int_{-1}^{1} (1 - x^2)^{-1/2} Q_0^2(x) dx = \int_{-1}^{1} (1 - x^2)^{-1/2} dx.$$

Using the same substitution we obtain

$$\langle Q_0, Q_0 \rangle = \int_0^{\pi} (1 - \cos^2\theta)^{-1/2} \sin\theta d\theta = \pi.$$

(c) For $f(x) = (1-x^2)^{1/2}$ the inner product is

$$\begin{aligned}\langle Q_k, f\rangle &= \int_{-1}^{1} (1-x^2)^{-1/2}(1-x^2)^{1/2} Q_k(x)dx \\ &= \int_{-1}^{1} Q_k(x)dx = \int_{-1}^{1} \cos(k \arccos x)dx.\end{aligned}$$

Again substituting $x = \cos\theta$ gives

$$\begin{aligned}\langle Q_k, f\rangle &= \int_0^{\pi} \cos k\theta \sin\theta d\theta \\ &= \frac{1}{2}\int_0^{\pi} [\sin(k+1)\theta - \sin(k-1)\theta]d\theta \\ &= \frac{1}{2}\left[-\frac{1}{k+1}\cos(k+1)\theta + \frac{1}{k-1}\cos(k-1)\theta\right]_0^{\pi}\end{aligned}$$

For odd k this evaluates to zero, since the expression in the square bracket evaluates to the same for $\theta = \pi$ and $\theta = 0$, because we always have an even multiple of π. For even k, π is multiplied by an odd number and thus

$$\langle Q_k, f\rangle = \frac{1}{k+1} - \frac{1}{k-1} + \frac{1}{k+1} - \frac{1}{k-1} = \frac{-2}{k^2-1}.$$

(d) For $k = 1$, the above calculation would lead to a division by zero. In this case $\sin(k-1)\theta = 0$ and thus

$$\langle Q_1, f\rangle = \frac{1}{2}\int_0^{\pi} \sin 2\theta d\theta = 0$$

(e) To summarize the linear least squares approximation to $f(x) = (1-x^2)^{1/2}$ is

$$\frac{2}{\pi}Q_0(x) + \sum_{m=1}^{n/2} \frac{-4}{\pi(4m^2-1)} Q_{2m}(x).$$

3.8 EXERCISE 3.15

(a) Define the divided difference of degree n, $f[x_0, x_1, \ldots, x_n]$. What is the divided difference of degree zero?

(b) Prove the recursive formula for divided differences

$$f[x_0, x_1, \ldots, x_k, x_{n+1}] = \frac{f[x_1, \ldots, x_{n+1}] - f[x_0, \ldots, x_n]}{x_{n+1} - x_0}.$$

(c) By considering the polynomials $p, q \in \mathbb{P}_k[x]$ *that interpolate* f *at* $x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ *and* $x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n$ *respectively, where* $i \neq j$*, construct a polynomial* r*, which interpolates* f *at* $x_0, \ldots, x_n$*. For the constructed* r *show that* $r(x_k) = f(x_k)$ *for* $k = 0, \ldots, n$*.*

(d) Deduce that, for any $i \neq j$*, we have*

$$f[x_0, \ldots, x_n] = \frac{f[x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n] - f[x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n]}{x_j - x_i}.$$

(e) Calculate the divided difference table for $x_0 = 0, x_1 = 1, x_2 = 2$ *and* $x_3 = 3$ *with data values* $f_0 = 0, f_1 = 1, f_2 = 8$ *and* $f_3 = 27$*.*

(f) Using the above formula, calculate the divided differences $f[x_0, x_2]$*,* $f[x_0, x_2, x_3]$ *and* $f[x_0, x_1, x_3]$*.*

Solution

(a) Let p interpolate $f_0, f_1, \ldots, f_n$. The polynomial p is unique and the coefficient of x^n in p is called the divided difference of degree n: $f[x_0, x_1, \ldots, x_n]$. The divided difference of degree zero is the coefficient of the zero degree interpolating polynomial, i.e. a constant. Hence $f[x_i] = f(x_i)$.

(b) The recursive formula for the divided differences can be proven in the following way: Let $p, q \in \mathbb{P}_n[x]$ be the polynomials that interpolate f at $x_0, \ldots, x_n$ and $x_1, \ldots, x_{n+1}$ respectively. Let

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

It can be easily seen that $r(x_i) = f(x_i)$ for $i = 0, \ldots, n+1$. Hence r is the unique interpolating polynomial of degree $n+1$ and the coefficient of x^{n+1} in r is given by the formula

$$f[x_0, x_1, \ldots, x_n, x_{n+1}] = \frac{f[x_1, \ldots, x_{n+1}] - f[x_0, \ldots, x_n]}{x_{n+1} - x_0}.$$

(c) For polynomials p, q that interpolate f at $x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ and $x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n$ respectively, we let

$$r(x) := \frac{x - x_i}{x_j - x_i}p(x) + \frac{x_j - x}{x_j - x_i}q(x).$$

For $x = x_i$, we have

$$r(x_i) := \frac{x_i - x_i}{x_j - x_i}p(x_i) + \frac{x_j - x_i}{x_j - x_i}q(x_i) = q(x_i) = f(x_i).$$

For $x = x_j$, we have

$$r(x_j) := \frac{x_j - x_i}{x_j - x_i} p(x_j) + \frac{x_j - x_j}{x_j - x_i} q(x_j) = p(x_j) = f(x_j).$$

For $x = x_k$, $k \neq i, k \neq j$, we have

$$r(x_k) := \frac{x_k - x_i}{x_j - x_i} p(x_k) + \frac{x_j - x_k}{x_j - x_i} q(x_k) = \frac{x_k - x_i + x_j - x_k}{x_j - x_i} f(x_k) = f(x_k).$$

Thus r as above interpolates f at $x_0, \ldots, x_n$.

(d) Since r interpolates f at $x_0, \ldots, x_n$, $f[x_0, \ldots, x_n]$ is the coefficient of x^n in r. This however is the difference of the coefficient of x^{n-1} in p and in q divided by $x_j - x_i$. On the other hand the coefficient of x^{n-1} in p and in q are the divided differences $f[x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$ and $f[x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n]$. It follows that

$$f[x_0, \ldots, x_n] = \frac{f[x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n] - f[x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n]}{x_j - x_i}.$$

For $i = 0$ and $j = n$, we recover the original recursive formula

$$f[x_0, x_1, \ldots, x_n, x_{n+1}] = \frac{f[x_1, \ldots, x_{n+1}] - f[x_0, \ldots, x_n]}{x_{n+1} - x_0}.$$

(e) The divided difference table for $x_0 = 0, x_1 = 1, x_2 = 2$ and $x_3 = 3$ with data values $f_0 = 0, f_1 = 1, f_2 = 8$ and $f_3 = 27$ is

$$\begin{array}{ccccccc}
f[x_0] & & & & & & \\
 & \searrow & & & & & \\
 & & f[x_0, x_1] & & & & \\
f[x_1] & \nearrow & & \searrow & & & \\
 & \searrow & & & f[x_0, x_1, x_2] & & \\
 & & f[x_1, x_2] & \nearrow & & \searrow & \\
f[x_2] & \nearrow & & \searrow & & & f[x_0, x_1, x_2, x_3]. \\
 & \searrow & & & f[x_1, x_2, x_3] & \nearrow & \\
 & & f[x_2, x_3] & \nearrow & & & \\
f[x_3] & \nearrow & & & & &
\end{array}$$

$$\begin{array}{ccccccc}
0 & & & & & & \\
 & \searrow & & & & & \\
 & & \frac{1-0}{1-0} = 1 & & & & \\
1 & \nearrow & & \searrow & & & \\
 & \searrow & & & \frac{7-1}{2-0} = 3 & & \\
 & & \frac{8-1}{2-1} = 7 & \nearrow & & \searrow & \\
8 & \nearrow & & \searrow & & & \frac{6-3}{3-0} = 1. \\
 & \searrow & & & \frac{19-7}{3-1} = 6 & \nearrow & \\
 & & \frac{27-8}{3-2} = 19 & \nearrow & & & \\
27 & \nearrow & & & & &
\end{array}$$

(f) Using

$$f[x_0, \ldots, x_n] = \frac{f[x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n] - f[x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n]}{x_j - x_i},$$

we first let $n = 2$, $i = 2$ and $j = 1$ to get

$$f[x_0, x_1, x_2] = \frac{f[x_0, x_1] - f[x_0, x_2]}{x_1 - x_2}.$$

Solving for $f[x_0, x_2]$, we arrive at

$$f[x_0, x_2] = f[x_0, x_1] - (x_1 - x_2)f[x_0, x_1, x_2] = 1 - (1 - 2)3 = 4.$$

Next we let $n = 3$, $i = 3$ and $j = 1$ to get

$$f[x_0, x_1, x_2, x_3] = \frac{f[x_0, x_1, x_2] - f[x_0, x_2, x_3]}{x_1 - x_3}.$$

Solving for $f[x_0, x_2, x_3]$, we arrive at

$$f[x_0, x_2, x_3] = f[x_0, x_1, x_2] - (x_1 - x_3)f[x_0, x_1, x_2, x_3] = 3 - (1 - 3)1 = 5.$$

Last we let $n = 3$, $i = 3$ and $j = 2$ to get

$$f[x_0, x_1, x_2, x_3] = \frac{f[x_0, x_1, x_2] - f[x_0, x_1, x_3]}{x_2 - x_3}.$$

Solving for $f[x_0, x_1, x_3]$, we arrive at

$$f[x_0, x_1, x_3] = f[x_0, x_1, x_2] - (x_2 - x_3)f[x_0, x_1, x_2, x_3] = 3 - (2 - 3)1 = 4.$$

3.9 EXERCISE 3.17

(a) Given a set of real values $f_0, f_1, \ldots, f_n$ at real data points $x_0, x_1, \ldots, x_n$, give a formula for the Lagrange cardinal polynomials and state their properties. Write the polynomial interpolant in the Lagrange form.

(b) How many operations are necessary to evaluate the polynomial interpolant in the Lagrange form at x?

(c) Prove that the polynomial interpolant is unique.

(d) Using the Lagrange form of interpolation, compute the polynomial $p(x)$ that interpolates the data $x_0 = 0$, $x_1 = 1$, $x_2 = 2$ and $f_0 = 1$, $f_1 = 2$, $f_2 = 3$. What is the degree of $p(x)$?

(e) What is a divided difference and a divided difference table and for which form of interpolant is it used? Give the formula for the interpolant. How many operations are necessary to evaluate the polynomial in this form?

(f) Prove the relation used in a divided difference table.

(g) Write down the divided difference table for the interpolation problem given in (d). How does it change with the additional data $f_3 = 5$ at $x_3 = 3$.

Solution

(a) For $k = 0, \ldots, n$ the Lagrange cardinal polynomials are

$$L_k(x) := \prod_{\substack{l=0 \\ l \neq k}}^{n} \frac{x - x_l}{x_k - x_l}, \qquad x \in \mathbb{R}.$$

The k-th Lagrange cardinal polynomial is monic and unique, has degree n and has the property $L_k(x_k) = 1$ and $L_k(x_j) = 0$ for $j \neq k$. The polynomial interpolant is given by

$$p(x) = \sum_{k=0}^{n} f_k L_k(x) = \sum_{k=0}^{n} f_k \prod_{\substack{l=0 \\ l \neq k}}^{n} \frac{x - x_l}{x_k - x_l}.$$

(b) Even if the products $\prod_{\substack{l=0 \\ l \neq k}}^{n} \frac{1}{x_k - x_l}$ are precalculated, to evaluate $p(x)$ at x requires for each data f_k, $k = 1, \ldots, n$, $n - 1$ subtractions and $n - 2$ multiplications. Thus the number of operations is $O(n^2)$.

(c) Uniqueness is proven in the following way. Suppose that two polynomials $p, q \in \mathbb{P}_n[x]$ satisfy $p(x_i) = q(x_i) = f_i$, $i = 0, \ldots, n$. Then the n-th degree polynomial $p - q$ vanishes at $n + 1$ distinct points. However, the only n-th degree polynomial with $n + 1$ or more zeros is the zero polynomial. Therefore $p = q$.

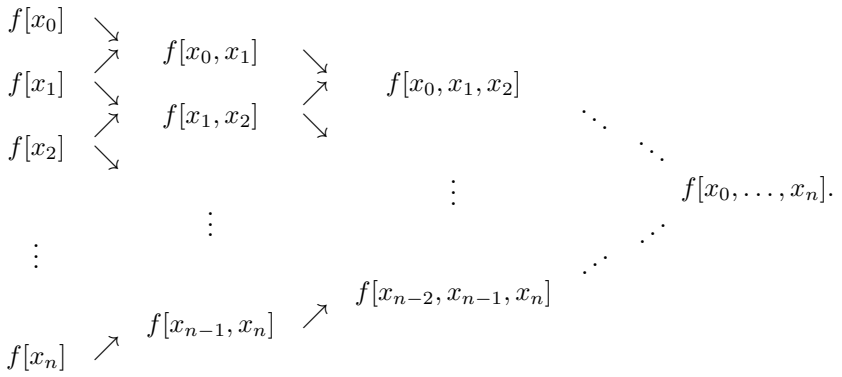
(d) For $x_0 = 0$, $x_1 = 1$, $x_2 = 2$ and $f_0 = 1$, $f_1 = 2$, $f_2 = 3$, the polynomial is

$$\begin{aligned} p(x) &= f_0 \frac{x - x_1}{x_0 - x_1} \frac{x - x_2}{x_0 - x_2} + f_1 \frac{x - x_0}{x_1 - x_0} \frac{x - x_2}{x_1 - x_2} + f_2 \frac{x - x_0}{x_2 - x_0} \frac{x - x_1}{x_2 - x_1} \\ &= 1 \frac{x - 1}{-1} \frac{x - 2}{-2} + 2 \frac{x}{1} \frac{x - 2}{-1} + 3 \frac{x}{2} \frac{x - 1}{1} \\ &= \frac{1}{2}(x - 1)(x - 2) - 2x(x - 2) + \frac{3}{2} x(x - 1) \\ &= x + 1. \end{aligned}$$

The degree of $p(x)$ is one.

(e) Given pairwise distinct points $x_0, x_1, \ldots, x_n \in [a, b]$, let $p \in \mathbb{P}_n[x]$ interpolate $f \in C^n[a, b]$ at these points. The coefficient of x^n in p is called the divided difference of degree n and denoted by $f[x_0, x_1, \ldots, x_n]$. The

divided difference table is



The divided difference table is used for the Newton form of interpolant. This has the form

$$p(x) := f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_{n-1}] \prod_{i=0}^{n-2} (x - x_i) + f[x_0, \ldots, x_n] \prod_{i=0}^{n-1} (x - x_i).$$

Provided that the divided differences are known, it can be evaluated at a given point x in just $O(n)$ operations as long as we employ the nested multiplication as in the Horner scheme

$$p_n(x) = \{\cdots\{\{f[x_0, \ldots, x_n](x - x_{n-1}) + f[x_0, \ldots, x_{n-1}]\} \times (x - x_{n-2}) + f[x_0, \ldots, x_{n-2}]\} \times (x - x_{n-3}) + \cdots\} + f[x_0].$$

(f) The divided difference table relies on the recurrence relation of divided differences

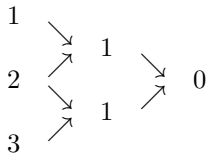
$$f[x_0, x_1, \ldots, x_k, x_{k+1}] = \frac{f[x_1, \ldots, x_{k+1}] - f[x_0, \ldots, x_k]}{x_{k+1} - x_0}.$$

This is proven by considering the the polynomials $p(x)$ and $q(x)$ that interpolate f at $x_0, \ldots, x_k$ and $x_1, \ldots, x_{k+1}$ respectively. Let

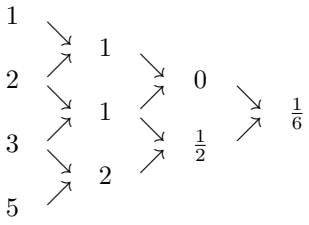
$$r(x) := \frac{(x - x_0)q(x) + (x_{k+1} - x)p(x)}{x_{k+1} - x_0} \in \mathbb{P}_{k+1}[x].$$

It can be easily seen that $r(x_i) = f(x_i)$ for $i = 0, \ldots, k + 1$. Hence r is the unique interpolating polynomial of degree $k + 1$ and the coefficient of x^{k+1} in r is given by the formula above.

(g)



With the additional data the difference table becomes



CHAPTER 4

Non-linear Systems Exercises

4.1 EXERCISE 4.1

Write a program which takes a polynomial of degree between 2 *and* 7 *as input, applies Newton's root finding method and colours the basins of attraction for each root a different colour. Try it out for the polynomial* $z^n - 1$ *for* $n = 2, \ldots, 7$.

Solution

```
function img = NewtonFractal(P)
% Calculates the basins of attractions when the Newton method for root
% finding is used on a polynomial of degree at most 7.
% P input argument, vector specifying the coefficents of the
% polynomial starting with the coefficient of the highest power
% img output argument, each pixel is coloured according to the root it
% converges to and shaded by the number of iterations necessary, the
% more iterations, the darker shade
NITER = 100;        % maximum number of iterations
threshold = .001;   % convergence criterion
pixelnum = 1000;    % resolution of image
colorArr = [7,3];   % seven colours specified by their RGB values
%RED
colorArr(1,1) = 1;colorArr(1,2) = 0;colorArr(1,3) = 0;
%GREEN
colorArr(2,1) = 0;colorArr(2,2) = 1;colorArr(2,3) = 0;
%BLUE
colorArr(3,1) = 0;colorArr(3,2) = 0;colorArr(3,3) = 1;
%YELLOW
colorArr(4,1) = 1;colorArr(4,2) = 1;colorArr(4,3) = 0;
%WHITE
colorArr(5,1) = 1;colorArr(5,2) = 0;colorArr(5,3) = 1;
%CYAN
```

```
colorArr(6,1) = 0;colorArr(5,2) = 1;colorArr(5,3) = 1;
%RED
colorArr(7,1) = 1;colorArr(6,2) = 1;colorArr(6,3) = 1;

% generate grid over the square [-1, 1] x [-1, 1]
[xs,ys] = meshgrid(linspace(-1,1,pixelnum), linspace(-1,1,pixelnum));
% grid points interpreted as complex numbers, which is an array of
% length pixelnum * pixelnum
solutions = xs(:) + 1i*ys(:);
% array of indices of grid points under consideration
select = 1:numel(xs);
% for each grid point initialise the necessary number of iterations
% to the maximum
niters = NITER*ones(numel(xs), 1);

% calculate the roots of the polynomial
Proots = roots(P);
if isempty(Proots)
    disp('Polynomial has no roots');
    return;
end
% calculate the coefficients of the derivative
Pderivative = zeros(length(P) - 1,1);
for it = 1:length(P)-1
    Pderivative(it)=(length(P)-it)*P(it);
end

for iteration = 1:NITER
    % each iteration considers the entire grid minus the grid points
    % where convergence has occured
    oldi = solutions(select);

    % in newton's method we have z_{i+1} = z_i - p(z_i) / p'(z_i)
    solutions(select) = oldi - polyval(P,oldi) ...
        ./ polyval(Pderivative,oldi);

    % check for convergence or NaN (in case of a division by zero)
    differ = (oldi - solutions(select));
    % logical array marking converged grid points
    converged = abs(differ) < threshold;
    % logical array marking problematic grid points
    problematic = isnan(differ);

    % if converence occured update the necessary number of iterations
    niters(select(converged)) = iteration;
    % for problematic grid points set the number of iterations to the
    % maximum + 1
    niters(select(problematic)) = NITER+1;
    %remove indices of converged or problematic points
    select(converged | problematic) = [];
end

Max = max(niters);
niters = reshape(niters,size(xs));
solutions = reshape(solutions,size(xs));

A = zeros(pixelnum, pixelnum);
```

```
B = uint8(round(A * 255));
RowCol = size(solutions);
rows = RowCol(1);
cols = RowCol(2);
for i1 = 1:rows
    for i2 =1:cols
        % to which root did the method converge
        tmp = abs(repmat(solutions(i1,i2), size(Proots))-Proots);
        rootIndex = find(tmp<threshold);
        if ~isempty(rootIndex)
        % color associated with roots and rate of convergence
        B(i1,i2,1)=colorArr(rootIndex,1) * ...
            (1-(niters(i1,i2)/(Max))) * 255;
        B(i1,i2,2)=colorArr(rootIndex,2) * ...
            (1-(niters(i1,i2)/(Max))) * 255;
        B(i1,i2,3)=colorArr(rootIndex,3) * ...
            (1-(niters(i1,i2)/(Max))) * 255;
        end
    end
end
```

Figure 4.1 shows the basins of attraction for Newton's method applied to $z^n - 1$ for $n = 2, \ldots, 7$.

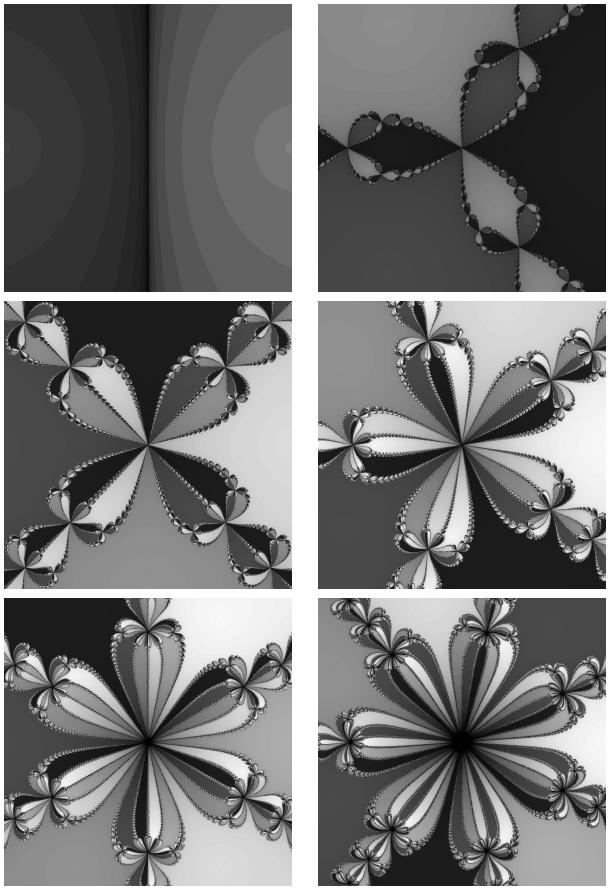


Figure 4.1 Basins of attraction for Newton's method applied to $z^n - 1$ (colour images can be produced by the code provided)

4.2 EXERCISE 4.3

Implement Brent's algorithm. It should terminate if either $f(b_n)$ or $f(s)$ is zero or if $|b_n - a_n|$ is small enough. Use the bisection rule if s is not between $(3a_n + b_n)/4$ and b_n for both linear and inverse quadratic interpolation or if any of Brent's conditions arises. Try your program on $f(x) = x^3 - x^2 - 4x + 4$ which has zeros at $-2, 1$ and 2. Start with the interval $[-4, 2.5]$ which contains all roots. List which method is used in each iteration.

Solution

```
function [ x,k,z ] = Brent( f,a,b,tol,max )
% implements Brent's algorithm to find a solution of f(x)=0;
% f input argument, function handle or vector of polynomial
% coefficients
% b input argument, initial iterate
% c input argument, initial contrapoint
% tol input argument, tolerance
% max input argument, maximum number of iterations
% x output argument, solution
% k output argument, number of iterations
% z output argument, holds information which method was used for each
% iteration; 1 -> Binary search
%            2 -> Inverse quadratic interpolation
%            3 -> Linear interpolation

% first check user inputs
if tol≤0;
    error('tol must be >0');
elseif max≤0;
    error('max must be >0');
elseif isa(f,'function_handle');
    % do nothing
elseif isa(f,'double');
    [n,m]=size(f);  % find the size of f
    if n≠1 && m≠1; % then f is not a vector
        error(['f must be a function handle or vector holding ',...
            'polynomial coefficients']);
    else
        v=poly2sym(f);  % this converts the vector of coefficients to
                        % an expression for the polynomial
        f=matlabFunction(v);    % this converts the expression to a
                                % function handle
    end
else
    error(['f must be a function handle or vector holding ',...
            'polynomial coefficients']);
end

Δ=tol/2;        % set Δ

fa=feval(f,a);      % evaluate f(a)
fb=feval(f,b);      % evaluate f(b)
```

```
k=1;                % initialise iteration counter

if fa==0;       % check incase the solution is at one of the endpoints
    x=a; return;
elseif fb==0;
    x=b; return;
elseif fa*fb>0;   % if f(a),f(b) have the same sign
        error('f(a),f(b) must have opposite signs');
end

if abs(fb)<abs(fa);      % then we need to interchange b and c, so the
                         % iterate is the best approximation to x
    swap=a;              % hold a
    a=b; a=swap;         % swap a and b
    fa=feval(f,a);       % revaluate f(b)
    fb=feval(f,b);       % revaluate f(c)
end

c=a;             % we will need an extra variables to hold the previous
d=a;             % iterates, in the first instance set these to a.
fc=fa; fd=fa;        % store f(c), f(d)

z=zeros(max,1);      % z will be a variable which holds information on
z(1)=1;              % whether the previous iteration used binary
                     % search, linear interpolation or IQI.



% start of algorithm
while abs(a-b)>tol && k≤max;

    if fa≠fb && fb≠fc && fc≠fa;  % then choose s according to
                                 % inverse quadratic interpolation
        s=a*fb*fc/((fa-fb)*(fa-fc))+fa*b*fc/((fb-fa)*(fb-fc))...
            +fa*fb*c/((fc-fa)*(fc-fb));
        temp=2;      % temporarily store this choice of interpolation
    else                         % if any of fa,fb,fc coincide,
                                 % choose s according to linear
                                 % interpolation
        s=b-fb*(b-a)/(fb-fa);
        temp=3;      % temporarily store this choice of interpolation
    end

    if or((s<(3*a+b)/4 & s<b),(s>(3*a+b)/4 & s>b)); % if s is not
                                                    % between (3*a+b)/4
                                                    % and b
        s=(a+b)/2; z(k+1)=1; % choose s according to binary search
                             % and set z=1
    elseif z(k)==1 && abs(b-c)<Δ;
        s=(a+b)/2; z(k+1)=1; % choose s according to binary search
                             % and set z=1
    elseif z(k)≠1 && abs(c-d)<Δ;
        s=(a+b)/2; z(k+1)=1; % choose s according to binary search
                             % and set z=1
    elseif z(k)==1 && abs(s-b)≥abs(b-a)/2;
        s=(a+b)/2; z(k+1)=1; % choose s according to binary search
                             % and set z=1
```

```
    elseif z(k)≠1 && abs(s−b)≥abs(b−d)/2;
        s=(a+b)/2; z(k+1)=1; % choose s according to binary search
                             % and set z=1
    else
        z(k+1)=temp;
    end

    fs=feval(f,s);      % evaluate f(s)
    if fs==0;           % check if we have found the solution
        x=s;
        return;
    end

    d=c;    % update d
    c=b;    % update c
    fa=feval(f,a);      % evaluate f(a)
    fb=feval(f,b);      % evaluate f(b)
    fd=feval(f,d);      % evaluate f(d)
    fc=feval(f,c);      % evaluate f(c)

    if fa*fs<0;   % if f(a) and f(s) have different signs
        b=s;        % update b
        fb=feval(f,b);     % update f(b)
    else
        a=s;        % update a
        fa=feval(f,a);     % update f(a)
    end

    if abs(fa)<abs(fb);     % then interchange a and b, so the iterate
                            % is the best approximation to x
        swap=a;             % hold a
        a=b; b=swap;        % swap b and a
        swap=fa;            % hold f(a)
        fa=fb;  fb=swap;    % swap f(a) and f(b)
    end

    if fb==0;      % check if we have found the solution
        x=a;
    end

    k=k+1;  % increment k
end

x=b;        % set x to the most recent iterate
z=z(1:k);   % shorten z

if k==max+1;
    disp('maximum number of iterations reached');
end

end
```

For $f(x) = x^3 - x^2 - 4x + 4$ the above implementation converges after 14 iterations to -2, when the accuracy is set to 0.01. The first iteration uses binary search, the next linear interpolation. Then follow five iterations using

inverse quadratic interpolation and four binary seraches. The last iterations are two linear interpolations followed by an inverse quadratic interpolation.

4.3 EXERCISE 4.5

Newton's method for finding the solution of $f(x) = 0$ *is given by*

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})},$$

where $x^{(n)}$ *is the approximation to the root* x^* *in the* n*-th iteration. The starting point* $x^{(0)}$ *is already close enough to the root.*

(a) By means of a sketch graph describe how the method works in a simple case and give an example where it might fail to converge.

(b) Using the Taylor expansion of $f(x^*) = 0$ *about* $x^{(n)}$*, relate the error in the next iteration to the error in the current iteration and show that the convergence of Newton's method is quadratic.*

(c) Generalize Newton's method to higher dimensions.

(d) Let

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x, y) = \begin{pmatrix} \frac{1}{2}x^2 + y \\ \frac{1}{2}y^2 + x \end{pmatrix}.$$

The roots lie at $(0, 0)$ *and* $(-2, -2)$*. Calculate the Jacobian of* $\mathbf{f}$ *and its inverse.*

(e) Why does Newton's method fail near $(1, 1)$ *and* $(-1, -1)$*?*

(f) Let $\mathbf{x}^{(0)} = (1, 0)$*. Calculate* $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ *and* $\mathbf{x}^{(3)}$ *and their Euclidean norms.*

(g) The approximations converge to $(0, 0)$*. Show that the speed of convergence agrees with the theoretical quadratic speed of convergence.*

Solution

(a) Newton's method given by

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}.$$

is geometrically the tangent to the curve f at the point $(x^{(n)}, f(x^{(n)}))$. It has the equation

$$y = f(x^{(n)}) + f'(x^{(n)})(x - x^{(n)}).$$

The point $x^{(n+1)}$ is the point of intersection of this tangent with the x-axis. The method fails if any of the iteration points happens to be a stationary point, i.e. a point where the first derivative vanishes. In this case the next iteration step is undefined, since the tangent there will be parallel to the x-axis and not intersect it. Even if the derivative is nonzero, but small the next approximation may be far worse.

(b) Let x^* be the root. The Taylor expansion of $f(x^*)$ about $x^{(n)}$ is

$$f(x^*) = f(x^{(n)}) + f'(x^{(n)})(x^* - x^{(n)}) + \frac{1}{2!}f''(\xi^{(n)})(x^* - x^{(n)})^2,$$

where $\xi^{(n)}$ lies between $x^{(n)}$ and x^*. Since x^* is the root, this equates to zero

$$f(x^{(n)}) + f'(x^{(n)})(x^* - x^{(n)}) + \frac{1}{2!}f''(\xi^{(n)})(x^* - x^{(n)})^2 = 0.$$

Let's assume that f' is bounded away from zero in a neighbourhood of x^* and $x^{(n)}$ lies in this neighbourhood. We can then divide the above equation by $f'(x^{(n)})$. After rearranging this becomes

$$\frac{f(x^{(n)})}{f'(x^{(n)})} + (x^* - x^{(n)}) = -\frac{1}{2}\frac{f''(\xi^{(n)})}{f'(x^{(n)})}(x^* - x^{(n)})^2.$$

Using the definition of Newton's method we can relate the error in the next iteration to the error in the current iteration

$$x^* - x^{(n+1)} = x^* - x^{(n)} + \frac{f(x^{(n)})}{f'(x^{(n)})} = -\frac{1}{2}\frac{f''(\xi^{(n)})}{f'(x^{(n)})}(x^* - x^{(n)})^2.$$

This shows that under certain conditions the convergence of Newton's method is quadratic.

(c) Newton's method readily generalizes to higher dimensional problems. Given a function $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^m$, we consider $\mathbf{f}$ as a vector of m functions

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix},$$

where $\mathbf{x} \in \mathbb{R}^m$. Let $\mathbf{h} = (h_1, \dots h_m)^T$ be a small perturbation vector. The multidimensional Taylor expansion of each function component f_i, $i = 1, \dots, m$ is

$$f_i(\mathbf{x} + \mathbf{h}) = f_i(\mathbf{x}) + \sum_{k=1}^{m} \frac{\partial f_i(\mathbf{x})}{\partial x_k} h_k + O(\|\mathbf{h}\|^2).$$

The Jacobian matrix $J_f(\mathbf{x})$ has the entries $[J_f(\mathbf{x})]_{i,k} = \frac{\partial f_i(\mathbf{x})}{\partial x_k}$ and thus we can write in matrix notation

$$\mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{f}(\mathbf{x}) + J_f(\mathbf{x})\mathbf{h} + O(\|\mathbf{h}\|^2).$$

Assuming that $\mathbf{x}$ is an approximation to the root, we want to improve the approximation by choosing $\mathbf{h}$. Thus we want to choose $\mathbf{h}$ such that $f(\mathbf{x}+\mathbf{h}) = 0$. Ignoring higher order terms, we solve the above equation for $\mathbf{h}$. The new approximation is set to $\mathbf{x}+\mathbf{h}$. More formally the Newton iteration in higher dimensions is

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} - J_f(\mathbf{x}^{(n)})^{-1}\mathbf{f}(\mathbf{x}^{(n)}).$$

(d) Given

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x,y) = \begin{pmatrix} \frac{1}{2}x^2 + y \\ \frac{1}{2}y^2 + x \end{pmatrix}$$

the Jacobian is

$$J_f(\mathbf{x}) = \begin{pmatrix} x & 1 \\ 1 & y \end{pmatrix}$$

and its inverse is

$$J_f(\mathbf{x})^{-1} = \frac{1}{xy-1}\begin{pmatrix} y & -1 \\ -1 & x \end{pmatrix}.$$

(e) Since the inverse of the Jacobian involves the division by $xy-1$, it becomes unbounded near $(1,1)$ and $(-1,-1)$.

(f) Starting with $\mathbf{x}^{(0)} = (1,0)$, we have

$$\mathbf{x}^{(1)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \frac{1}{-1}\begin{pmatrix} 0 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix}$$

with norm $\|\mathbf{x}^{(1)}\| = 1/2$. Next

$$\mathbf{x}^{(2)} = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix} - \frac{1}{-1}\begin{pmatrix} \frac{1}{2} & -1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} \frac{1}{2} \\ \frac{1}{8} \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix} + \begin{pmatrix} \frac{1}{8} \\ -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{8} \\ 0 \end{pmatrix}$$

with norm $\|\mathbf{x}^{(2)}\| = 1/8$. Next

$$\mathbf{x}^{(3)} = \begin{pmatrix} \frac{1}{8} \\ 0 \end{pmatrix} - \frac{1}{-1}\begin{pmatrix} 0 & -1 \\ -1 & \frac{1}{8} \end{pmatrix}\begin{pmatrix} \frac{1}{2^7} \\ \frac{1}{8} \end{pmatrix} = \begin{pmatrix} \frac{1}{8} \\ 0 \end{pmatrix} + \begin{pmatrix} -\frac{1}{8} \\ -\frac{1}{2^7} + \frac{1}{2^6} \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{2^7} \end{pmatrix}$$

with norm $\|\mathbf{x}^{(3)}\| = 1/2^7$.

(g) We clearly have $\|\mathbf{x}^{(1)}\| = \frac{1}{2} = \frac{1}{2}\|\mathbf{x}^{(0)}\|^2$, $\|\mathbf{x}^{(2)}\| = \frac{1}{8} = \frac{1}{2}\|\mathbf{x}^{(1)}\|^2$ and $\|\mathbf{x}^{(3)}\| = \frac{1}{2^7} = \frac{1}{2}\|\mathbf{x}^{(2)}\|^2$.

4.4 EXERCISE 4.7

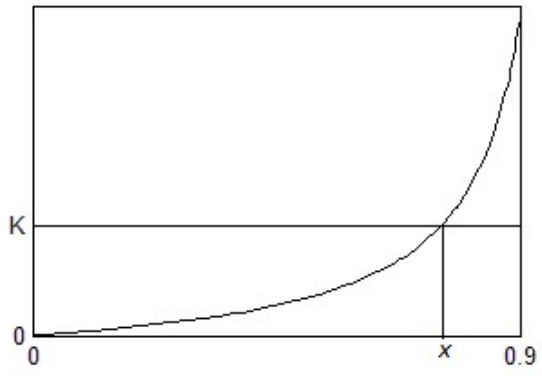
The following reaction occurs when water vapor is heated

$$H_2O \rightleftharpoons H_2 + \frac{1}{2}O_2.$$

The fraction $x \in [0, 1]$ *of* H_2O *that is consumed satisfies the equation*

$$K = \frac{x}{1-x}\sqrt{\frac{2p_t}{2+x}}, \tag{4.1}$$

where K *and* p_t *are given constants. The following figure illustrates this:*



(a) Rephrase the problem of determining x *as finding the root of a function* $f(x)$ *and state* $f(x)$*. Sketch a graph illustrating the rephrased problem.*

(b) Describe the bisection method to find the root of a function. Comment on the robustness and speed of convergence of the method.

(c) Given an approximation $x^{(n)}$ *to the root* x^* *of the function* $f(x)$*, give the formula how Newton's method calculates the next approximation* $x^{(n+1)}$*, explain what this means geometrically and expand your sketch with an example how Newton's method works.*

(d) What is the order of convergence of Newton's method?

(e) What happens to the right hand side of equation (4.1) if x *approaches* 1 *and what does this mean for Newton's method, if the starting point is chosen close to 1?*

(f) The derivative of $f(x)$ *at* 0 *is* 1*. What is the next approximation if* 0 *is chosen as the starting point? Depending on* K *what problem might this cause?*

(g) Give another example to demonstrate when Newton's method might fail to converge?

Solution

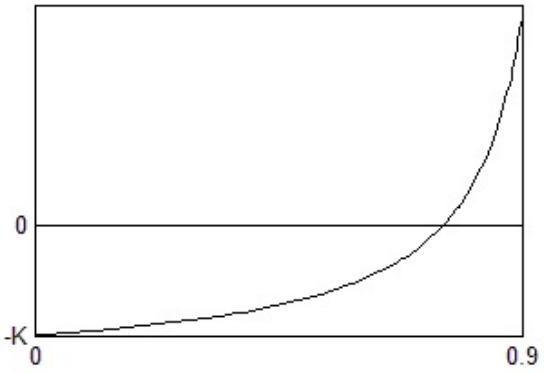
(a) Finding the fraction x satisfying the equation

$$K = \frac{x}{1-x}\sqrt{\frac{2p_t}{2+x}},$$

is the same as finding the root of the function

$$f(x) = \frac{x}{1-x}\sqrt{\frac{2p_t}{2+x}} - K.$$

The following figure shows the adjusted graph:



(b) If for a given interval $[a, b]$ $f(a)$ and $f(b)$ have opposite signs, then f must have at least one zero in the interval by the intermediate value theorem, if f is continuous. The bisection method can be used to find the zero. It is also known as binary search method.

We repeatedly bisect the interval and select the interval in which the root must lie. At each step we calculate the midpoint $m = (a + b)/2$ and the function value $f(m)$. Unless m is itself a root (improbable, but not impossible), there are two cases: If $f(a)$ and $f(m)$ have opposite signs, then the method sets m as the new value for b. Otherwise if $f(m)$ and $f(b)$ have opposite signs, then the method sets m as the new a. The algorithm terminates, when $b - a$ is sufficiently small.

It is robust, i.e. it is guaranteed to converge although at a possibly slow rate. Suppose the calculation is performed in binary. In every step the width of the interval containing a zero is reduced by 50%. Therefore at worst the method will add one binary digit of accuracy in each step. So the iterations are linearly convergent.

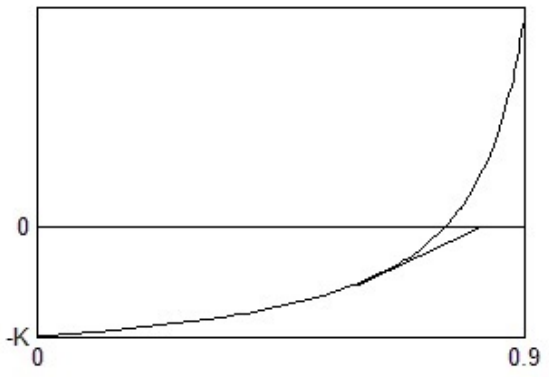
(c) Newton's method given by

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}.$$

is geometrically the tangent to the curve f at the point $(x^{(n)}, f(x^{(n)}))$. It has the equation

$$y = f(x^{(n)}) + f'(x^{(n)})(x - x^{(n)}).$$

The point $x^{(n+1)}$ is the point of intersection of this tangent with the x-axis. This is illustrated by the following figure:



(d) Under certain conditions the convergence of Newton's method is quadratic. The conditions are that there exists a neighbourhood U of the root where f' is bounded away from zero and where f'' is finite and that the starting point lies sufficiently close to x^*.

(e) If x approaches 1, then

$$\frac{x}{1-x}\sqrt{\frac{2p_t}{2+x}}$$

approaches infinity, since the denominator $1 - x$ is close to zero. If the starting point in Newton's method is chosen close to 1, then the tangent is close to a vertical line which means the next approximation is close to the previous approximation. This means that convergence is slow. With every next approximation the angle between the tangent and the vertical increases and convergence will speed up.

(f) If 0 is the starting point $x^{(0)}$, then the next approximation is

$$x^{(1)} = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})} = 0 - \frac{-K}{1} = K.$$

If $K > 1$, then the next approximation $x^{(1)}$ lies outside the interval $[0, 1]$ (where we know the root lies).

(g) Either of the following is accepted as answer:

- Newton's method fails if any of the iteration points happens to be a stationary point, i.e. a point where the first derivative vanishes. In this case the next iteration step is undefined, since the tangent there will be parallel to the x-axis and not intersect it. Even if the derivative is nonzero, but small the next approximation may be far worse.
- For some functions it can happen that the iteration points enter an infinite cycle. Take for example the polynomial $f(x) = x^3 - 2x + 2$. If 0 is chosen as the starting point, the first iteration produces 1, while the next iteration produces 0 again and so forth.

CHAPTER 5

Numerical Integration Exercises

5.1 EXERCISE 5.1

The mid-point rule $(b-a)f(\frac{1}{2}(a+b))$ *is exact for polynomials of degree* 1. *Use Peano's kernel theorem to find a formula for* $L(f)$. *(Hint: This is similar to the trapezium rule, except that it is harder to prove that* $K(\theta)$ *does not change sign in* $[a, b]$.*)*

Solution

The kernel is given by

$$K(\theta) = L[(x-\theta)_+] = \int_a^b (x-\theta)_+ dx - (b-a)(\frac{a+b}{2} - \theta)_+.$$

For $\theta \in [\frac{a+b}{2}, b]$ the second term of the difference is zero and thus $K(\theta) = \frac{1}{2}(b-\theta)^2 \geq 0$.

For $\theta \in [a, \frac{a+b}{2}]$ we have

$$\begin{aligned} K(\theta) &= \tfrac{1}{2}(b-\theta)^2 - \tfrac{1}{2}(b-a)(a+b-2\theta) \\ &= \tfrac{1}{2}(b^2 - 2b\theta + \theta^2 - ab - b^2 + 2b\theta + a^2 + ab - 2a\theta) \\ &= \tfrac{1}{2}(\theta - a)^2 \geq 0. \end{aligned}$$

Thus the kernel does not change sign. Next we need to calculate the integral

of the kernel

$$\begin{aligned}\int_a^b K(\theta)d\theta &= \int_a^{(a+b)/2} \frac{1}{2}(\theta-a)^2 d\theta + \int_{(a+b)/2}^b \frac{1}{2}(b-\theta)^2 d\theta \\ &= \left[\frac{1}{6}(\theta-a)^3\right]_a^{(a+b)/2} + \left[-\frac{1}{6}(b-\theta)^3\right]_{(a+b)/2}^b \\ &= \frac{1}{24}(b-a)^3.\end{aligned}$$

The error bound which was deduced from first principles in the lecture follows.

5.2 EXERCISE 5.3

Implement the Gauss-Legendre quadrature for $n = 2, \ldots, 5$ and approximate $\int_{-1}^{1} x^j dx$ for $j = 1, \ldots, 10$ and compare the results to the true solution. Interpret your results.

Solution

```
function [ Q ] = GaussLegendre( f,a,b,n )
% employs Gauss–Legendre rule to integrate f over [a,b]
% f input argument, function handle
% a,b input arguments, integration bounds, a<b
% n input argument, number of abscissae
% Q output argument, value of integral

% first check user inputs
if isa(f,'function_handle')==0;
    error('f must be a function handle');
elseif a≥b;
    error('a must be <b');
elseif mod(n,1)≠0 || n≤0; % if n does not equal zero modulo 1
                          % then it is not an integer value
    error('n must be a positive integer');
end

syms x;     % create a symbolic variable
legendre=legendreP(n,x);        % look up Legendre polynomial
legendre=sym2poly(legendre);    % convert symbolic expression to
                                % vector of polynomial coefficents
knots=roots(legendre);  % calculate roots of the Legendre polynomial

% generate the weights by integrating the Lagrange
% interpolating polynomials
for k=1:n;
    syms L; % initialise symbolic variable for kth Lagrange polynomial
    L=1;    % set to unity
    for j=1:n;
        if j==k;
            % do nothing
        else
```

```
            L=L*(x-knots(j))/(knots(k)-knots(j)); % add another term
                                                  % to the product
        end
    end
    L=sym2poly(L);      % convert symbolic expression to a vector of
                        % polynomial coefficients
    I=zeros(n+1,1);     % initialise vector to hold the integral
    for j=1:n;          % construct the polynomial coeffients of the
                        % integral of L
        I(j)=L(j)/(n+1-j);
    end
    weights(k)=polyval(I,1)-polyval(I,-1); % evaluate I at the
                                           % endpoints to determine
                                           % the weight
end
Q=0;    % initialise Q
% evaluate the quadrature
for i=1:n;
    Q=Q+weights(i)*feval(f,(b-a)*knots(i)/2+(a+b)/2);
end
Q=(b-a)/2*Q;
end
```

The following table shows the error each Gauss-Legendre quadrature makes when approximating the given integral.

	$n=2$	$n=3$	$n=4$	$n=5$
$\int_{-1}^{1} x dx = 0$	$2.2 * 10^{-16}$	$5.6 * 10^{-17}$	$1.1 * 10^{-16}$	$5.6 * 10^{-17}$
$\int_{-1}^{1} x^2 dx = \frac{2}{3}$	$-1.1 * 10^{-16}$	0	$1.1 * 10^{-16}$	$1.1 * 10^{-16}$
$\int_{-1}^{1} x^3 dx = 0$	$2.2 * 10^{-16}$	$1.1 * 10^{-16}$	$1.7 * 10^{-17}$	$5.6 * 10^{-17}$
$\int_{-1}^{1} x^4 dx = \frac{2}{5}$	-0.18	$-5.6 * 10^{-17}$	$5.6 * 10^{-16}$	0
$\int_{-1}^{1} x^5 dx = 0$	$1.2 * 10^{-16}$	$1.1 * 10^{-16}$	$-1.2 * 10^{-17}$	$-9.4 * 10^{-17}$
$\int_{-1}^{1} x^6 dx = \frac{2}{7}$	-0.21	-0.046	$8.3 * 10^{-16}$	0
$\int_{-1}^{1} x^7 dx = 0$	$5.9 * 10^{-17}$	$9.7 * 10^{-17}$	$-2.2 * 10^{-16}$	$-2.7 * 10^{-16}$
$\int_{-1}^{1} x^8 dx = \frac{2}{9}$	-0.20	-0.078	-0.011	0
$\int_{-1}^{1} x^9 dx = 0$	$2.4 * 10^{-17}$	$6.9 * 10^{-17}$	$-2.5 * 10^{-16}$	$-3.7 * 10^{-16}$
$\int_{-1}^{1} x^{10} dx = \frac{2}{11}$	-0.17	-0.095	-0.026	-0.0029

The error is very small as long as the integrated power is less than $2n-1$, but increases significantly if it is larger than the degree of polynomial for which the quadrature is correct. This is not true for integrals of odd powers. These are odd functions for which $f(-x) = -f(x)$. The integral over the interval $[-1, 1]$ of any odd function evaluates to zero. Similarly, since the Legendre-Gauss abscissae are distributed symmetrically around zero, the quadrature rule evaluates to zero. The small errors are due to rounding errors.

5.3 EXERCISE 5.5

Consider the numerical evaluation of an integral of the form

$$I = \int_a^b f(x)w(x)dx.$$

(a) Define Gaussian quadrature and state how the abscissae are obtained. Give

a formula for the weights. If f is a polynomial, what is the maximum degree of f for which the Gaussian quadrature rule is correct.

(b) In the following let the interval be $[a, b] = [-2, 2]$ and $w(x) = 4 - x^2$. Thus we want to approximate the integral

$$\int_{-2}^{2} (4 - x^2) f(x) dx.$$

Let the number of abscissae be 2. Calculate the abscissae.

(c) Calculate the weights.

(d) To approximate the integral

$$\int_{-1}^{1} (1 - x^2) f(x) dx.$$

by a Gaussian quadrature the orthogonal polynomials are the Jacobi polynomials for $\alpha = 1$ and $\beta = 1$. For $n = 2$ the abscissae are $x_1 = -1/\sqrt{5}$ and $x_2 = 1/\sqrt{5}$. The weights are $w_1 = w_2 = 2/3$. The interval of integration is changed from $[-1, 1]$ to $[-2, 2]$. What are the new abscissae and weights? Explain why the weights are different to the weights derived in the previous part.

Solution

(a) Gaussian quadrature approximates the integral in the following way

$$\int_a^b f(x) w(x) d(x) \approx \sum_{i=1}^{n} w_i f(x_i).$$

The abscissae $x_1, \ldots, x_n$ are the roots of the n-th orthogonal polynomial p_n, i.e.

$$\int_a^b p_n(x) p(x) w(x) dx = 0$$

for all polynomials of degree less than or equal to $n - 1$. Let L_i be the i-th Lagrange interpolating polynomial for these abscissae, i.e. L_i is the unique polynomial of degree $n-1$ such that $L_i(x_i) = 1$ and $L_i(x_k) = 0$ for $k \neq i$. The weights for the quadrature rule are then calculated according to

$$w_i = \int_a^b L_i(x) w(x) dx.$$

If f is a polynomial, the maximum degree of f for the quadrature rule to be correct is $2n - 1$.

(b) In the case $[a, b] = [-2, 2]$, $w(x) = 4 - x^2$ and $n = 2$ we seek weights w_1 and w_2 and abscissae x_1 and x_2 such that

$$\int_{-2}^{2} (4 - x^2) f(x) dx \approx w_1 f(x_1) + w_2 f(x_2).$$

We know that the abscissae x_1 and x_2 are the zeros of a quadratic polynomial p which is orthogonal to 1 and x with respect to the weight function $w(x)$. Let $p = x^2 + ax + b$,

$$\begin{aligned}
0 &= \int_{-2}^{2} (4 - x^2) p(x) dx \\
&= \int_{-2}^{2} (-x^4 - ax^3 + (4 - b)x^2 + 4ax + 4b) dx \\
&= \left[-\frac{1}{5}x^5 - \frac{a}{4}x^4 + \frac{4 - b}{3}x^3 + 2ax^2 + 4bx \right]_{-2}^{2} \\
&= -\frac{1}{5}2^5 + \frac{4 - b}{3}2^3 + 4b2 + \frac{1}{5}(-2)^5 - \frac{4 - b}{3}(-2)^3 + 4b(-2) \\
&= -\frac{1}{5}2^6 + \frac{4 - b}{3}2^4 + 4b2^2 \\
&= 2^4(-\frac{4}{5} + \frac{4 - b}{3} + b).
\end{aligned}$$

From this we deduce $b = -\frac{4}{5}$. To deduce a we consider

$$\begin{aligned}
0 &= \int_{-2}^{2} (4 - x^2) x p(x) dx \\
&= \int_{-2}^{2} (-x^5 - ax^4 + (4 - b)x^3 + 4ax^2 + 4bx) dx \\
&= \left[-\frac{1}{6}x^6 - \frac{a}{5}x^5 + \frac{4 - b}{4}x^4 + \frac{4a}{3}x^3 + 2bx^2 \right]_{-2}^{2} \\
&= -\frac{a}{5}2^5 + \frac{4a}{3}2^3 + \frac{a}{5}(-2)^5 - \frac{4a}{3}(-2)^3 = \frac{17}{15}2^6 a.
\end{aligned}$$

Therefore $a = 0$ and $p(x) = x^2 - \frac{4}{5}$ and the abscissae are $x_1 = -2/\sqrt{5}$ and $x_2 = 2/\sqrt{5}$.

(c) To determine the weights we use the fact that the quadrature has to be correct when integrating 1 and x. Thus

$$\begin{aligned}
w_1 + w_2 &= \int_{-2}^{2} (4 - x^2) dx = \frac{32}{3} \\
\frac{2}{\sqrt{5}}(-w_1 + w_2) &= \int_{-2}^{2} (4 - x^2) x dx = 0.
\end{aligned}$$

and $w_1 = w_2 = 16/3$.

(d) If one has a quadrature rule for the interval $[c, d]$, it can be adapted to the interval $[a, b]$ with a simple change of variables. Let $t(x)$ be the linear transformation taking $[c, d]$ to $[a, b]$,

$$t(x) = a + \frac{b-a}{d-c}(x-c).$$

If x_i and w_i, $i = 0, \ldots, n$, are the abscissae and weights of a quadrature approximating integrals over $[c, d]$, then the abscissae and weights of the quadrature over $[a, b]$ are

$$\hat{x}_i = t(x_i) \text{ and } \hat{w}_i = \frac{b-a}{d-c} w_i.$$

In our case $[c, d] = [-1, 1]$ and $[a, b] = [-2, 2]$ and $t(x) = 2x$. The new abscissae and weights are

$$\hat{x}_1 = -2/\sqrt{5}, \hat{x}_2 = 2/\sqrt{5} \text{ and } \hat{w}_1 = \hat{w}_2 = 4/3.$$

The weights differ since the weight functions do not correspond.

5.4 EXERCISE 5.7

(a) Describe what is meant by a composite rule of integration.

(b) Give two examples of composite rules and their formulae.

(c) Let a quadrature rule be given on $[c, d]$ by

$$Q_n(f) = \sum_{j=1}^{n} w_j f(x_j) \approx \int_c^d f(x)dx.$$

We denote by $(M \times Q_n)$ the composite rule Q_n applied to M subintervals of $[a, b]$. Give the formula for $(M \times Q_n)$.

(d) Describe the difference between open and closed quadrature rules and how this affects the composite rule.

(e) Show that if Q_n is a quadrature rule that integrates constants exactly, i.e. $Q_n(1) = \int_c^d 1dx = d - c$ and if f is bounded on $[a, b]$ and is Riemann integrable, then

$$\lim_{M \to \infty} (M \times Q_n)(f) = \int_a^b f(x)dx.$$

(f) Let $[c, d] = [-1, 1]$. Give the constant, linear and quadratic monic polynomials which are orthogonal with respect to the inner product given by

$$\langle f, g \rangle = \int_{-1}^{1} f(x)g(x)dx$$

and check that they are orthogonal to each other.

(g) Give the abscissae of the two-point Gauss-Legendre rule on the interval $[-1, 1]$.

(h) The weights of the two-point Gauss-Legendre rule are 1 *for both abscissae. State the two point Gauss-Legendre rule and give the formula for the composite rule on* $[a, b]$ *employing the two-point Gauss-Legendre rule.*

Solution

(a) A composite rule is constructed by splitting the integral into a set of panels and applying (usually) the same quadrature rule in each subinterval and summing the results.

(b) Possible examples for composite rules are:

- composite midpoint rule:

$$\int_a^b f(x)dx \approx h\sum_{i=1}^{N} f(a + (i - \frac{1}{2})h),$$

- composite trapezium rule:

$$\int_a^b f(x)dx \approx \frac{h}{2}f(a) + h\sum_{i=1}^{N-1} f(a + ih) + \frac{h}{2}f(b),$$

- composite Simpson rule:

$$\int_a^b f(x)dx \approx \frac{h}{6}\left[f(a) + 2\sum_{i=1}^{N-1} f(a + ih) + 4\sum_{i=1}^{N} f(a + (i - \frac{1}{2})h) + f(b)\right],$$

- composite rectangle rule:

$$\int_a^b f(x)dx \approx h\sum_{i=1}^{N} f(a + ih).$$

(c) For the quadrature rule given on $[c, d]$ by

$$Q_n(f) = \sum_{j=1}^{n} w_j f(x_j).$$

the composite rule on $[a, b]$ is given by

$$(M \times Q_n)(f) = \frac{b - a}{M(d - c)}\sum_{i=1}^{M}\sum_{j=1}^{n} w_j f(x_{ij}),$$

where x_{ij} is the j-th abscissa in the i-th subinterval calculated as $x_{ij} = t_i(x_j)$, where t_i is the transformation taking $[c, d]$ of length $d - c$ to $[a + (i - 1)(b - a)/M, a + i(b - a)/M]$ of length $(b - a)/M$.

(d) If Q_n is an open rule, it does not include the endpoints. If Q_n is a closed rule, it includes both end-points. If Q_n is an open rule, then $(M \times Q_n)$ uses Mn points. However, if Q_n is a closed rule, then $M \times Q_n$ uses only $(n-1)M+1$ points, which is $M-1$ less function evaluations.

(e) If Q_n is a quadrature rule that integrates constants exactly, i.e. $Q_n(1) = \int_c^d 1dx = d-c$, then

$$Q_n(1) = \sum_{j=1}^{n} w_j = d - c.$$

That is the weights sum to $d-c$. Swapping the summations and taking the limit in

$$(M \times Q_n)(f) = \frac{b-a}{M(d-c)} \sum_{i=1}^{M} \sum_{j=1}^{n} w_j f(x_{ij}),$$

gives

$$\lim_{M\to\infty} (M \times Q_n)(f) = \frac{1}{d-c} \sum_{j=1}^{n} w_j \lim_{M\to\infty} \left[\frac{b-a}{M} \sum_{i=1}^{M} f(x_{ij}) \right].$$

The term in the square brackets is a Riemann sum and thus converges to the integral $\int_a^b f(x)dx$, since x_{ij} lies in the i-th subinterval of length $(b-a)/M$ for each j. Since the weights sum to $d-c$, it follows that

$$\lim_{M\to\infty} (M \times Q_n)(f) = \int_a^b f(x)dx.$$

(f) The constant monic orthogonal polynomial is $p_0(x) = 1$. The linear monic orthogonal polynomial is $p_1(x) = x$ and we have

$$\langle p_0, p_1 \rangle = \int_{-1}^{1} x dx = \left[\frac{x^2}{2} \right]_{-1}^{1} = 0.$$

Let $p_2 = x^2 + a$. To determine a we calculate

$$0 = \langle p_0, p_2 \rangle = \int_{-1}^{1} (x^2 + a)dx = \left[\frac{x^3}{3} + ax \right]_{-1}^{1} = \frac{1}{3} + a - (\frac{-1}{3} - a) = \frac{2}{3} + 2a.$$

Hence $a = -1/3$ and $p_2(x) = x^2 - 1/3$. By construction p_2 is orthogonal to p_0. Check

$$\langle p_1, p_2 \rangle = \int_{-1}^{1} (x^3 - \frac{x}{3})dx = \left[\frac{x^4}{4} - \frac{x^2}{6} \right]_{-1}^{1} = \frac{1}{4} - \frac{1}{6} - \frac{1}{4} + \frac{1}{6} = 0.$$

(g) The abscissae of the two-point Gauss-Legendre rule on $[-1, 1]$ are the roots of p_2 which are $\pm 1/\sqrt{3}$.

(h) The two point Gauss-Legendre rule is

$$\int_{-1}^{1} f(x)d(x) \approx f(-\frac{1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}}).$$

The transformation t_i taking $[-1, 1]$ of length 2 to the i-th subinterval $[a + (i-1)(b-a)/M, a + i(b-a)/M]$ of length $(b-a)/M$ is

$$t_i(x) = a + \frac{b-a}{2M}(x + 2i - 1).$$

The composite rule on $[a, b]$ is given by

$$\frac{b-a}{2M}\sum_{i=1}^{M}[f(a + \frac{b-a}{2M}(-\frac{1}{\sqrt{3}} + 2i - 1)) + f(a + \frac{b-a}{2M}(\frac{1}{\sqrt{3}} + 2i - 1)).$$

5.5 EXERCISE 5.9

The integral

$$\int_0^2 f(x)dx$$

shall be approximated by a two point Gaussian quadrature formula.

(a) Find the monic quadratic polynomial $g(x)$ which is orthogonal to all linear polynomials with respect to the scalar product

$$\langle f, g\rangle = \int_0^2 f(x)g(x)dx,$$

where $f(x)$ denotes an arbitrary linear polynomial.

(b) Calculate the zeros of the polynomial found in (a) and explain how they are used to construct a Gaussian quadrature rule.

(c) Describe how the weights are calculated for a Gaussian quadrature rule and calculate the weights to approximate $\int_0^2 f(x)dx$.

(d) For which polynomials is the constructed quadrature rule correct?

(e) State the functional $L(f)$ acting on f describing the error when the integral $\int_0^2 f(x)dx$ is approximated by the quadrature rule.

(f) Define the Peano kernel and state the Peano kernel theorem.

(g) Calculate the Peano kernel for the functional $L(f)$ in (e).

(h) The Peano kernel does not change sign in $[0, 2]$ (not required to be proven). Derive an expression for $L(f)$ of the form constant times a derivative of f. (Hint: $(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$)

Solution

(a) The integral

$$\int_0^2 f(x)dx$$

shall be approximated by a Gaussian quadrature formula. Let x^2+ax+b be a general monic quadratic polynomial. It is orthogonal to all linear polynomials if it is orthogonal to 1 and x:

$$0 = \int_0^2 (x^2+ax+b)dx = [\frac{x^3}{3} + \frac{ax^2}{2} + bx]_0^2 = \frac{8}{3} + 2a + 2b,$$

$$0 = \int_0^2 (x^2+ax+b)xdx = [\frac{x^4}{4} + \frac{ax^3}{3} + \frac{bx^2}{2}]_0^2 = 4 + \frac{8}{3}a + 2b.$$

Subtracting the first equation from the second gives

$$\frac{12}{3} - \frac{8}{3} + (\frac{8}{3} - 2)a = 0$$

and thus $a = -2$. Inserting this into any of the two equations gives $b = 2/3$.

(b) The zeros of $x^2 - 2x + 2/3$ are

$$x_{1,2} = \frac{-(-2) \pm \sqrt{(-2)^2 - 4\frac{2}{3}}}{2} = 1 \pm \frac{1}{\sqrt{3}}.$$

They are used as the abscissae in the two point Gaussian quadrature rule.

(c) The weights for a Gaussian quadrature are calculated by finding the Lagrange polynomials which are one at one of the abscissae and zero at all the others and integrating them over the range. In this case we have to find two linear polynomials, one interpolating the data $(1-1/\sqrt{3},1),(1+1/\sqrt{3},0)$ and one interpolating $(1-1/\sqrt{3},0),(1+1/\sqrt{3},1)$. The first one is given by

$$\frac{x-(1+1/\sqrt{3})}{(1-1/\sqrt{3})-(1+1/\sqrt{3})} = (-\sqrt{3}x + \sqrt{3} + 1)/2$$

while the other one is

$$\frac{x-(1-1/\sqrt{3})}{(1+1/\sqrt{3})-(1-1/\sqrt{3})} = (\sqrt{3}x - \sqrt{3} + 1)/2.$$

Integrating gives

$$\frac{1}{2}\int_0^2 (-\sqrt{3}x + \sqrt{3} + 1)dx = \frac{1}{2}[-\sqrt{3}\frac{x^2}{2} + (\sqrt{3}+1)x]_0^2 = 1,$$

$$\frac{1}{2}\int_0^2 (\sqrt{3}x - \sqrt{3} + 1)dx = \frac{1}{2}[\sqrt{3}\frac{x^2}{2} - (\sqrt{3}-1)x]_0^2 = 1,$$

(d) The quadrature rule

$$\int_0^2 f(x)dx \approx f(1-\frac{1}{\sqrt{3}}) + f(1+\frac{1}{\sqrt{3}})$$

is correct for all cubic polynomials.

(e) The functional describing the error when approximating the integral $\int_0^2 f(x)dx$ by the quadrature rule is

$$L(f) = \int_0^2 f(x)dx - [f(1-\frac{1}{\sqrt{3}}) + f(1+\frac{1}{\sqrt{3}})].$$

(f) The Peano kernel K of L is the function defined by

$$K(\theta) := L[(x-\theta)_+^k] \text{ for } x \in [a,b].$$

where k is the largest integer such that $L(p) = 0$ for all $p \in \mathbb{P}_k[x]$.

Peano Kernel Theorem: *Let L be a linear functional such that $L(p) = 0$ for all $p \in \mathbb{P}_k[x]$. Provided that the exchange of L with the integration is valid, then for $f \in C^{k+1}[a,b]$*

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

(g) $L(f) = 0$ for all polynomials of degree 3 or less. Thus $k = 3$ and the Peano kernel is

$$K(\theta) = \int_0^2 (x-\theta)_+^3 dx - [(1-\frac{1}{\sqrt{3}}-\theta)_+^3 + (1+\frac{1}{\sqrt{3}}-\theta)_+^3].$$

Now

$$\int_0^2 (x-\theta)_+^3 dx = [(x-\theta)_+^4/4]_0^2 = (2-\theta)_+^4/4 - (0-\theta)_+^4/4.$$

The second term is zero since $\theta \in [0,2]$.

For $0 \le \theta \le 1 - 1/\sqrt{3}$

$$K(\theta) = (2-\theta)^4/4 - [(1-\frac{1}{\sqrt{3}}-\theta)^3 + (1+\frac{1}{\sqrt{3}}-\theta)^3].$$

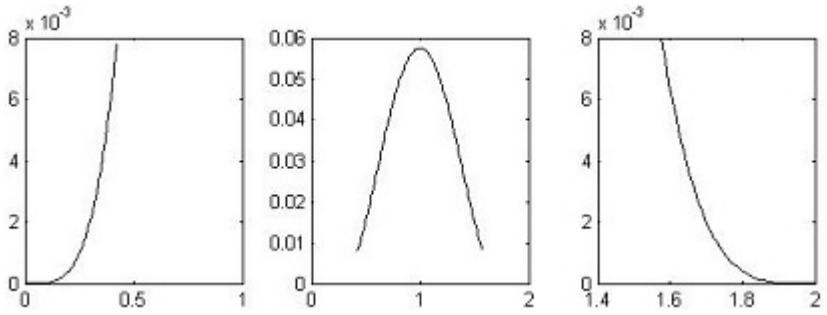
For $1 - 1/\sqrt{3} \le \theta < 1 + 1/\sqrt{3}$

$$K(\theta) = (2-\theta)^4/4 - (1+\frac{1}{\sqrt{3}}-\theta)^3.$$

For $1 + 1/\sqrt{3} \le \theta < 2$

$$K(\theta) = (2-\theta)^4/4.$$

The figure shows the pieces of the Peano kernel. It is not required to prove that it does not change sign.



(h) Since $K(\theta)$ does not change sign we have

$$L(f) == \frac{1}{3!}\int_0^2 K(\theta)d\theta f^{(k+1)}(\xi).$$

Next step is to integrate the kernel

$$\begin{aligned}
\int_0^2 K(\theta)d\theta &= \int_0^2 (2-\theta)^4/4d\theta - \int_0^{1+\sqrt{3}} (1+\frac{1}{\sqrt{3}}-\theta)^3 d\theta \\
&\quad -\int_0^{1-\sqrt{3}} (1-\frac{1}{\sqrt{3}}-\theta)^3 d\theta \\
&= [-(2-\theta)^5/20]_0^2 - [-(1+\frac{1}{\sqrt{3}}-\theta)^4/4]_0^{1+\sqrt{3}} \\
&\quad -[-(1-\frac{1}{\sqrt{3}}-\theta)^4/4]_0^{1-\sqrt{3}} \\
&= 2^5/20 - (1+\frac{1}{\sqrt{3}})^4/4 - (1-\frac{1}{\sqrt{3}})^4/4 \\
&= 2^5/20 - (1+\frac{4}{\sqrt{3}}+\frac{6}{3}+\frac{4}{3\sqrt{3}}+\frac{1}{9})/4 \\
&\quad -(1-\frac{4}{\sqrt{3}}+\frac{6}{3}-\frac{4}{3\sqrt{3}}+\frac{1}{9})/4 = 2/45
\end{aligned}$$

Therefore

$$L(f) = \frac{2}{3! * 45} = \frac{1}{135}.$$

CHAPTER 6

ODEs Exercises

6.1 EXERCISE 6.1

Let $h = \frac{1}{M}$, where M is a positive integer. The following ODEs are given

$$y' = -\frac{y}{1+t} \qquad \textit{and} \qquad y' = \frac{2y}{1+t}, \qquad 0 \le t \le 1,$$

with starting conditions $y_0 = y(0) = 1$ in both cases. Forward Euler is used to calculate the estimates y_n, $n = 1, \ldots M$. By using induction and by canceling as many terms as possible in the resultant products, deduce simple explicit expressions for y_n, $n = 1, \ldots M$, which should be free from summations and products. By considering the limit for $h \to 0$, deduce the exact solutions of the equations. Verify that the errors $|y_n - y(t_n)|$ is at most $O(h)$.

Solution

For the first equation forward Euler gives

$$\begin{aligned} y_{n+1} &= y_n - h\frac{y_n}{1+t_n} = y_n\left(1 - \frac{h}{1+nh}\right) = \frac{y_n(1+nh-h)}{1+nh} \\ &= y_0 \prod_{j=0}^{n} \frac{(1+jh-h)}{(1+jh)} = \frac{1-h}{1+nh}, \end{aligned}$$

where we use $y_0 = 1$. We let h tend to zero and pick n for each h so that $nh \to t$. Then

$$\frac{1-h}{1+nh} \to \frac{1}{1+t}$$

which can easily be verified to be the analytic solution. The error is

$$|y_n - y(t_n)| = \left|\frac{1-h}{1+nh-h} - \frac{1}{1+nh}\right| = \frac{nh^2}{(1+nh)(1+nh-h)}$$

which is $O(h)$ when $t_n = nh$ is confined to a finite interval.

For the second equation we have

$$\begin{aligned} y_{n+1} &= y_n + h\frac{2y_n}{1+t_n} = y_n\frac{1+nh+2h}{1+nh} \\ &= y_0\prod_{j=0}^{n}\frac{(1+jh+2h)}{(1+jh)} = \frac{(1+nh+2h)(1+nh+h)}{1+h}. \end{aligned}$$

Taking the limit as before, we obtain

$$\frac{(1+nh+2h)(1+nh+h)}{1+h} \to (1+t)^2$$

which again can be easily verified to be the analytic solution. We have

$$|y_n - y(t_n)| = \left|\frac{(1+nh+h)(1+nh)}{1+h} - (1+nh)^2\right| = \frac{nh^2(1+nh)}{1+h} = O(h).$$

6.2 EXERCISE 6.3

Implement the backward Euler method in MATLAB or a different programming language of your choice.

Solution

The backward Euler method can be implemented as follows

```
function [x,y]=euler_backward(f,tinit,yinit,tfinal,n)
% Euler backward method
% Calculation of h from tinit, tfinal, and n
h=(tfinal-tinit)/n;
% Initialization of t and y as column vectors
t=[tinit zeros(1,n)];
y=[yinit zeros(1,n)];
% Calculation of t and y
for i=1:n
t(i+1)=t(i)+h;
ynew=y(i)+h*(f(t(i),y(i)));
y(i+1)=y(i)+h*f(t(i+1),ynew);
end
end
```

The right hand side f needs to be defined via a function handle. For example if $y' = \frac{t}{y}$, which has the analytic solution $y(t) = \sqrt{x^2+1}$ for $y(0) = 1$, the function handle can be defined anonymously by

```
f=@(t,y) t./y;
```

Note that this is a very simple implementation to solve the implicit system. Only one step of a direct iteration is executed. A better solution would be the

use of Newton-Raphson. In this case another function handle for the derivative of f with respect to y needs to be defined.

```
df=@(t,y) -t./(y*y);
```

A few iterations of Newton-Raphson are

```
for k=1:5
 ynew = ynew -(h * df(t(i),ynew) - 1)\(h * f(t(i),ynew) - ynew + y(i));
end;
y(i+1) = ynew;
```

In a more sophisticated implementation the iterations are stopped depending on how the error compares with an estimate of the local truncation error.

6.3 EXERCISE 6.5

Show that the method given by

$$\mathbf{y}_{n+2} - 3\mathbf{y}_{n+1} + 2\mathbf{y}_n = \frac{1}{12}h\left[13\mathbf{f}(t_{n+2}, \mathbf{y}_{n+2}) - 20\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) - 5\mathbf{f}(t_n, \mathbf{y}_n)\right]$$

is at least of order 2 *just like the* 2*-step Adams-Bashforth method.*

Solution

The associated polynomials with this method are

$$\rho(w) = w^2 - 3w + 2 \text{ and } \sigma(w) = \frac{13}{12}w^2 - \frac{5}{3}w - \frac{5}{12}.$$

To have order to we require $\rho(e^z) - z\sigma(e^z) = O(z^3)$. Thus

$$\begin{aligned} e^{2z} - 3e^z + 2 - z(\tfrac{13}{12}e^{2z} - \tfrac{5}{3}e^z - \tfrac{5}{12}) = \;& 1 + 2z + 2z^2 - 3 - 3z - \tfrac{3}{2}z^2 + 2 \\ & -z\left(\tfrac{13}{12} + \tfrac{13}{6}z - \tfrac{5}{3} - \tfrac{5}{3}z - \tfrac{5}{12}\right) + O(z^3) \\ = \;& O(z^3), \end{aligned}$$

since all other terms vanish.

6.4 EXERCISE 6.7

The stiff differential equation

$$y'(t) = -10^6(y - t^{-1}) - t^{-2}, \qquad t \geq 1, \qquad y(1) = 1,$$

has the analytical solution $y(t) = t^{-1}$, $t \geq 1$. *Let it be solved numerically by forward Euler* $y_{n+1} = y_n + h_n f(t_n, y_n)$ *and by backward Euler* $y_{n+1} =$

$y_n + h_n f(t_{n+1}, y_{n+1})$, where $h_n = t_{n+1} - t_n$ is allowed to depend on n and to be different for the two methods. Suppose that at a point $t_n \geq 1$ an accuracy of $|y_n - y(t_n)| \leq 10^{-6}$ is achieved and that we want to achieve the same accuracy in the next step, i.e. $|y_{n+1} - y(t_{n+1})| \leq 10^{-6}$. Show that forward Euler can fail if $h_n = 2 \times 10^{-6}$, but that backward Euler always achieves the desired accuracy if $h_n \leq t_n t_{n+1}^2$. (Hint: Find relations between $y_{n+1} - y(t_{n+1})$ and $y_n - y(t_n)$.)

Solution

For the forward Euler method we have

$$\begin{aligned} y_{n+1} - y(t_{n+1}) &= y_n - 10^6 h_n (y_n - \frac{1}{t_n}) - \frac{h_n}{t_n^2} - \frac{1}{t_{n+1}} \\ &= (1 - 10^6 h_n)(y_n - \frac{1}{t_n}) + \frac{1}{t_n} - \frac{h_n}{t_n^2} - \frac{1}{t_{n+1}} \\ &= (1 - 10^6 h_n)(y_n - y(t_n)) - \frac{h_n^2}{t_n^2 t_{n+1}}. \end{aligned}$$

For $h_n = 2 \times 10^{-6}$ the last equality becomes $-(y_n - y(t_n)) - \frac{h_n^2}{t_n^2 t_{n+1}}$. If $10^{-6} \geq y_n - y(t_n) > 10^{-6} - \frac{h_n^2}{t_n^2 t_{n+1}}$, we then have $|y_{n+1} - y(t_{n+1})| > 10^{-6}$.

On the other hand, for the backward Euler we obtain

$$y_{n+1} = y_n - 10^6 h_n (y_{n+1} - \frac{1}{t_{n+1}}) - \frac{h_n}{t_{n+1}^2} = y_n - 10^6 h_n (y_{n+1} - y(t_{n+1})) - \frac{h_n}{t_{n+1}^2}.$$

Solving for $y_{n+1} - y(t_{n+1})$ gives

$$\begin{aligned} (1 + 10^6 h_n)(y_{n+1} - y(t_{n+1})) &= y_n - y(t_{n+1}) - \frac{h_n}{t_{n+1}^2} \\ &= y_n - y(t_n) + \frac{1}{t_n} - \frac{1}{t_{n+1}} - \frac{h_n}{t_{n+1}^2} \\ &= y_n - y(t_n) + \frac{h_n^2}{t_n t_{n+1}^2}. \end{aligned}$$

Thus we have

$$|y_{n+1} - y(t_{n+1})| \leq \frac{1}{1 + 10^6 h_n} \left(|y_n - y(t_n)| + \frac{h_n^2}{t_n t_{n+1}^2} \right).$$

Using $|y_n - y(t_n)| \leq 10^{-6}$ and $h_n \leq t_n t_{n+1}^2$, it follows

$$|y_{n+1} - y(t_{n+1})| \leq \frac{1}{1 + 10^6 h_n} (10^{-6} + h_n) = 10^{-6}.$$

6.5 EXERCISE 6.9

Calculate the actual values of the coefficients of the 3-step Adams-Bashforth method.

Solution

The Adams-Bashforth methods are explicit multistep methods. Thus the formula is

$$\sum_{l=0}^{3} \rho_l \mathbf{y}_{n+l} = h \sum_{l=0}^{2} \sigma_l \mathbf{f}(t_{n+l}, \mathbf{y}_{n+l}).$$

For the Adams methods we have for the left hand side the associated polynomial

$$\rho(w) = \sum_{l=0}^{3} \rho_l w^l = w^2(w-1) = w^3 - w^2.$$

For the right hand side we can write the associated polynomial as

$$\sigma(w) = \sum_{l=0}^{2} \sigma_l w^l = \sigma_2 w^2 + \sigma_1 w + \sigma_0.$$

The coefficients $\sigma_0, \sigma_1, \sigma_2$ have to be chosen such that $\rho(e^z) - z\sigma(e^z) = O(z^4)$. Hence we have

$$\begin{aligned} e^{3z} - e^{2z} - z(\sigma_2 e^{2z} + \sigma_1 e^z + \sigma_0) = \quad & 1 + 3z + \tfrac{9}{2}z^2 + \tfrac{9}{2}z^3 - 1 - 2z - 2z^2 - \tfrac{4}{3}z^3 \\ & -z\sigma_2\left(1 + 2z + 2z^2\right) - z\sigma_1\left(1 + z + \tfrac{1}{2}z^2\right) \\ & -z\sigma_0 + O(z^4) \\ = \quad & z(1 - \sigma_0 - \sigma_1 - \sigma_2) + z^2(\tfrac{5}{2} - \sigma_1 - 2\sigma_2) \\ & +z^3(\tfrac{19}{6} - \tfrac{1}{2}\sigma_1 - 2\sigma_2) + O(z^4). \end{aligned}$$

This gives three equations which can be solved easily to give $\sigma_1 = -\frac{4}{3}, \sigma_2 = \frac{23}{12}$ and $\sigma_0 = \frac{5}{12}$.

6.6 EXERCISE 6.11

Show that the truncation error of methods given by

$$\begin{array}{c|cc} 0 & & \\ \alpha & \alpha & \\ \hline & (1 - \frac{1}{2\alpha}) & \frac{1}{2\alpha} \end{array} \tag{6.1}$$

is minimal for $\alpha = \frac{2}{3}$. Also show that no such method has order 3 or above.

Solution

Using the short hand notation $f_t = \frac{\partial f}{\partial t}$ and $f_y = \frac{\partial f}{\partial y}$, the Taylor expansion of y about t with step h is

$$y(t+h) = y + hf + \frac{1}{2}h^2(f_t + ff_y) + \frac{1}{6}h^3(f_{tt} + 2f_{ty}f + f_{yy}f^2 + f_t f_y + f_y^2 f) + O(h^4)$$

The steps of the method are

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + \alpha h, y_n + \alpha h k_1) = f(t_n + \alpha h, y_n + \alpha h f(t_n, y_n)). \end{aligned}$$

The next step is given by

$$\begin{aligned} y_{n+1} = &\; y_n + h[(1 - \tfrac{1}{2\alpha})k_1 + \tfrac{1}{2\alpha}k_2] \\ = &\; y_n + h[(1 - \tfrac{1}{2\alpha})f(t_n, y_n) + \tfrac{1}{2\alpha}f(t_n + \alpha h, y_n + \alpha h f(t_n, y_n))] \\ = &\; y_n + h(1 - \tfrac{1}{2\alpha} + \tfrac{1}{2\alpha})f + h^2\alpha\tfrac{1}{2\alpha}[f_t + f_y f] \\ &\; + \tfrac{1}{2}h^3\alpha^2\tfrac{1}{2\alpha}[f_{tt} + 2f_{ty}f + f_{yy}f^2] + O(h^4) \\ = &\; y_n + hf + \tfrac{1}{2}h^2[f_t + f_y f] + \tfrac{1}{4}\alpha h^3[f_{tt} + 2f_{ty}f + f_{yy}f^2] + O(h^4) \end{aligned}$$

So the first three terms agree with the first three terms of the Taylor expansion of $y(t + h)$ and cancel when the local truncation error is calculated. The h^3 term of the local truncation error is given by

$$-\frac{1}{4}\alpha h^3(f_{tt} + 2f_{ty}f + f_{yy}f^2) + \frac{1}{6}h^3(f_{tt} + 2f_{ty}f + f_{yy}f^2 + f_t f_y + f_y^2 f)$$

For the choice $\alpha = \frac{2}{3}$ all terms apart from $f_t f_y$ and $f_y^2 f$ vanish. Thus unless f is independent of y no higher order can be achieved.

6.7 EXERCISE 6.13

Consider the predictor-corrector pair given by

$$\begin{aligned} \mathbf{y}^P_{n+3} &= -\tfrac{1}{2}\mathbf{y}_n + 3\mathbf{y}_{n+1} - \tfrac{3}{2}\mathbf{y}_{n+2} + 3h\mathbf{f}(t_{n+2}, \mathbf{y}_{n+2}), \\ \mathbf{y}^C_{n+3} &= \tfrac{1}{11}[2\mathbf{y}_n - 9\mathbf{y}_{n+1} + 18\mathbf{y}_{n+2} + 6h\mathbf{f}(t_{n+3}, \mathbf{y}_{n+3})]. \end{aligned}$$

The predictor is as in exercise 6.4.The corrector is the three step backwards differentiation formula. Show that both methods are third order, and that the estimate of the error of the corrector formula by Milne's device has the value $\frac{6}{17}|\mathbf{y}^P_{n+3} - \mathbf{y}^C_{n+3}|$.

Solution

Letting D be the differential operator with respect of t. Putting the true solution into the formula of the predictor, we obtain the local truncation error

$\mathbf{e}^P$ as

$$\begin{aligned}
\mathbf{e}^P &= \mathbf{y}(t_{n+3}) + \tfrac{1}{2}\mathbf{y}(t_n) - 3\mathbf{y}(t_{n+1}) + \tfrac{3}{2}\mathbf{y}(t_{n+2}) - 3h\mathbf{f}(t_{n+2}, \mathbf{y}(t_{n+2})) \\
&= \left[e^{3hD} + \tfrac{1}{2} - 3e^{hD} + \tfrac{3}{2}e^{2hD} - 3hDe^{2hD}\right]\mathbf{y} \\
&= \big[1 + 3hD + \tfrac{9}{2}h^2D^2 + \tfrac{9}{2}h^3D^3 + \tfrac{27}{8}h^4D^4 \\
&\quad + \tfrac{1}{2} - 3 - 3hD - \tfrac{3}{2}h^2D^2 - \tfrac{1}{2}h^3D^3 - \tfrac{1}{8}h^4D^4 \\
&\quad + \tfrac{3}{2} + 3hD + 3h^2D^2 + 2h^3D^3 + h^4D^4 \\
&\quad - 3hD\left(1 + 2hD + 2h^2D^2 + \tfrac{4}{3}h^3D^3\right) + O(h^5D^5)\big]\mathbf{y} \\
&= \tfrac{1}{4}h^4\mathbf{y}''''(t_n) + O(h^5).
\end{aligned}$$

On the other hand the local truncation error of the corrector is

$$\begin{aligned}
\mathbf{e}^C &= \mathbf{y}(t_{n+3}) - \tfrac{1}{11}\left(2\mathbf{y}(t_n) - 9\mathbf{y}(t_{n+1}) + 18\mathbf{y}(t_{n+2}) + 6h\mathbf{f}(t_{n+3}, \mathbf{y}(t_{n+3}))\right) \\
&= \left[e^{3hD} - \tfrac{1}{11}\left(2 - 9e^{hD} + 18e^{2hD} + 6hDe^{3hD}\right)\right]\mathbf{y} \\
&= \big[1 + 3hD + \tfrac{9}{2}h^2D^2 + \tfrac{9}{2}h^3D^3 + \tfrac{27}{8}h^4D^4 \\
&\quad - \tfrac{1}{11}\left(2 - 9 - 9hD - \tfrac{9}{2}h^2D^2 - \tfrac{3}{2}h^3D^3 - \tfrac{3}{8}h^4D^4\right. \\
&\quad 18 + 36hD + 36h^2D^2 + 24h^3D^3 + 12h^4D^4 \\
&\quad \left. + 6hD\left(1 + 3hD + \tfrac{9}{2}h^2D^2 + \tfrac{9}{2}h^3D^3\right)\right) + O(h^5D^5)\big]\mathbf{y} \\
&= -\tfrac{3}{22}h^4\mathbf{y}''''(t_n) + O(h^5).
\end{aligned}$$

Thus we can write

$$\begin{aligned}
\mathbf{y}^P_{n+3} &\approx \mathbf{y}(t_{n+3}) + \tfrac{1}{4}h^4\mathbf{y}''''(t_n), \\
\mathbf{y}^C_{n+3} &\approx \mathbf{y}(t_{n+3}) - \tfrac{3}{22}h^4\mathbf{y}''''(t_n).
\end{aligned}$$

It follows that $h^4\mathbf{y}''''(t_n)$ is approximately $\frac{44}{17}(\mathbf{y}^P_{n+3} - \mathbf{y}^C_{n+3})$. The error in the corrector is then $-\frac{3}{22}\frac{44}{17}(\mathbf{y}^P_{n+3} - \mathbf{y}^C_{n+3}) = \frac{6}{17}(\mathbf{y}^P_{n+3} - \mathbf{y}^C_{n+3})$.

6.8 EXERCISE 6.15

Consider the scalar ordinary differential $y' = f(y)|$, that is f is independent of t. It is solved by the following Runge-Kutta method

$$\begin{aligned}
k_1 &= f(y_n), \\
k_2 &= f(y_n + (1-\alpha)hk_1 + \alpha hk_2), \\
y_{n+1} &= y_n + \tfrac{h}{2}(k_1 + k_2),
\end{aligned}$$

where α is a real parameter.

(a) Express the first, second and third derivative of y in terms of f.

(b) Perform the Taylor expansion of $y(t_{n+1})$ using the expressions found in the previous part and explain what it means for the method to be of order p.

(c) Determine p for the given Runge-Kutta method.

(d) Define A-stability, stating explicitly the linear test equation.

(e) Suppose the Runge-Kutta method is applied to the linear test equation. Show that then

$$y_{n+1} = R(h\lambda)y_n$$

and derive $R(h\lambda)$ explicitly.

(f) Show that the method is A-stable if and only if $\alpha = \frac{1}{2}$.

Solution

(a) In the case when f is independent of t, i.e. $y' = f(y)$, we have

$$\begin{aligned} y'' &= f'(y)y' = f'(y)f(y), \\ y''' &= f''(y)[f(y)]^2 + [f'(y)]^2 f(y). \end{aligned}$$

(b) The Taylor expansion is

$$\begin{aligned} y(t_{n+1}) &= y + hf(y) + \tfrac{1}{2}h^2 f'(y)f(y) \\ &\quad + \tfrac{1}{6}h^3 \left[f''(y)[f(y)]^2 + [f'(y)]^2 f(y)\right] + O(h^4). \end{aligned}$$

The method is of order p if the expansions of $y(t_{n+1})$ and of y_{n+1} given by the method agree for all terms up to h^p such that the local error $y(t_{n+1}) - y_{n+1}$ is $O(h^{p+1})$.

(c) First we expand the stages of the Runge-Kutta method

$$\begin{aligned} k_1 &= f(y_n), \\ k_2 &= f(y_n) + hf'(y_n)\left[(1-\alpha)k_1 + \alpha k_2\right] \\ &\quad + \tfrac{1}{2}h^2 f''(y_n)\left[(1-\alpha)k_1 + \alpha k_2\right]^2 + O(h^3). \end{aligned}$$

Now

$$\begin{aligned} (1-\alpha)k_1 + \alpha k_2 &= (1-\alpha)f(y_n) + \alpha f(y_n) \\ &\quad + \alpha h f'(y_n)\left[(1-\alpha)k_1 + \alpha k_2\right] + O(h^2) \\ &= f(y_n) + \alpha h f'(y_n)\left[(1-\alpha)k_1 + \alpha k_2\right] + O(h^2). \end{aligned}$$

Inserting this into the expression for k_2 we obtain

$$\begin{aligned} k_2 &= f(y_n) + hf'(y_n)f(y_n) \\ &\quad + h^2\left[\alpha f'(y_n)^2\left[(1-\alpha)k_1 + \alpha k_2\right] + \tfrac{1}{2}f''(y_n)\left[(1-\alpha)k_1 + \alpha k_2\right]^2\right] \\ &\quad + O(h^3). \end{aligned}$$

Also $(1-\alpha)k_1 + \alpha k_2 = f(y_n) + O(h)$ which can be inserted into the last expression for k_2 to give

$$\begin{aligned} k_2 &= f(y_n) + hf'(y_n)f(y_n) + \\ &\quad h^2\left[\alpha[f'(y_n)]^2 f(y_n) + \tfrac{1}{2}f''(y_n)[f(y_n)]^2\right] + O(h^3). \end{aligned}$$

Now we can obtain the expansion for y_{n+1}

$$\begin{aligned} y_{n+1} &= y_n + hf(y_n) + \tfrac{1}{2}h^2 f'(y_n)f(y_n) \\ &\quad + \tfrac{1}{2}h^3\left[\alpha[f'(y_n)]^2 f(y_n) + \tfrac{1}{2}f''(y_n)[f(y_n)]^2\right] + O(h^4). \end{aligned}$$

Assuming $y_n = y(t_n)$ then comparing the two Taylor expansions we see that the order is 2 regardless of the value of α.

(d) Suppose that a numerical method is applied to the linear test equation $y' = \lambda y$ with initial condition $y(0) = 1$ and produces the solution sequence $\{y_n\}_{n\in\mathbb{Z}^+}$ for constant h. We call the set

$$D = \{h\lambda \in \mathbb{C} : \lim_{n\to\infty} y_n = 0\}$$

the *linear stability domain* of the method. The set of $\lambda \in \mathbb{C}$ for which $y(t) \overset{t\to\infty}{\to} 0$ is the *exact stability set* and is the left half-plane $\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Re}z < 0\}$. We say that the method is *A-stable* if $\mathbb{C}^- \subseteq D$.

(e) Since $f(y) = \lambda y$, we have

$$\begin{aligned} k_1 &= \lambda y_n, \\ k_2 &= \lambda\left[y_n + (1-\alpha)hk_1 + \alpha hk_2\right]. \end{aligned}$$

Inserting the expression for k_1 into the expression for k_2 and solving for k_2 gives

$$k_2 = \frac{\lambda[1 + (1-\alpha)h\lambda]y_n}{1 - \alpha h\lambda}.$$

Therefore

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{2}(k_1 + k_2) \\ &= \left[1 + \frac{h}{2}(\lambda + \frac{\lambda[1 + (1-\alpha)h\lambda]}{1 - \alpha h\lambda}\right] y_n \\ &= \frac{1 + (1-\alpha)h\lambda + (\frac{1}{2} - \alpha)h^2\lambda^2}{1 - \alpha h\lambda} y_n. \end{aligned}$$

Thus

$$R(h\lambda) = \frac{1 + (1-\alpha)h\lambda + (\frac{1}{2} - \alpha)h^2\lambda^2}{1 - \alpha h\lambda}.$$

(f) The method is A-stable if and only if $|R(h\lambda)| < 1$ for all $h\lambda$ in the left complex half plane. For $\alpha = \frac{1}{2}$, then

$$R(h\lambda) = \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda},$$

which is the stability function of the trapezoidal rule. The trapezoidal rule is A-stable and thus is the Runge-Kutta method in this case. For $\alpha \neq \frac{1}{2}$, we consider the real axis. As $h\lambda$ approaches $-\infty$ the $(h\lambda)^2$ term will cause the numerator to grow quicker than the denominator and hence $|R(h\lambda)|$ becomes unbounded and thus A-stability is impossible.

6.9 EXERCISE 6.17

Consider the multistep method for numerical solution of the differential equation $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$*:*

$$\sum_{l=0}^{s} \rho_l \mathbf{y}_{n+l} = h \sum_{l=0}^{s} \sigma_l \mathbf{f}(t_{n+l}, \mathbf{y}_{n+l}), \qquad n = 0, 1, \ldots.$$

(a) Describe in general what it means that a method is of order p?

(b) Define generally the convergence of a method.

(c) Define the stability region and A-stability in general.

(d) Describe how to determine the stability region of the multistep method.

(e) Show that the method is of order p if

$$\sum_{l=0}^{s} \rho_l = 0, \qquad \sum_{l=0}^{s} l^k \rho_l = k \sum_{l=0}^{s} l^{k-1} \sigma_l, \qquad k = 1, 2, \ldots, p,$$

(f) Give the conditions on $\rho(w) = \sum_{l=0}^{s} \rho_l w^l$ *that ensure convergence.*

(g) Hence determine for what values of θ *and* $\sigma_0, \sigma_1, \sigma_2$ *the two-step method*

$$\mathbf{y}_{n+2} - (1-\theta)\mathbf{y}_{n+1} - \theta \mathbf{y}_n = h[\sigma_0 \mathbf{f}(t_n, \mathbf{y}_n) + \sigma_1 \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) + \sigma_2 \mathbf{f}(t_{n+2}, \mathbf{y}_{n+2})]$$

is convergent and of order 3*.*

Solution

(a) If

$$\mathbf{y}_{n+1} = \phi_h(t_n, \mathbf{y}_0, \mathbf{y}_1, \dots \mathbf{y}_{n+1})$$

for the time step h, then the order is the largest integer p such that

$$\mathbf{y}(t_{n+1}) - \phi_h(t_n, \mathbf{y}(t_0), \mathbf{y}(t_1), \dots \mathbf{y}(t_{n+1})) = O(h^{p+1}).$$

(b) Let $t^* > 0$ be fixed. A method which produces for every $h > 0$ the solution sequence $\mathbf{y}_n = \mathbf{y}_n(h)$, $n = 0, 1, \dots, \lfloor t^*/h \rfloor$, converges, if, as $h \to 0$ and $n_k(h)h \overset{k\to\infty}{\longrightarrow} t$, it is true that $\mathbf{y}_{n_k} \to \mathbf{y}(t)$, the exact solution, uniformly for $t \in [0, t^*]$.

(c) Suppose that a numerical method is applied to the test equation $y' = \lambda y$ with initial condition $y(0) = 1$ and produces the solution sequence $\{y_n\}_{n\in\mathbb{Z}^+}$ for constant h. We call the set

$$D = \{h\lambda \in \mathbb{C} : \lim_{n\to\infty} y_n = 0\}$$

the *linear stability domain* of the method. The set of $\lambda \in \mathbb{C}$ for which $y(t) \overset{t\to\infty}{\to} 0$ is the *exact stability set* and is the left half-plane $\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Re} z < 0\}$. We say that the method is *A-stable* if $\mathbb{C}^- \subseteq D$.

(d) When the multistep method is applied to the test equation $y' = \lambda y$, $y(0) = 1$, it reads

$$\sum_{l=0}^{s} (\rho_l - h\lambda\sigma_l) y_{n+l} = 0.$$

This recurrence relation has the characteristic polynomial

$$\tau(w) = \sum_{l=0}^{s} (\rho_l - h\lambda\sigma_l) w^l.$$

Let its zeros be $w_1(h\lambda), \dots, w_{N(h\lambda)}(h\lambda)$ with multiplicities $\mu_1(h\lambda), \dots, \mu_{N(h\lambda)}(h\lambda)$ respectively, where the multiplicities sum to the order of the polynomial τ. The solutions of the recurrence relation are given by

$$y_n = \sum_{j=1}^{N(h\lambda)} \sum_{i=0}^{\mu_j(h\lambda)-1} n^i w_j(h\lambda)^n \alpha_{ij}(h\lambda),$$

where $\alpha_{ij}(h\lambda)$ are independent of n but depending on the starting values $y_0, \dots, y_{s-1}$. Hence the linear stability domain is the set of all $h\lambda \in \mathbb{C}$ such that all the zeros of the characteristic polynomial satisfy $|w_j(h\lambda)| \le 1$ and if $|w_j(h\lambda)| = 1$, then $\mu_j(h\lambda) = 1$.

(e) The order is obtained from

$$\sum_{l=0}^{s} \rho_l \mathbf{y}(t_{n+l}) - h \sum_{l=0}^{s} \sigma_l \mathbf{y}'(t_{n+l}) =$$
$$\sum_{l=0}^{s} \rho_l \sum_{k=0}^{\infty} \frac{(lh)^k}{k!} \mathbf{y}^{(k)}(t_n) - h \sum_{l=0}^{s} \sigma_l \sum_{k=0}^{\infty} \frac{(lh)^k}{k!} \mathbf{y}^{(k+1)}(t_n)$$
$$\left(\sum_{l=0}^{s} \rho_l \right) \mathbf{y}(t_n) + \sum_{k=1}^{\infty} \frac{1}{k!} \left(\sum_{l=0}^{s} l^k \rho_l - k \sum_{l=0}^{s} l^{k-1} \sigma_l \right) h^k \mathbf{y}^{(k)}(t_n).$$

Thus the difference is $O(h^{p+1})$ if

$$\sum_{l=0}^{s} \rho_l = 0, \qquad \sum_{l=0}^{s} l^k \rho_l = k \sum_{l=0}^{s} l^{k-1} \sigma_l, \qquad k = 1, 2, \ldots, p.$$

(f) The multistep method is convergent if and only if it is of order $p \geq 1$ and the polynomial $\rho(w) = \sum_{l=0}^{s} \rho_l w^l$ obeys the root condition which means all its zeros lie within $|w| \leq 1$ and all zeros of unit modulus are simple zeros. In this case the method is sometimes called zero-stable.

(g) For the multistep method given we have $\rho(w) = w^2 + (\theta - 1)w - \theta$ which has roots 1 and $-\theta$. Thus the root condition is satisfied, if $\theta \in (-1, 1]$.

To achieve order 3 we have $\rho_0 = -\theta, \rho_1 = \theta - 1$ and $\rho_2 = 1$ and hence $\rho_0 + \rho_1 + \rho_2 = 1$ and thus the first condition is fulfilled.

The other order conditions give the equations

$$\begin{aligned} \sigma_0 + \sigma_1 + \sigma_2 &= \rho_1 + 2\rho_2 = \theta + 1 \\ 2\sigma_1 + 4\sigma_2 &= \rho_1 + 4\rho_2 = \theta + 3 \\ 3\sigma_1 + 12\sigma_2 &= \rho_1 + 8\rho_2 = \theta + 7. \end{aligned}$$

Subtracting the third equation from 3 times the second equation yields $3\sigma_1 = 2\theta + 2$. Inserting this into the third equation gives $12\sigma_2 = -\theta + 5$. Thus $\sigma_1 = 2(\theta + 1)/3$ and $\sigma_2 = (5 - \theta)/12$. Inserting this into the first equation we have $\sigma_0 = (5\theta - 1)/12$.

6.10 EXERCISE 6.19

We consider the autonomous scalar differential equation

$$\frac{d}{dt} y(t) = y'(t) = f(y(t)), \qquad y(0) = y_0.$$

Note that f is independent of t.

(a) Express the second and third derivative of y in terms of f and its derivatives. Write the Taylor expansion of $y(t+h)$ in terms of f and its derivatives up to $O(h^4)$.

(b) The differential equation is solved by the Runge-Kutta scheme

$$\begin{aligned} k_1 &= hf(y_n), \\ k_2 &= hf(y_n + k_1), \\ y_{n+1} &= y_n + \tfrac{1}{2}(k_1 + k_2). \end{aligned}$$

Show that the scheme is of order 2.

(c) Define the linear stability domain and A-stability for a general numerical method, stating explicitly the linear test equation on which the definitions are based.

(d) Apply the Runge-Kutta scheme given in (b) to the linear test equation from part (c) and find an expression for the linear stability domain of the method. Is the method A-stable?

(e) We now modify the Runge-Kutta scheme in the following way

$$\begin{aligned} k_1 &= hf(y_n), \\ k_2 &= hf(y_n + a(k_1 + k_2)), \\ y_{n+1} &= y_n + \tfrac{1}{2}(k_1 + k_2), \end{aligned}$$

where $a \in \mathbb{R}$. Apply it to the test equation and find a rational function R such that $y_{n+1} = R(h\lambda)y_n$.

(f) Explain the maximum modulus principle and use it to find the values of a such that the method given in (e) is A-stable.

Solution

(a) Given the autonomous scalar differential equation

$$\frac{d}{dt}y(t) = y'(t) = f(y(t)), \qquad y(0) = y_0,$$

we have

$$\begin{aligned} \frac{d^2}{dt^2}y(t) = y''(t) &= \frac{d}{dt}[f(y(t))] = \frac{d}{dy}f(y(t))\frac{d}{dt}y(t) = f'(y(t))f(y(t)), \\ \frac{d^3}{dt^3}y(t) = y'''(t) &= \frac{d}{dt}[f'(y(t))f(y(t))] \\ &= \frac{d}{dt}[f'(y(t))]f(y(t)) + f'(y(t))\frac{d}{dt}[f(y(t))] \\ &= \frac{d}{dy}f'(y(t))\frac{d}{dt}y(t)f(y(t)) + f'(y(t))\frac{d}{dy}f(y(t))\frac{d}{dt}y(t) \\ &= f''(y(t))[f(y(t))]^2 + [f'(y(t))]^2 f(y(t)). \end{aligned}$$

Hence

$$\begin{aligned} y(t+h) &= y(t) + hf(y(t) + \frac{1}{2}h^2 f'(y(t))f(y(t)) + \\ &\quad + \frac{1}{6}h^3 \left[f''(y(t))[f(y(t))]^2 + [f'(y(t))]^2 f(y(t)) \right] + O(h^4). \end{aligned}$$

(b) The differential equation is solved by the Runge-Kutta Scheme

$$\begin{aligned} k_1 &= hf(y_n), \\ k_2 &= hf(y_n + k_1), \\ y_{n+1} &= y_n + \tfrac{1}{2}(k_1 + k_2). \end{aligned}$$

Expanding k_2 gives

$$\begin{aligned} k_2 &= hf(y_n + k_1) = h[f(y_n) + k_1 f'(y_n) + \frac{1}{2}k_1^2 f''(y_n) + O(k_1^3)] \\ &= hf(y_n) + h^2 f(y_n)f'(y_n) + \frac{1}{2}h^3[f(y_n)]^2 f''(y_n) + O(h^4). \end{aligned}$$

Assuming $y_n = y(t_n)$ we get

$$\begin{aligned} & y(t_n) + \frac{1}{2}(k_1 + k_2) = \\ & = y(t_n) + \frac{1}{2}\Big(hf(y(t_n)) + hf(y(t_n)) + h^2 f(y(t_n))f'(y(t_n)) \\ & + \frac{1}{2}h^3[f(y(t_n))]^2 f''(y(t_n)) + O(h^4) \Big) \\ & = y(t_n) + hy'(t_n) + \frac{1}{2}h^2 y''(t_n) + \frac{1}{4}h^3[f(y(t_n))]^2 f''(y(t_n)) + O(h^4) \\ & = y(t_n + h) + O(h^3), \end{aligned}$$

since the h^3 term does not match the third derivative term in the Taylor expansion of $y(t_n + h)$. Thus the method is of order 2.

(c) Suppose that a numerical method is applied to the linear test equation $y' = \lambda y$ with initial condition $y(0) = 1$ and produces the solution sequence $\{y_n\}_{n\in\mathbb{Z}^+}$ for constant h. We call the set

$$D = \{h\lambda \in \mathbb{C} : \lim_{n\to\infty} y_n = 0\}$$

the linear stability domain of the method. The set of $h\lambda \in \mathbb{C}$ for which $y(t) \overset{t\to\infty}{\to} 0$ is the exact stability set and is the left half-plane $\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Re}z < 0\}$. We say that the method is A-stable if $\mathbb{C}^- \subseteq D$.

(d)

$$\begin{aligned} k_1 &= hf(y_n) = h\lambda y_n, \\ k_2 &= hf(y_n + k_1) = h\lambda(y_n + k_1) = h\lambda y_n + (h\lambda)^2 y_n, \\ y_{n+1} &= y_n + \tfrac{1}{2}(k_1 + k_2) = y_n + \tfrac{1}{2}(h\lambda y_n + h\lambda y_n + (h\lambda)^2 y_n) \\ &= y_n[1 + h\lambda + \tfrac{1}{2}(h\lambda)^2]. \end{aligned}$$

Thus

$$y_n = \left[1 + h\lambda + \frac{1}{2}(h\lambda)^2\right]^n y_0$$

and the linear stability domain is

$$D = \{h\lambda \in \mathbb{C} : |1 + h\lambda + \frac{1}{2}(h\lambda)^2| < 1\}.$$

The method is not A-stable, since for $h\lambda = -2$ the factor becomes $1 + h\lambda + \frac{1}{2}(h\lambda)^2 = 1$ and thus $y_{n+1} = y_n = \ldots = y_0 = 1$ and the sequence does not tend to zero. So -2 does not lie in the linear stability domain, but $-2 \in \mathbb{C}^-$.

(e) For the modified Runge-Kutta scheme

$$\begin{aligned} k_1 &= hf(y_n) = h\lambda y_n, \\ k_2 &= hf(y_n + a(k_1 + k_2)) = h\lambda(y_n + a(k_1 + k_2)) \\ &= h\lambda y_n + a(h\lambda)^2 y_n + ah\lambda k_2. \end{aligned}$$

Solving the second equation for k_2 we get

$$k_2 = \frac{1 + ah\lambda}{1 - ah\lambda} h\lambda y_n.$$

Next

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{1}{2}(k_1 + k_2) \\ &= y_n + \tfrac{1}{2}(h\lambda y_n + \frac{1 + ah\lambda}{1 - ah\lambda} h\lambda y_n) \\ &= y_n \left[1 + \tfrac{1}{2} h\lambda \frac{1 - ah\lambda + 1 + ah\lambda}{1 - ah\lambda}\right] \\ &= y_n \left[1 + \frac{h\lambda}{1 - ah\lambda}\right] \\ &= y_n \frac{1 + (1 - a)h\lambda}{1 - ah\lambda} \end{aligned}$$

Thus the rational function R is given by

$$R(z) = \frac{1 + (1 - a)z}{1 - az}.$$

(f) According to the maximum modulus principle, if g is an analytic function in the closed complex domain V, then $|g|$ attains its maximum on the boundary ∂V. We let $g = R$. If $a = 0$, R has no singularities. Otherwise its only singularity is the pole $1/a$, which is the root of the denominator. Thus R is only analytic in $V = \{z \in \mathbb{C} : \mathrm{Re} z \leq 0\}$, the closure of $\mathbb{C}^-$,

if $a \geq 0$. Then it attains its maximum on $\partial V = i\mathbb{R}$ and the following statements are equivalent

$$\text{A-stability} \quad \Leftrightarrow \quad |R(z)| < 1, \quad z \in \mathbb{C}^- \quad \Leftrightarrow \quad |R(\mathrm{i}t)| \leq 1, \quad t \in \mathbb{R}.$$

We have

$$|R(it)| = \left| \frac{1 + (1-a)\mathrm{i}t}{1 - a\mathrm{i}t} \right| = \frac{1 + (1-a)^2 t^2}{1 + a^2 t^2} = 1 + \frac{(1-2a)t^2}{1 + a^2 t^2}.$$

Thus to have $|R(\mathrm{i}t)| \leq 1$ we need $a \geq 1/2$.

CHAPTER 7

Numerical Differentiation Exercises

7.1 EXERCISE 7.1

List the assumptions made in the analysis and give an example where at least one of these assumptions does not hold. What does this mean in practice for the approximation of derivatives?

Solution

Firstly, it was assumed that h is small enough such that the $O(h^2)$ term can be neglected in the expression for the discretization error. Secondly it was assumed that the calculation of the approximation does not introduce any further error. Only the error in the representation of $f(x)$ and $f(x+h)$ was considered. Another assumption is that $f(x)$ and $f(x+h)$ can be evaluated to similar accuracy. We have seen when analyzing the condition of a problem that this might not be the case. Further we assumed that both $f(x) + f(x+h) = O(1)$ and $f''(x) = O(1)$. If there are singularities nearby or dramatic changes any of the assumptions may not hold. As an example consider $f(x) = x^{30}$ and evaluating $f'(x)$ for $x = 1$. The order of $f(1) = 1$ is $O(1)$ and we could argue this still for a point $1 + h$ close-by. However, $f''(1) = 30 \times 29 = 870 = O(100)$ which is larger by two magnitudes. Because of this h has to be chosen smaller.

CHAPTER 8

PDEs Exercises

8.1 EXERCISE 8.1

Consider the PDE

$$au_{xx} + 2bu_{xy} + cu_{yy} = f,$$

where $a > 0, b, c > 0$ *and* f *are functions of* x, y, u, u_x *and* u_y. *At* (x, y) *the PDE is*

elliptic, if $b^2 - ac < 0$,

hyperbolic, if $b^2 - ac > 0$ *and*

parabolic, if $b^2 - ac = 0$.

Show that this definition is equivalent to the eigenvalue definition given in lectures.

Solution

The associated matrix is

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

It is symmetric and thus has real eigenvalues specified by

$$(a - \lambda)(c - \lambda) - b^2 = 0 \quad \Leftrightarrow \quad \lambda^2 - (a + c)\lambda + ac - b^2.$$

The eigenvalues are

$$\lambda_{12} = \frac{1}{2}(a + c) \pm \frac{1}{2}\sqrt{(a + c)^2 + 4(b^2 - ac)}$$

Then

- Both eigenvalues are positive for $b^2 - ac < 0 \Rightarrow$ elliptic.
- We have one positive and one negative eigenvalue if $b^2 - ac > 0 \Rightarrow$ hyperbolic.
- There is a zero eigenvalue for $b^2 - ac = 0 \Rightarrow$ parabolic

8.2 EXERCISE 8.3

The Crank–Nicolson formula is applied to the heat equation $u_t = u_{xx}$ *on a rectangular mesh* $(m\Delta x, n\Delta t)$, $m = 0, 1, ..., M + 1$, $n = 0, 1, 2, ...,$ *where* $\Delta x = 1/(M + 1)$. *We assume zero boundary conditions* $u(0, t) = u(1, t) = 0$ *for all* $t \geq 0$. *Prove that the estimates* $u_n^m \approx u(m\Delta x, n\Delta t)$ *satisfy the equation*

$$\sum_{m=1}^{M} [(u_m^{n+1})^2 - (u_m^n)^2] = -\frac{1}{2}\mu \sum_{m=1}^{M+1} (u_m^{n+1} + u_m^n - u_{m-1}^{n+1} - u_{m-1}^n)^2, \qquad n = 1, 2, \ldots.$$

This shows that $\sum_{m=1}^{M} (u_m^n)^2$ *is monotonically decreasing with increasing* n *and the numerical solution mimics the decaying behaviour of the exact solution.*

Solution

Firstly, rearranging the Crank-Nicolson formula we have

$$u_m^{n+1} - u_m^n = \frac{1}{2}\mu \left(u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1} + u_{m-1}^n - 2u_m^n + u_{m+1}^n\right)$$

Now

$$\begin{aligned}
&(u_m^{n+1})^2 - (u_m^n)^2 = (u_m^{n+1} - u_m^n)(u_m^{n+1} + u_m^n) \\
&= \tfrac{1}{2}\mu \left(u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1} + u_{m-1}^n - 2u_m^n + u_{m+1}^n\right)(u_m^{n+1} + u_m^n) \\
&= -\tfrac{1}{2}\mu(u_m^{n+1} + u_m^n - u_{m-1}^{n+1} - u_{m-1}^n)(u_m^{n+1} + u_m^n) \\
&\quad + \tfrac{1}{2}\mu(u_{m+1}^{n+1} + u_{m+1}^n - u_m^{n+1} - u_m^n)(u_m^{n+1} + u_m^n)
\end{aligned}$$

Summing over $m = 1, \ldots, M$ we have

$$\begin{aligned}
&\sum_{m=1}^{M} [(u_m^{n+1})^2 - (u_m^n)^2] = \\
&-\frac{1}{2}\mu \sum_{m=1}^{M} (u_m^{n+1} + u_m^n - u_{m-1}^{n+1} - u_{m-1}^n)(u_m^{n+1} + u_m^n) \\
&+\frac{1}{2}\mu \sum_{m=1}^{M} (u_{m+1}^{n+1} + u_{m+1}^n - u_m^{n+1} - u_m^n)(u_m^{n+1} + u_m^n).
\end{aligned}$$

We can extend the summation to $M + 1$ in the first sum due to the zero boundary conditions. In the second sum we change the summation index m

by 1. Thus

$$\sum_{m=1}^{M}[(u_m^{n+1})^2-(u_m^n)^2]=$$
$$-\frac{1}{2}\mu\sum_{m=1}^{M+1}(u_m^{n+1}+u_m^n-u_{m-1}^{n+1}-u_{m-1}^n)(u_m^{n+1}+u_m^n)$$
$$+\frac{1}{2}\mu\sum_{m=1}^{M+1}(u_m^{n+1}+u_m^n-u_{m-1}^{n+1}-u_{m-1}^n)(u_{m-1}^{n+1}+u_{m-1}^n)=$$
$$-\frac{1}{2}\mu\sum_{m=1}^{M+1}(u_m^{n+1}+u_m^n-u_{m-1}^{n+1}-u_{m-1}^n)^2$$

8.3 EXERCISE 8.5

Determine the order of the local error of the finite difference approximation to $\partial^2 u/\partial x\partial y$ which is given by the computational stencil

$$\frac{1}{(\Delta x)^2}\begin{array}{ccc} -\frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 \\ \frac{1}{4} & 0 & -\frac{1}{4} \end{array}$$

Solution

Using the same operators as in the previous exercise and inserting the true solution into the computational stencil we obtain

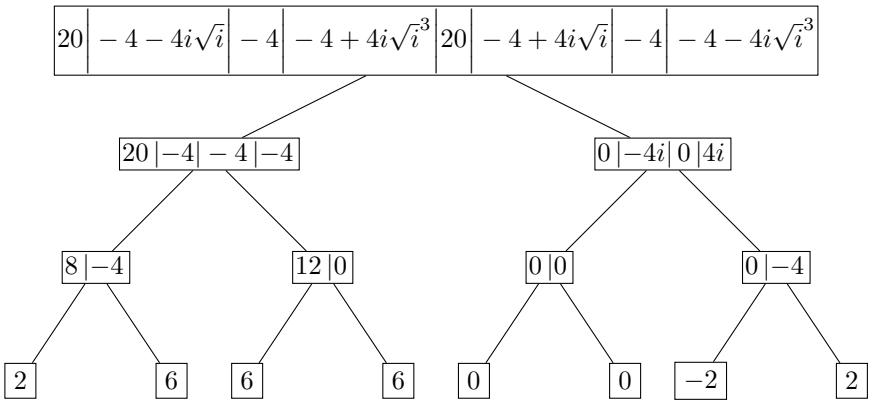
$$
\begin{aligned}
&\tfrac{1}{4}\left[u(x+\Delta x, y+\Delta x) - u(x+\Delta x, y-\Delta x)\right.\\
&\left.+u(x-\Delta x, y-\Delta x) - u(x-\Delta x, y+\Delta x)\right]\\
&= \tfrac{1}{4}\left[e^{\Delta x D_x} e^{\Delta x D_y} - e^{\Delta x D_x} e^{-\Delta x D_y} + e^{-\Delta x D_x} e^{-\Delta x D_y} - e^{-\Delta x D_x} e^{\Delta x D_y}\right] u(x,y)\\
&= \tfrac{1}{4}\left[\left(e^{\Delta x D_x} - e^{-\Delta x D_x}\right)\left(e^{\Delta x D_y} - e^{-\Delta x D_y}\right)\right] u(x,y)\\
&= \tfrac{1}{4}\left[\left(2\Delta x D_x + \tfrac{1}{3}(\Delta x D_x)^3 + O((\Delta x)^5)\right)\right.\\
&\left.\times\left(2\Delta x D_y + \tfrac{1}{3}(\Delta x D_y)^3 + O((\Delta x)^5)\right)\right] u(x,y)\\
&= \left[(\Delta x)^2 D_x D_y + \tfrac{1}{6}(\Delta x)^4 (D_x^3 D_y + D_x D_y^3) + O((\Delta x)^6)\right] u(x,y).
\end{aligned}
$$

Thus the order is $O((\Delta x)^2)$ because of the division by $(\Delta x)^2$.

8.4 EXERCISE 8.7

Let $(\hat{x}_0, \hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4, \hat{x}_5, \hat{x}_6, \hat{x}_7) = (2, 0, 6, -2, 6, 0, 6, 2)$*. By applying the inverse of the FFT algorithm, calculate* $x_l = \sum_{j=0}^{7} \omega_8^{jl} \hat{x}_j$ *for* $l = 0, 2, 4, 6$*, where* $\omega_8 = \exp \frac{2i\pi}{8}$.

Solution



8.5 EXERCISE 8.9

The function $u(x, y) = 18x(1-x)y(1-y), 0 \le x, y \le 1$*, is the solution of the Poisson equation* $u_{xx} + u_{yy} = 36(x^2 + y^2 - x - y) = f(x, y)$*, subject to zero boundary conditions. Let* $\Delta x = 1/6$ *and seek the solution of the five-point*

method

$$u_{m-1,n}+u_{m+1,n}+u_{m,n-1}+u_{m,n+1}-4u_{m,n} = (\Delta x)^2 f(mh, nh), \;\; 1 \le m, n \le 5,$$

where $u_{m,n}$ *is zero if one of* m, n *is* 0 *or* 6. *Let the multigrid method be applied, using only this fine grid and a coarse grid of mesh size* $1/3$, *and let every* $u_{m,n}$ *be zero initially. Calculate the* 25 *residuals of the starting vector on the fine grid. Then, following the restriction procedure, find the residuals for the initial calculation on the coarse grid. Solve the equations on the coarse grid exactly. The resultant estimates of* u *at the four interior points of the coarse grid all have the value* $5/6$. *By applying the prolongation operator to these estimates, find the* 25 *starting values of* $u_{m,n}$ *for the subsequent iterations of Jacobi on the fine grid. Further, show that if one Jacobi iteration is performed, then* $u_{3,3} = 23/24$ *occurs, which is the estimate of* $u(1/2, 1/2) = 9/8$.

Solution

Since the initial values are zero, the residuals are given by $(\Delta x)^2 f$ at the 25 interior points

$-\frac{10}{36}$	$-\frac{13}{36}$	$-\frac{14}{36}$	$-\frac{13}{36}$	$-\frac{10}{36}$
$-\frac{13}{36}$	$-\frac{16}{36}$	$-\frac{17}{36}$	$-\frac{16}{36}$	$-\frac{13}{36}$
$-\frac{14}{36}$	$-\frac{17}{36}$	$-\frac{18}{36}$	$-\frac{17}{36}$	$-\frac{14}{36}$
$-\frac{13}{36}$	$-\frac{16}{36}$	$-\frac{17}{36}$	$-\frac{16}{36}$	$-\frac{13}{36}$
$-\frac{10}{36}$	$-\frac{13}{36}$	$-\frac{14}{36}$	$-\frac{13}{36}$	$-\frac{10}{36}$

Restriction gives 4 values at the interior points of the coarse grid

	$-\frac{5}{3}$		$-\frac{5}{3}$	
	$-\frac{5}{3}$		$-\frac{5}{3}$	

The recurrence relations on the coarse grid simplify due to the zero boundary values

$$\begin{aligned}
u_{01}+u_{21}+u_{10}+u_{12}-4u_{11} = -\tfrac{5}{3} \;\; &\Rightarrow \;\; u_{21}+u_{12}-4u_{11} = -\tfrac{5}{3} \\
u_{02}+u_{22}+u_{11}+u_{13}-4u_{12} = -\tfrac{5}{3} \;\; &\Rightarrow \;\; u_{22}+u_{11}-4u_{12} = -\tfrac{5}{3} \\
u_{11}+u_{31}+u_{20}+u_{22}-4u_{21} = -\tfrac{5}{3} \;\; &\Rightarrow \;\; u_{11}+u_{22}-4u_{21} = -\tfrac{5}{3} \\
u_{12}+u_{32}+u_{21}+u_{23}-4u_{22} = -\tfrac{5}{3} \;\; &\Rightarrow \;\; u_{12}+u_{21}-4u_{22} = -\tfrac{5}{3}
\end{aligned}$$

Due to symmetry the solutions are $u_{11} = u_{21} = u_{21} = u_{22} = \frac{5}{6}$. Prolongating these 4 values gives

$\frac{5}{24}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{5}{24}$
$\frac{5}{12}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{5}{12}$
$\frac{5}{12}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{5}{12}$
$\frac{5}{12}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{5}{12}$
$\frac{5}{24}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{5}{24}$

One Jacobi iteration is

$$u_{33} = -\frac{1}{4}(-\frac{5}{6} - \frac{5}{6} - \frac{5}{6} - \frac{5}{6} - -\frac{18}{36}) = \frac{23}{24}$$

while $u(\frac{1}{2}, \frac{1}{2}) = 18\frac{1}{2}(1 - \frac{1}{2})\frac{1}{2}(1 - \frac{1}{2}) = \frac{9}{8}$.

8.6 EXERCISE 8.11

Let $F(t) = e^{tA}e^{tB}$ be the first order Beam-Warming splitting of $e^{t(A+B)}$. Generally the splitting error is of the form t^2C for some matrix C. If C has large eigenvalues the splitting error can be large even for small t. Show that

$$F(t) = e^{t(A+B)} + \int_0^t e^{(t-\tau)(A+B)} \left(e^{\tau A}B - Be^{\tau A}\right) e^{\tau B} d\tau.$$

(Hint: Find explicitly $G(t) = F'(t) - (A+B)F(t)$ and use variation of constants to find the solution of the linear matrix ODE $F' = (A+B)F + G$, $F(0) = I$.)

Suppose that a matrix norm $\|\cdot\|$ is given and that there exist real constants c_A, c_B and c_{A+B} such that

$$\|e^{tA}\| \leq e^{c_A t}, \qquad \|e^{tB}\| \leq e^{c_B t}, \qquad \|e^{t(A+B)}\| \leq e^{c_{A+B} t}.$$

Prove that

$$\|F(t) - e^{t(A+B)}\| \leq 2\|B\| \frac{e^{(c_A+c_B)t} - e^{c_{A+B}t}}{c_A + c_B - c_{A+B}}.$$

Hence, for $c_A, c_B \leq 0$, the splitting error remains relatively small even for large t. ($e^{c_{A+B}t}$ is an intrinsic error.)

Solution

Since $F(t) = e^{tA}e^{tB}$, the derivative is $F'(t) = Ae^{tA}e^{tB} + e^{tA}Be^{tB}$ by the product rule. Subtracting $(A+B)F(t)$ gives

$$\begin{aligned} G(t) &= Ae^{tA}e^{tB} + e^{tA}Be^{tB} - (A+B)e^{tA}e^{tB} \\ &= e^{tA}Be^{tB} - Be^{tA}e^{tB}. \end{aligned}$$

The ODE $F' = (A+B)F+G$, $F(0) = I$ has the solution of the form $e^{t(A+B)}c(t)$ which has the derivative

$$\begin{aligned} F'(t) &= (A+B)e^{t(A+B)}c(t) + e^{t(A+B)}c'(t) \\ &= (A+B)F(t) + G(t) \end{aligned}$$

Thus

$$c'(t) = e^{-t(A+B)}G = e^{-t(A+B)}(e^{tA}B - Be^{tA})e^{tB}$$

which has the solution

$$c(t) = c + \int_0^t e^{-\tau(A+B)}(e^{\tau A}B - Be^{\tau A})e^{\tau B}d\tau.$$

Hence F is given by

$$F(t) = ce^{t(A+B)} + \int_0^t e^{(t-\tau)(A+B)}\left(e^{\tau A}B - Be^{\tau A}\right)e^{\tau B}d\tau.$$

The initial condition $F(0) = I$ specifies the additive constant as $c = 1$.

For the second part of the question

$$\begin{aligned} \|F(t) - e^{t(A+B)}\| &= \|\int_0^t e^{(t-\tau)(A+B)}\left(e^{\tau A}B - Be^{\tau A}\right)e^{\tau B}d\tau\| \\ &\leq \int_0^t \|e^{(t-\tau)(A+B)}\|\,\|e^{\tau A}B - Be^{\tau A}\|\,\|e^{\tau B}\|d\tau \\ &\leq \int_0^t \|e^{t(A+B)}\|\,\|e^{-\tau(A+B)}\|\,2\,\|B\|\,\|e^{\tau A}\|\,\|e^{\tau B}\|d\tau \end{aligned}$$

Now $\|e^{tA}\| \leq e^{c_A t}$, $\|e^{tB}\| \leq e^{c_B t}$, $\|e^{t(A+B)}\| \leq e^{c_{A+B}t}$ and hence

$$\begin{aligned} \|F(t) - e^{t(A+B)}\| &\leq e^{c_{A+B}t}2\|B\|\int_0^t e^{(c_A+c_B-c_{A+B})\tau}d\tau \\ &= 2\|B\|e^{c_{A+B}t}\frac{1}{c_A+c_B-c_{A+B}}\left[e^{(c_A+c_B-c_{A+B})\tau}\right]_0^t \\ &= 2\|B\|\frac{e^{(c_A+c_B)t} - e^{c_{A+B}t}}{c_A+c_B-c_{A+B}}. \end{aligned}$$

8.7 EXERCISE 8.13

The diffusion equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a(x)\frac{\partial u}{\partial x}\right), \qquad 0 \leq x \leq 1, \qquad t \geq 0,$$

with the initial condition $u(x,0) = \phi(x)$, $0 \leq x \leq 1$ and zero boundary conditions for $x = 0$ and $x = 1$, is solved by the finite difference method

$$u_m^{n+1} = u_m^n + \mu\left[a_{m-\frac{1}{2}}u_{m-1}^n - (a_{m-\frac{1}{2}} + a_{m+\frac{1}{2}})u_m^n + a_{m+\frac{1}{2}}u_{m+1}^n\right],$$

where $m = 1, \ldots, M$, $\mu = \Delta t/(\Delta x)^2$ is constant, $\Delta x = \frac{1}{M+1}$ and u_m^n approximates $u(m\Delta x, n\Delta t)$. The notation $a_\alpha = a(\alpha\Delta x)$ is employed.

(a) Assuming sufficient smoothness of the function a, show that the local error of the method is at least $O((\Delta x)^3)$. State which expansions and substitutions you are using.

(b) Remembering the zero boundary conditions, write the method as

$$\mathbf{u}^{n+1} = A\mathbf{u}^n$$

giving a formula for the entries $A_{k,l}$ of A. From the structure of A what can you say about the eigenvalues of A?

(c) Describe the eigenvalue analysis of stability.

(d) Assume that there exist finite positive constants a_- and a_+ such that for $0 \le x \le 1$ $a(x)$ lies in the interval $[a_-, a_+]$. Prove that the method is stable for $0 < \mu \le \frac{1}{2a_+}$. (Hint: You may use without proof the Gerschgorin theorem: All eigenvalues of the matrix A are contained in the union of the Gerschgorin discs given for each $k = 1, \ldots, M$ by

$$\{z \in \mathbb{C} : |z - A_{k,k}| \le \sum_{l=1, l\neq k}^{M} |A_{k,l}|\}.)$$

Solution

(a) We employ the differential operators

$$D_t = \frac{\partial}{\partial t} \text{ and } D_x = \frac{\partial}{\partial x}.$$

The partial differential equation can then be rewritten as

$$D_t u = D_x[aD_x u] = D_x a D_x u + aD_x^2 u.$$

The local error is given by

$$
\begin{aligned}
&(e^{\Delta t D_t} - I)u - \mu \Big[e^{-\frac{1}{2}\Delta x D_x} a e^{-\Delta x D_x} u - (e^{-\frac{1}{2}\Delta x D_x} + e^{\frac{1}{2}\Delta x D_x})au \\
&+e^{\frac{1}{2}\Delta x D_x} a e^{\Delta x D_x} u\Big] \\
&= (e^{\Delta t D_t} - I)u - \mu \Big[-(e^{-\frac{1}{2}\Delta x D_x} + e^{\frac{1}{2}\Delta x D_x})au \\
&+e^{-\frac{1}{2}\Delta x D_x} a(I - \Delta x D_x + \tfrac{1}{2}(\Delta x D_x)^2 - \tfrac{1}{6}(\Delta x D_x)^3 + O((\Delta x)^4))u \\
&+e^{\frac{1}{2}\Delta x D_x} a(I + \Delta x D_x + \tfrac{1}{2}(\Delta x D_x)^2 + \tfrac{1}{6}(\Delta x D_x)^3 + O((\Delta x)^4))u\Big] \\
&= (e^{\Delta t D_t} - I)u - \mu \Big[(e^{-\frac{1}{2}\Delta x D_x} - e^{\frac{1}{2}\Delta x D_x})a(\Delta x D_x + \tfrac{1}{6}(\Delta x D_x)^3)u \\
&+(e^{-\frac{1}{2}\Delta x D_x} + e^{\frac{1}{2}\Delta x D_x})a(\tfrac{1}{2}(\Delta x D_x)^2)u\Big] + O((\Delta x)^4) \\
&= (e^{\Delta t D_t} - I)u \\
&-\mu \left[(2\tfrac{1}{2}\Delta x D_x + 2\tfrac{1}{6}(\tfrac{1}{2}\Delta x D_x)^3 + O((\Delta x)^5))a(\Delta x D_x + \tfrac{1}{6}(\Delta x D_x)^3)u\right. \\
&\left.+ (2 + 2\tfrac{1}{2}(\tfrac{1}{2}\Delta x D_x)^2 + O((\Delta x D_x)^4))a(\tfrac{1}{2}(\Delta x D_x)^2)u\right] + O((\Delta x)^4) \\
&= \Delta t D_t u + O((\Delta t)^2) - \tfrac{\Delta t}{(\Delta x)^2}\left[(\Delta x)^2 D_x a D_x u + (\Delta x)^2 a D_x^2 u + O((\Delta x)^4)\right] \\
&= \Delta t(D_t u - D_x a D_x u - a D_x^2 u) + O(\Delta t(\Delta x)^2) = O((\Delta x)^4),
\end{aligned}
$$

where we used $O((\Delta t)^2) = O((\Delta x)^4)$.

(b) In matrix form we have $\mathbf{u}^{n+1} = A\mathbf{u}^n$. Note that the dimensions of A and $\mathbf{u}^n$, $n = 0, \ldots$, depend on Δx. The entries of A are

$$
\begin{aligned}
A_{k,k} &= 1 - \mu(a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}}), \\
A_{k,k-1} &= \mu a_{k-\frac{1}{2}}, \\
A_{k,k+1} &= \mu a_{k+\frac{1}{2}}, \\
A_{k,l} &= 0 \text{ for } |k-l| > 1.
\end{aligned}
$$

The matrix A is symmetric and therefore has real eigenvalues.

(c) By induction we have

$$\|\mathbf{u}^n\| = \|A^n \mathbf{u}^0\| \leq \|A^n\| \|\mathbf{u}^0\| \leq \|A\|^n \|\mathbf{u}^0\|.$$

Thus the numerical solution does not grow as long as $\|A\| \leq 1$. For normal matrices the Euclidean matrix norm is the spectral radius. Hence we have stability for all starting vectors $\mathbf{u}^0$, if all eigenvalues reside in the closed complex unit disc, since then $\|A\| \leq 1$.

(d) For the given A the Gerschgorin discs for each $k = 1, \ldots, M$ are

$$\{z \in \mathbb{C} : |z - 1 + \mu(a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}})| \leq \mu(a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}})\}.$$

The centre of the Gerschgorin discs lies at $1 - \mu(a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}})$ with radius

$\mu(a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}})$, since a is a positive function. All Gerschgorin discs lie within the unit disc if and only if $1 - 2\mu(a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}}) \geq -1$ for all $k = 1, \ldots, M$. Now $a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}} \leq 2a_+$. Hence we need $1 - 4\mu a_+ \geq -1$ and therefore $\mu \leq \frac{1}{2a_+}$.

8.8 EXERCISE 8.15

Consider the advection equation

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x}$$

for $x \in [0, 1]$ *and* $t \in [0, T]$.

(a) What does it mean, if a partial differential equation is well posed?

(b) Define stability for time marching algorithms for PDEs.

(c) Derive the eigenvalue analysis of stability.

(d) Define the forward difference operator Δ_+*, the central difference operator* δ *and the averaging operator* μ_0 *and calculate the operator defined by* $\delta\mu_0$.

(e) In the solution of partial differential equations often matrices occur which are constant on the diagonals. Let A *be an* $M \times M$ *matrix of the form*

$$A = \begin{pmatrix} a & b & & \\ -b & a & & \\ & \ddots & \ddots & b \\ & & -b & a \end{pmatrix},$$

that is $A_{i,i} = a, A_{i,i+1} = b, A_{i+1,i} = -b$ *and* $A_{i,j} = 0$ *otherwise. The eigenvectors of* A *are* $\mathbf{v}_1, \ldots, \mathbf{v}_M$ *where the* j*-th component of* $\mathbf{v}_k$ *is given by* $(\mathbf{v}_k)_j = \imath^j \sin \frac{\pi jk}{M+1}$*, where* $\imath = \sqrt{-1}$*. Calculate the eigenvalues of* A *by evaluating* $A\mathbf{v}_k$ *(Hint:* $\sin(x \pm y) = \sin x \cos y \pm \cos x \sin y$*).*

(f) The advection equation is approximated by the following Crank-Nicolson scheme

$$u_m^{n+1} - u_m^n = \frac{1}{4}\mu(u_{m+1}^{n+1} - u_{m-1}^{n+1}) + \frac{1}{4}\mu(u_{m+1}^n - u_{m-1}^n),$$

where $\mu = \Delta t/\Delta x$ *and* $\Delta x = 1/(M+1)$*. Assuming zero boundary conditions, that is* $u(0, t) = u(1, t) = 0$*, show that the scheme can be written in the form*

$$B\mathbf{u}^{n+1} = C\mathbf{u}^n,$$

where $\mathbf{u}^n = \begin{pmatrix} u_1^n & \ldots & u_M^n \end{pmatrix}^T$*. Specify the matrices* B *and* C.

(g) Calculate the eigenvalues of $A = B^{-1}C$ *and their moduli.*

(h) Deduce the range of μ *for which the method is stable.*

Solution

(a) A PDE problem is said to be well posed if

- a solution to the problem exists,
- the solution is unique, and
- the solution (in a compact time interval) depends in a uniformly bounded manner on the initial conditions.

(b) A numerical method for a PDE is stable if for zero boundary conditions it produces a uniformly bounded approximation of the solution in any bounded interval of the form $0 \leq t \leq T$, when $\Delta x \to 0$ and the Courant number μ is constant.

(c) If a numerical method (for a PDE with zero boundary conditions) can be written in the form

$$\mathbf{u}_{\Delta x}^{n+1} = A_{\Delta x}\mathbf{u}_{\Delta x}^{n},$$

where $\mathbf{u}_{\Delta x}^{n} \in \mathbb{R}^M$ and $A_{\Delta x}$ is an $M \times M$ matrix. By induction we have $\mathbf{u}_{\Delta x}^{n} = (A_{\Delta x})^n \mathbf{u}_{\Delta x}^{0}$. For any vector norm $\|\cdot\|$ and the induced metric norm $\|A\| = \sup \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$ we have

$$\|\mathbf{u}_{\Delta x}^{n}\| = \|(A_{\Delta x})^n \mathbf{u}_{\Delta x}^{0}\| \leq \|(A_{\Delta x})^n\| \|\mathbf{u}_{\Delta x}^{0}\| \leq \|(A_{\Delta x})\|^n \|\mathbf{u}_{\Delta x}^{0}\|.$$

Stability can now be defined as preserving the boundedness of $\mathbf{u}_{\Delta x}^{n}$ with respect to the chosen norm $\|\cdot\|$, and it follows from the inequality above that the method is stable if

$$\|A_{\Delta x}\| \leq 1.$$

Usually the norm $\|\cdot\|$ is chosen to be the Euclidean length. For normal matrices (i.e. matrices which have a complete set of orthonormal eigenvectors) the Euclidean norm of the matrix equals the spectral radius, i.e. $\|A\| = \rho(A)$, which is the maximum modulus of the eigenvalues and we arrive at the eigenvalue analysis of stability. That is the method is stable, if the maximum modulus of the eigenvalues of $A_{\Delta x}$ is less than one.

(d) The forward difference operator Δ_+ is given by $\Delta_+ f(x) = f(x+h) - f(x)$, the central difference operator δ is defined by $\delta f(x) = f(x+\frac{1}{2}h) - f(x-\frac{1}{2}h)$ and lastly the averaging operator μ_0 is defined by $\mu_0 f(x) = \frac{1}{2}(f(x+\frac{1}{2}h) + f(x-\frac{1}{2}h))$. The operator defined by $\delta\mu_0$ is $\delta\mu_0 f(x) = \frac{1}{2}(f(x+h) - f(x-h))$.

(e) Given

$$A = \begin{pmatrix} a & b & & \\ -b & a & & \\ & \ddots & \ddots & b \\ & & -b & a \end{pmatrix},$$

and $\mathbf{v}_k$ with $(\mathbf{v}_k)_j = i^j \sin\frac{\pi jk}{M+1}$, the j-th component of $A\mathbf{v}_k$ is given by

$$\begin{aligned}(A\mathbf{v}_k)_j = &\quad -bi^{j-1}\sin\tfrac{\pi(j-1)k}{M+1} + ai^j \sin\tfrac{\pi jk}{M+1} + b\sin i^{j+1}\sin\tfrac{\pi(j+1)k}{M+1}\\ = &\quad ai^j\sin\tfrac{\pi jk}{M+1} + bi^j[-i^{-1}(\sin\tfrac{\pi jk}{M+1}\cos\tfrac{\pi k}{M+1} - \cos\tfrac{\pi jk}{M+1}\sin\tfrac{\pi k}{M+1})\\ &\quad +i(\sin\tfrac{\pi jk}{M+1}\cos\tfrac{\pi k}{M+1} + \cos\tfrac{\pi jk}{M+1}\sin\tfrac{\pi k}{M+1})]\\ = &\quad (a+2ib\cos\tfrac{\pi k}{M+1})i^j\sin\tfrac{\pi jk}{M+1},\end{aligned}$$

since $-i^{-1} = i$ and where we used $\sin(x\pm y) = \sin x\cos y \pm \cos x\sin y$ with $x = \frac{\pi jk}{M+1}$ and $y = \frac{\pi k}{M+1}$. Thus the eigenvalues are $\lambda_k = a + 2ib\cos\frac{\pi k}{M+1}$, $k = 1,\ldots,n$.

(f) Rearranging the Crank-Nicolson scheme given by

$$u_m^{n+1} - u_m^n = \frac{1}{4}\mu(u_{m+1}^{n+1} - u_{m-1}^{n+1}) + \frac{1}{4}\mu(u_{m+1}^n - u_{m-1}^n),$$

leads to

$$u_m^{n+1} - \frac{1}{4}\mu(u_{m+1}^{n+1} - u_{m-1}^{n+1}) = u_m^n + \frac{1}{4}\mu(u_{m+1}^n - u_{m-1}^n).$$

Hence the scheme can be written in the form

$$B\mathbf{u}^{n+1} = C\mathbf{u}^n,$$

where

$$B = \begin{pmatrix} 1 & -\frac{1}{4}\mu & & \\ \frac{1}{4}\mu & 1 & & \\ & \ddots & \ddots & -\frac{1}{4}\mu \\ & & \frac{1}{4}\mu & 1 \end{pmatrix}, \qquad C = \begin{pmatrix} 1 & \frac{1}{4}\mu & & \\ -\frac{1}{4}\mu & 1 & & \\ & \ddots & \ddots & \frac{1}{4}\mu \\ & & -\frac{1}{4}\mu & 1 \end{pmatrix}.$$

(g) The matrices B and C are of the form given in the first part of the question and therefore have the same set of eigenvectors $\mathbf{v}_1,\ldots,\mathbf{v}_M$. These are also eigenvectors of B^{-1}, since

$$B\mathbf{v}_k = (1 + 2i(-\frac{1}{4}\mu)\cos\frac{\pi k}{n+1})\mathbf{v}_k = (1 - \frac{1}{2}i\mu\cos\frac{\pi k}{M+1})\mathbf{v}_k$$

and it follows that

$$B^{-1}\mathbf{v}_k = \frac{1}{1 - \frac{1}{2}i\mu\cos\frac{\pi k}{M+1}}\mathbf{v}_k.$$

By the same argument the eigenvalues of $A = B^{-1}C$ are

$$\lambda_k = \frac{1 + \frac{1}{2}i\mu\cos\frac{\pi k}{M+1}}{1 - \frac{1}{2}i\mu\cos\frac{\pi k}{M+1}}.$$

Since the numerator of λ_k is the complex conjugate of the denominator of λ_k it follows that $|\lambda_k| = 1$ for all μ.

(h) Therefore the method is stable for all $\mu > 0$.

8.9 EXERCISE 8.17

Assume a numerical scheme is of the form

$$\sum_{k=-r}^{s} \alpha_k u_{m+k}^{n+1} = \sum_{k=-r}^{s} \beta_k u_{m+k}^{n}, \qquad m \in \mathbb{Z}, n \in \mathbb{Z}^+,$$

where the coefficients α_k and β_k are independent of m and n.

(a) The approximations u_m^n, $m \in \mathbb{Z}$, are an infinite sequences of numbers. Define the Fourier transform $\hat{u}^n(\theta)$ of this sequence.

(b) Derive the Fourier analysis of stability. In the process give a definition of the amplification factor.

(c) Prove that the method is stable if the amplification factor is less than or equal to 1 in modulus.

(d) Find the range of parameters μ such that the method

$$(1-2\mu)u_{m-1}^{n+1} + 4\mu u_m^{n+1} + (1-2\mu)u_{m+1}^{n+1} = u_{m-1}^n + u_{m+1}^n$$

is stable, where $\mu = \Delta t/\Delta x^2 > 0$ is the Courant number. (Hint: Substitute $x = \cos\theta$ and check whether the amplification factor can become unbounded and consider the gradient of the amplification factor.)

(e) Suppose the above method is used to solve the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}.$$

Express the local error as a power of Δx.

Solution

(a) Let $\mathbf{u}^n = (u_m^n)_{m\in\mathbb{Z}}$ be a sequence of numbers. The Fourier transform of this sequence is the function

$$\hat{u}^n(\theta) = \sum_{m\in\mathbb{Z}} u_m^n e^{-im\theta}, \qquad -\pi \le \theta \le \pi.$$

(b) Multiplying

$$\sum_{k=-r}^{s} \alpha_k u_{m+k}^{n+1} = \sum_{k=-r}^{s} \beta_k u_{m+k}^{n}, \qquad m \in \mathbb{Z}, n \in \mathbb{Z}^+$$

by $e^{-im\theta}$ and summing over $m \in \mathbb{Z}$, gives for the left-hand side

$$\begin{aligned}
&\sum_{m\in\mathbb{Z}} e^{-im\theta} \sum_{k=-r}^{s} \alpha_k u_{m+k}^{n+1} = \sum_{k=-r}^{s} \alpha_k \sum_{m\in\mathbb{Z}} e^{-im\theta} u_{m+k}^{n+1} \\
&= \sum_{k=-r}^{s} \alpha_k \sum_{m\in\mathbb{Z}} e^{-i(m-k)\theta} u_m^{n+1} = \left(\sum_{k=-r}^{s} \alpha_k e^{ik\theta} \right) \sum_{m\in\mathbb{Z}} e^{-im\theta} u_m^{n+1} \\
&= \left(\sum_{k=-r}^{s} \alpha_k e^{ik\theta} \right) \hat{u}^{n+1}(\theta).
\end{aligned}$$

Similarly the right hand side is $\left(\sum_{k=-r}^{s} \beta_k e^{ik\theta}\right)$. It follows that

$$\hat{u}^{n+1}(\theta) = H(\theta)\hat{u}^n(\theta) \qquad \text{where} \qquad H(\theta) = \frac{\sum_{k=-r}^{s} \beta_k e^{ik\theta}}{\sum_{k=-r}^{s} \alpha_k e^{ik\theta}}.$$

The function H is called the amplification factor.

(c) The definition of stability says that there exists $C > 0$ such that $\|\mathbf{u}^n\| \le C$ for all $n \in \mathbb{Z}$. Since the Fourier transform is an isometry, this is equivalent to $\|\hat{u}^n\| \le C$ for all $n \in \mathbb{Z}$. Iterating we deduce

$$\hat{u}^{n+1}(\theta) = [H(\theta)]^{n+1}\hat{u}^0(\theta), \qquad \theta \in [-\pi, \pi], n \in \mathbb{Z}^+.$$

If $|H(\theta)| \le 1$ for all $\theta \in [-\pi, \pi]$, then by the above equation $|\hat{u}^n(\theta)| \le |\hat{u}^0(\theta)|$ and it follows that

$$\begin{aligned}
\|\hat{u}^n\|^2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |\hat{u}^n(\theta)|^2 d\theta \le \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(\theta)|^{2n} |\hat{u}^0(\theta)|^2 d\theta \\
&\le \frac{1}{2\pi}\int_{-\pi}^{\pi} |\hat{u}^0(\theta)|^2 d\theta = \|\hat{u}^0\|^2
\end{aligned}$$

and hence we have stability.

(d) For

$$(1-2\mu)u_{m-1}^{n+1} + 4\mu u_m^{n+1} + (1-2\mu)u_{m+1}^{n+1} = u_{m-1}^n + u_{m+1}^n$$

we have $r = s = 1$ and $\alpha_{-1} = \alpha_1 = (1-2\mu)$, $\alpha_0 = 4\mu$ and $\beta_{-1} = \beta_1 = 1$, $\beta_0 = 0$. This gives

$$\begin{aligned}
H(\theta) &= \frac{e^{-i\theta} + e^{i\theta}}{4\mu + (1-2\mu)(e^{-i\theta} + e^{i\theta})} = \frac{2\cos\theta}{4\mu + (1-2\mu)2\cos\theta} \\
&= \frac{\cos\theta}{2\mu + (1-2\mu)\cos\theta}.
\end{aligned}$$

Since $\theta \in [-\pi, pi]$, it sufficient to find the range of μ such that $y(x) \in [-1, 1]$ for $x \in [-1, 1]$, where

$$y(x) = \frac{x}{2\mu + (1-2\mu)x}.$$

The derivative is

$$y'(x) = \frac{2\mu + (1-2\mu)x - x(1-2\mu)}{(2\mu + (1-2\mu)x)^2} = \frac{2\mu}{(2\mu + (1-2\mu)x)^2} > 0.$$

Thus the function is monotonically increasing for all x. However, the denominator becomes zero for

$$x = \frac{-2\mu}{1-2\mu} = \frac{1}{1-\frac{1}{2\mu}}$$

and we have a vertical asymptote where y tends to infinity. For $\mu \in [0, 1/4]$ this asymptote lies between -1 and 0 and thus the method is not stable for this choice of μ. For $\mu > 1/4$, we have $y(1) = 1$ and $y(-1) = 1/(1-4\mu)$. Since y is monotonically increasing, we have $y(x) \in [1/(1-4\mu), 1]$ for $x \in [-1, 1]$. To have stability we need $1/(1-4\mu) \geq -1$. From $\mu > 1/4$ follows $1 - 4\mu < 0$ and hence

$$1/(1-4\mu) \geq -1 \Leftrightarrow 1 \leq -1 + 4\mu \Leftrightarrow \frac{1}{2} \leq \mu.$$

Thus the method is stable for all $\mu \geq 1/2$.

(e) To determine the local error we use the operators $D_t = \frac{\partial}{\partial t}$ and $D_x = \frac{\partial}{\partial x}$. For the diffusion equation we have $D_t = D_x^2$ and since the Courant number is constant $\Delta t = \mu \Delta x^2$. A shift can then be expressed as $e^{\Delta t D_t}$ or $e^{\pm \Delta x D_x}$. Thus applying the numerical scheme to the true solution

$$\begin{aligned}
&\left[e^{\Delta t D_t}\left[(1-2\mu)\left(e^{-\Delta x D_x} + e^{+\Delta x D_x}\right) + 4\mu\right]\right. \\
&\left. - \left(e^{-\Delta x D_x} + e^{+\Delta x D_x}\right)\right] u(x,t) \\
&= \left[\left(e^{-\Delta x D_x} + e^{+\Delta x D_x}\right)\left(e^{\Delta t D_t} - 2\mu e^{\Delta t D_t} - 1\right) + 4\mu e^{\Delta t D_t}\right] u(x,t) \\
&= \left[\left(e^{-\Delta x D_x} + e^{+\Delta x D_x}\right)\left(1 + \Delta t D_t + \tfrac{1}{2}\Delta t^2 D_t^2 + O(\Delta t^3)\right.\right. \\
&\left.\left. -2\mu e^{\Delta t D_t} - 1\right) + 4\mu e^{\Delta t D_t}\right] u(x,t) \\
&= \left[\left(2 + \Delta x^2 D_x^2 + \tfrac{1}{12}\Delta x^4 D_x^4 + O(\Delta x^6)\right)\left(\mu \Delta x^2 D_x^2 + \tfrac{1}{2}\mu^2 \Delta x^4 D_x^4\right.\right. \\
&\left.\left. + O(\Delta x^6) - 2\mu e^{\Delta t D_t}\right) + 4\mu e^{\Delta t D_t}\right] u(x,t) \\
&= \left[\left(\Delta x^2 D_x^2 + \tfrac{1}{12}\Delta x^4 D_x^4 + O(\Delta x^6)\right)\left(\mu \Delta x^2 D_x^2 + \tfrac{1}{2}\mu^2 \Delta x^4 D_x^4\right.\right. \\
&\left.\left. + O(\Delta x^6) - 2\mu e^{\Delta t D_t}\right) + 2\mu \Delta x^2 D_x^2 + \mu^2 \Delta x^4 D_x^4 + O(\Delta x^6)\right. \\
&\left. - 4\mu e^{\Delta t D_t} + 4\mu e^{\Delta t D_t}\right] u(x,t) \\
&= \left[\left(\Delta x^2 D_x^2 + \tfrac{1}{12}\Delta x^4 D_x^4 + O(\Delta x^6)\right)\left(\mu \Delta x^2 D_x^2 + \tfrac{1}{2}\mu^2 \Delta x^4 D_x^4\right.\right. \\
&\left.\left. + O(\Delta x^6) - 2\mu - 2\mu \Delta t D_t - \mu \Delta t^2 D_t^2 + O(\Delta t^3)\right) + 2\mu \Delta x^2 D_x^2\right. \\
&\left. + \mu^2 \Delta x^4 D_x^4 + O(\Delta x^6)\right] u(x,t)
\end{aligned}$$

$$
\begin{aligned}
&= \left[\left(\Delta x^2 D_x^2 + \tfrac{1}{12}\Delta x^4 D_x^4 + O(\Delta x^6)\right)\left(\mu(1-2\mu)\Delta x^2 D_x^2\right.\right.\\
&\left.+\mu^2(\tfrac{1}{2}-\mu)\Delta x^4 D_x^4 + O(\Delta x^6) - 2\mu\right) + 2\mu\Delta x^2 D_x^2\\
&\left.+\mu^2\Delta x^4 D_x^4 + O(\Delta x^6)\right] u(x,t)\\
&= \left[-2\mu\Delta x^2 D_x^2 + 2\mu\Delta x^2 + \mu(1-2\mu)\Delta x^4 D_x^4 - \tfrac{1}{12}2\mu\Delta x^4 D_x^4\right.\\
&\left.+\mu^2\Delta x^4 D_x^4 + O(\Delta x^6)\right] u(x,t)\\
&= (\tfrac{5}{6}\mu - \mu^2)\Delta x^4 D_x^4 + O(\Delta x^6)
\end{aligned}
$$

8.10 EXERCISE 8.19

We consider the diffusion equation with variable diffusion coefficient

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}(a\frac{\partial u}{\partial x}),$$

where $a(x)$, $x \in [-1,1]$ is a given differentiable function. The initial condition for $t = 0$ is given, that is $u(x,0) = u_0(x)$ and we have zero boundary conditions for $x = -1$ and $x = 1$, that is $u(-1,t) = 0$ and $u(1,t) = 0$, $t \geq 0$.

(a) Given space discretization step Δx and time discretization step Δt, the following finite difference method is used

$$u_m^{n+1} = u_m^n + \mu\left[a_{m-1/2}u_{m-1}^n - (a_{m-1/2} + a_{m+1/2})u_m^n + a_{m+1/2}u_{m+1}^n\right],$$

where $a_{m\pm1/2} = a(-1 + m\Delta x \pm \Delta x/2)$ and u_m^n approximates $u(-1 + m\Delta x, n\Delta t)$ and $\mu = \Delta t/(\Delta x)^2$ is constant. Show that the local error is at least $O(\Delta x^4)$.

(b) Derive the matrix A such that the numerical method given in (a) is written as

$$\mathbf{u}^{n+1} = A\mathbf{u}^n.$$

(c) Since the boundary conditions are zero, the solution may be expressed in terms of periodic functions. Therefore the differential equation is solved by spectral methods letting

$$u(x,t) = \sum_{n=-\infty}^{\infty} \hat{u}_n(t)e^{i\pi nx} \quad \textit{and} \quad a(x) = \sum_{n=-\infty}^{\infty} \hat{a}_n e^{i\pi nx}.$$

Calculate the first derivative of u with regards to x.

(d) Using convolution calculate the product

$$a(x)\frac{\partial u(x,t)}{\partial x}.$$

(e) By differentiating the result in (d) again with regards to x and truncating, deduce the system of ODEs for the coefficients $\hat{u}_n(t)$. Specify the matrix B such that

$$\frac{d}{dt}\hat{\mathbf{u}}(t) = B\hat{\mathbf{u}}(t)$$

(f) Let $a(x)$ be constant, that is $a(x) = \hat{a}_0$. What are the matrices A and B with this choice of $a(x)$?

(g) Let

$$a(x) = \cos \pi x = \frac{1}{2}(e^{\imath \pi x} + e^{-\imath \pi x}).$$

What are the matrices A and B with this choice of $a(x)$? (Hint: $\cos(x - \pi) = -\cos x$ and $\cos(x - y) + \cos(x + y) = 2 \cos x \cos y$.)

Solution

(a) We consider the diffusion equation with variable diffusion coefficient

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}(a\frac{\partial u}{\partial x}).$$

Let D_t and D_x denote the operators representing differentiation in the t and x direction respectively. We can rewrite the equation as

$$D_t u = D_x(aD_x u) = D_x a D_x u + aD_x^2 u.$$

Note $\Delta t = \mu \Delta x^2$. Applying the numerical method to the true solution gives

$$\begin{aligned}\left[e^{\Delta t D_t} - I\right] u = \quad & \mu \left[e^{-(\Delta x/2)D_x} a e^{-\Delta x D_x} u - (e^{-(\Delta x/2)D_x} + e^{(\Delta x/2)D_x}) a u \right. \\ & \left. + e^{(\Delta x/2)D_x} a e^{\Delta x D_x} u\right].\end{aligned}$$

The left hand side simplifies to

$$\Delta t D_t u + O(\Delta t^2) = \Delta t D_t u + O(\Delta x^4).$$

The term in the square brackets on the right hand side becomes

$$\begin{aligned}
&e^{-(\Delta x/2)D_x}a(1-\Delta xD_x+\frac{1}{2}\Delta x^2D_x^2-\frac{1}{6}\Delta x^3D_x^3)u\\
&+e^{(\Delta x/2)D_x}a(1+\Delta xD_x+\frac{1}{2}\Delta x^2D_x^2+\frac{1}{6}\Delta x^3D_x^3)u\\
&-(e^{-(\Delta x/2)D_x}+e^{(\Delta x/2)D_x})a+O(\Delta x^4)\\
&=(\Delta xD_x+\frac{1}{6}\Delta x^3D_x^3)u(e^{(\Delta x/2)D_x}-e^{-(\Delta x/2)D_x})a\\
&+\frac{1}{2}\Delta x^2D_x^2u(e^{-(\Delta x/2)D_x}+e^{(\Delta x/2)D_x})a+O(\Delta x^4)\\
&=(\Delta xD_x+\frac{1}{6}\Delta x^3D_x^3)u(\Delta xD_x+2\frac{1}{6}(\Delta x/2)^3D_x^3)a\\
&+\frac{1}{2}\Delta x^2D_x^2u(2+2\frac{1}{2}(\Delta x/2)^2D_x^2)a+O(\Delta x^4)\\
&=\Delta x^2(D_xuD_xa+aD_x^2u)+O(\Delta x^4)
\end{aligned}$$

Using $\Delta t=\mu\Delta x^2$ and the original differential equation we see that the local error is $O(\Delta x^4)$.

(b) The matrix A such that the numerical method given above is written as

$$\mathbf{u}^{n+1}=A\mathbf{u}^n$$

has entries

$$\begin{aligned}
A_{kk} &= 1-\mu(a_{k-1/2}+a_{k+1/2}),\\
A_{kk-1} &= \mu a_{k-1/2},\\
A_{kk+1} &= \mu a_{k+1/2},
\end{aligned}$$

with $A_{kj}=0$ if $|k-j|\geq 2$. Thus it is tridiagonal.

(c) Letting

$$u(x,t)=\sum_{n=-\infty}^{\infty}\hat{u}_n(t)e^{i\pi nx}\quad\text{and}\quad a(x)=\sum_{n=-\infty}^{\infty}\hat{a}_ne^{i\pi nx},$$

the first derivative of u with regards to x is

$$\frac{\partial u(x,t)}{\partial x}=\sum_{n=-\infty}^{\infty}\hat{u}_n(t)i\pi ne^{i\pi nx}.$$

(d) Using convolution we have

$$a(x)\frac{\partial u(x,t)}{\partial x}=\sum_{n=-\infty}^{\infty}\left(\sum_{m=-\infty}^{\infty}\hat{a}_{n-m}i\pi m\hat{u}_m(t)\right)e^{i\pi nx}.$$

(e) Differentiating again we deduce the following system ODEs for the coefficients $\hat{u}_n$

$$\hat{u}_n'(t) = -\pi^2 \sum_{m=-N/2+1}^{N/2} nm\hat{a}_{n-m}\hat{u}_m(t), \qquad n = -N/2+1, \ldots, N/2.$$

The matrix B has the entries

$$B_{nm} = -\pi^2 nm\hat{a}_{n-m}.$$

(f) For $a(x)$ constant, that is $a(x) = \hat{a}_0$, A has the entries

$$\begin{aligned} A_{kk} &= 1 - 2\mu\hat{a}_0, \\ A_{kk-1} &= \mu\hat{a}_0, \\ A_{kk+1} &= \mu\hat{a}_0, \end{aligned}$$

with $A_{kj} = 0$ if $|k-j| \geq 2$. It is a tridiagonal, symmetric Toeplitz (TST) matrix that is it is constant along the diagonals. The matrix B on the other hand is diagonal with diagonal entries

$$B_{nn} = -\pi^2 n^2 \hat{a}_0.$$

(g) For

$$a(x) = \cos \pi x = \frac{1}{2}(e^{\imath\pi x} + e^{-\imath\pi x}).$$

A has entries

$$\begin{aligned} A_{kk} &= 1 - \mu(\cos\pi(-1 + k\Delta x - \Delta x/2) + \cos\pi(-1 + k\Delta x + \Delta x/2)) \\ &= 1 + \mu(\cos(k-1/2)\Delta x\pi + \cos(k+1/2)\Delta x\pi) \\ &= 1 + 2\mu\cos k\Delta x\pi \cos \Delta x\pi/2, \\ A_{kk-1} &= -\mu\cos(k-1/2)\Delta x\pi, \\ A_{kk+1} &= -\mu\cos(k+1/2)\Delta x\pi, \end{aligned}$$

with $A_{kj} = 0$ if $|k-j| \geq 2$. On the other hand $\hat{a}_{-1} = \hat{a}_1 = 1/2$ while $\hat{a}_n = 0$ for all $n \neq -1, 1$. Thus B has entries

$$\begin{aligned} B_{kk} &= 0, \\ B_{kk-1} = B_{kk+1} &= -\tfrac{1}{2}\pi^2 k(k-1), \end{aligned}$$

with $B_{kj} = 0$ if $|k-j| \geq 2$. Here B is a TST matrix.