MATH 266 June 2002

NUMERICAL ANALYSIS, SOLUTION OF LINEAR EQUATIONS

TIME ALLOWED: TWO HOURS AND A HALF

Instructions to candidates

Full marks may be obtained for FIVE complete answers.

All questions will be marked but only the best **five** counted

This examination contributes 70% towards the final mark. The balance comes from coursework which consists of a set of mini projects each of which contains some theory and some computer practical work.

1. a) Describe the standard computer representation of floating point numbers in binary form. Explain the significance of the parameters E_{max} and E_{min}. Write down in binary form the largest and smallest sized numbers which can be accurately represented in single precision with 24 bits in the mantissa and $E_{max} = 128$ and $E_{min} = -125$. How would the computer represent a number too large to be represented normally? What would happen if such a number arose in some calculation?

b) Show that the hexadecimal number 1.6A09E6 is approximately equal to $\sqrt{2}$. Write this number in binary form and round it to 24 bits. Is this rounded number less than or greater than $\sqrt{2}$?

c) Two numbers a and b have associated absolute errors of $\pm\epsilon$ and $\pm\eta$ respectively. Write down the absolute and relative errors in $a + b$, $a - b$, $a * b$ and a/b.

d) The accurate values of the two roots of the quadratic equation

$$x^2 - 31.13x - 0.05$$

are 31.13160608 and −0.00160608.

Calculate the values of the two roots using **5** digit rounding arithmetic at **every** stage of the calculation. What are the percentage errors in these approximate values of the roots?

[20 marks]

2. a) Describe the Newton-Raphson method to find a solution to the equation $f(x) = 0$, explaining carefully the strengths and weaknesses of the method.

Show that if e_n, the error at the nth stage of the iteration process is small enough and $f'(x) \neq 0$ at the root, the error at the $n+1$ th stage is proportional to e_n^2.

b) Show that the equation

$$x^3 - 7x + 2 = 0$$

has a root in each of the intervals (-4,0), (0,1) and (1,3).

Starting at the point $x = 1.5$ perform two iterations of the Newton-Raphson process. Explain what is happening.

[20 marks]

3. Find a lower triangular matrix L with all its diagonal elements equal to 1 and zeros everywhere above the leading diagonal and an upper triangular matrix U with zeros everywhere below the leading diagonal such that the matrix A given below can be written as the product $A = L.U$.

$$A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 0 & 3 & 2 & 1 \end{pmatrix}.$$

Show that the matrix U can be written as the product $D.M$, where D is a diagonal matrix whose diagonal elements are 1, -3, -8/3 and 11/2 and M is the transpose of the matrix L.

Find the inverses of the matrices L, D and M [You may assume that the inverse of the matrix M is the transpose of the inverse of the matrix L.] Hence or otherwise determine the values of a, b and c, such that the inverse of the matrix A is

$$A^{-1} = \begin{pmatrix} a & b & 4/11 & -7/22 \\ b & c & 5/22 & 4/11 \\ 4/11 & 5/22 & c & b \\ -7/22 & 4/11 & b & a \end{pmatrix}.$$

Explain the significance of the condition number of a matrix. Using whichever norm you like, find the condition number for the matrix A.

[20 marks]

4. State Gerschgorin's circle theorem on the eigenvalues of a matrix.

Draw a diagram to show the Gerschgorin circles for the matrix A below. Show that the largest eigenvalue is real and find the largest and smallest value it could have.

$$A = \begin{pmatrix} 76.3 & 10.1 & 3.5 & 1.8 & -11.4 \\ 7.6 & 95.9 & -4.8 & 13.7 & 12.1 \\ -2.3 & -4.7 & 224.7 & 1.6 & 8.9 \\ 0.0 & 3.7 & -9.8 & 85.7 & -11.2 \\ 2.7 & -6.5 & -9.6 & 1.1 & 8.2 \end{pmatrix}$$

Describe the power method for obtaining the largest eigenvalue of a matrix and explain the theory of how it works. What would happen if the largest sized eigenvalue was complex?

Describe the method of inverse iteration for finding eigenvalues of a matrix. Explain how you would use it to evaluate the eigenvalue of smallest size of the matrix A.

[20 marks]

5. Describe the Jacobi method and the Gauss-Seidel method for finding a solution to the set of linear equations $A\,\mathbf{x} = \mathbf{b}$. What conditions are sufficient to ensure that these methods converge?

Show that both the Jacobi and Gauss-Seidel methods will converge for the matrix A and column vector $\mathbf{b}$, where

$$A = \begin{pmatrix} 151.1 & -6.04 & 7.5 \\ 6.7 & 101.7 & 6.7 \\ 5.9 & -3.9 & 98.9 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 321.0 \\ 107.0 \\ -97.0 \end{pmatrix}.$$

Find the matrix T and column vector $\mathbf{c}$ such that the Jacobi iteration method can be written in the form:

$$\mathbf{x}^{n+1} = T.\mathbf{x}^{n} + \mathbf{c}.$$

Find a norm for T, and hence estimate the number of iterations needed to reduce the difference between $\mathbf{x}^n$ and the solution $\mathbf{x}$ by a factor of 10^5.

[20 marks]

6. a) Describe Euler's Method for finding an approximation to the solution of the differential equation $dy/dt = f(t, y)$ with the initial condition $y(a) = \alpha$. Show that the error in the value of $y(b)$, $b > a$ is proportional to $h = (b-a)/N$, where h is the step length and N is the number of steps between a and b, provided that the function $f(t, y)$ is suitably well behaved.

Show that

$$\frac{d^2y}{dt^2} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} f.$$

Describe the Second Order Taylor Series method for finding an approximate solution to the differential equation $dy/dt = f(t, y)$. Deduce that the error in $y(b)$ is proportional to h^2, where h is the step length.

Describe the Modified Euler Method for finding an approximate solution to the differential equation $dy/dx = f(x, y)$. Show that for a step of length h the approximation agrees with that of the Second Order Taylor Method to order h^2.

b) Describe the procedure you would adopt to use a shooting method to find a solution to the boundary value problem

$$\frac{d^2y}{dx^2} = 2\frac{dy}{dx} + \frac{y}{1+x^2} + e^{-x^2}, \quad y(a) = \alpha, \quad y(b) = \beta,$$

where $a < b$.

[20 marks]

7.

a) Determine the constants α, β and t, such that the quadrature rule for the integral

$$\int_{-1}^{1} f(x)\,dx = \alpha f(-t) + \beta f(0) + \alpha f(t)$$

is exact for $f(x) = 1$, $f(x) = x$, $f(x) = x^2$, $f(x) = x^3$, $f(x) = x^4$ and $f(x) = x^5$. Find the error in using this formula to evaluate $\int_{-1}^{1} x^6 dx$.

How would you adapt this formula to evaluate an approximation to the integral

$$\int_a^b f(x)dx?$$

If this formula is adapted to evaluate the integral $\int_a^b f(x)\,dx$ using 2n strips of width $h = (b-a)/2n$, the error in the result is

$$\frac{16}{700}\frac{f^{(6)}(\xi)}{6!}h^6(b-a),$$

where $a < \xi < b$.

Find the value of h which will ensure that the error in evaluating the integral for $\ln 2$

$$\ln 2 = \int_1^2 \frac{dx}{x}$$

is not greater than 2.5E-08.

Would the above quadrature formula be suitable for the evaluation of the integral

$$I = \int_0^1 \sqrt{x} e^x \, dx?$$

Explain how you would reformulate the integral I in order to be able to use a quadrature formula such as the one above to obtain an accurate numerical approximation to the value of this integral.

[20 marks]