2SM46 June 1999

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.

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}. 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) Find the Decimal representation of the hexadecimal number 3.243F6A9 and show that it is approximately equal to π.

Write down the binary, hexadecimal and decimal representations of the number π^* which represents π stored in single precision in the computer memory. Evaluate the relative error of this stored number. Write down the next larger representable number in binary, hexadecimal and decimal form. [You may assume that in the single precision representation the mantissa has 24 bits.]

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

d) Use the standard formula for the solutions of a quadratic equation to find as accurately as you can the two roots of the equation :

$$x^2 - 21x + 0.01 = 0$$

Show that if you round to six digits in your calculations at every step, one of the solutions you obtain is accurate to only about one digit. Show how to obtain a more accurate value for the smaller of the two roots. [You may use the identity $a - b = (a^2 - b^2)/(a + b)$.]

Show that a very good approximation to the larger root can be obtaind by writing the quadratic equation in the form

$$x = 21 - \frac{0.01}{x}$$

and putting $x = 21$ on the right hand side.

[20 marks]

2. a) Describe the method of Simple Iteration for finding a solution to the equation $x = g(x)$. If α is a solution and $e_n = x_n - \alpha$ is the error of the nth iterate, find a relation between e_{n+1} and e_n which is accurate to first order in e_n when e_n is small. What does this tell us about the convergence of the process?

Show that the equation $x = 3 + \ln x$ has two solutions, one for $x < 1$ and the other for $x > 1$. Show that the method of Simple Iteration converges for one of these roots and diverges for the other. Write the equation in a different form so that Simple Iteration now converges to the root for which it diverged in the original form.

b) Show how to derive the formula

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}$$

for the Secant Rule for finding a solution to the equation $f(x) = 0$. Explain the advantages and disadvantages of this method compared to the Newton Raphson method.

[20 marks]

3. 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 Mid Point Method for finding an approximate solution to $dy/dt = f(t, y)$. Show that when this method is applied over one step, it agrees with the Second Order Taylor method to order h^2.

b) Derive approximations for d^2y/dx^2 and dy/dx in terms of y_{i+1}, y_i and y_{i-1}, the values of $y(x_{i+1})$, $y(x_i)$ and $y(x_{i-1})$, where $x_{i+1} = x_i + h$ and $x_{i-1} = x_i - h$. What are the errors in these approximations? Show how these expressions can be used to approximate the solution of the linear boundary problem

$$\frac{d^2y}{dx^2} = p(x)\frac{dy}{dx} + q(x)y + r(x), \qquad y(a) = \alpha, \qquad y(b) = \beta$$

by the solution of a set of linear equations.

[20 marks]

4. 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 $A = L.U$.

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

Show that the matrix U can be written as the product $D.M$, where D is a diagonal matrix whose diagonal elements are 2, 3/2, 4/3, 5/4 and 6/5 and M is the transpose of the matrix L.

Find the inverses of the matrices L, D and M and hence or otherwise show that the inverse of the matrix A is

$$A^{-1} = \begin{pmatrix} 5/6 & 2/3 & 1/2 & 1/3 & 1/6 \\ 2/3 & 4/3 & 1 & 2/3 & 1/3 \\ 1/2 & 1 & 3/2 & 1 & 1/2 \\ 1/3 & 2/3 & 1 & 4/3 & 2/3 \\ 1/6 & 1/3 & 1/2 & 2/3 & 5/6 \end{pmatrix}.$$

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]

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

Use Gerschgorin's theorem to show that the matrix A below has five distinct eigenvalues. Show that these eigenvalues are all real and find the intervals within which they lie.

$$A = \begin{pmatrix} 105.9 & 32.3 & -7.7 & 3.2 & 0.0 \\ 32.3 & 275.5 & -4.5 & 4.3 & 1.0 \\ -7.7 & -4.5 & 17.7 & 1.9 & 2.2 \\ -1.4 & -2.3 & 7.7 & -111.9 & 2.2 \\ 3.2 & 4.3 & 1.9 & -3.7 & -35.5 \end{pmatrix}$$

Describe the power method for obtaining the largest eigenvalue of a matrix and explain the theory of how it works. Using the smallest possible value for the largest eigenvalue and the largest possible value for the next largest eigenvalue obtained from Gerschgorin's theorem, estimate the rate of convergence of the power method.

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

[20 marks]

6. a) Determine the constants α, β and γ such that the quadrature rule for the integral

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

is exact for $f(x) = 1$, $f(x) = x$, $f(x) = x^2$ and $f(x) = x^3$.

Write down the composite form of Simpson's Rule for integrating a function $f(x)$ over an interval of $2n$ steps of length h. If the error per step is given by $h^5 f^{(4)}(\xi)/180$ where ξ is somewhere in the interval over which the integral is being taken, find the number of steps needed to evaluate the integral $\int_2^3 dx/x$ with an absolute error of less than 0.000001. What would the absolute error be if the integral was evaluated with twice the number of strips?

b) Show that the quadrature formula

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

is exact for $f(x)$ any odd power of x. Find the values of α, β and a for which the formula is exact for any polynomial of order 5 or less. What is the absolute error in using the above formula when $f(x) = x^6$?

[20 marks]