2SM46 June 1998

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 why numbers larger than a certain amount are not representable. What happens if a particular calculation would produce a number which is too large to be represented? Explain why numbers of smaller size than a certain amount are not representable accurately. What happens if a calculation produces such a number?

b) Write down the binary representation of the smallest representable single precision number greater than 1 and the largest representable single precision number smaller than 1. [You may assume that in the single precision representation the mantissa has 24 bits.] Write down the decimal equivalents of these numbers.

c) Find the representations in hexadecimal form and binary form of the decimal number x =23.3.

Write down the binary and hexadecimal representations of the number x^* which represents the decimal number 23.3 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.]

d) Show that the equations

$$9x + y = 21$$

$$4x + 0.444y = 9.332$$

have as solution $x = 2$ and $y = 3$.

Use the method of Gaussian Elimination to find the solution to these equations, rounding to six digits at each stage. Explain why the errors in this solution are so large.

[20 marks]

2. a) Show that the equation $x = e^x/4$ has one solution for $0 < x < 1$ and another for $2 < x < 3$. Show that the method of Simple Iteration converges for one of these solutions but diverges for the other. Find the value of the convergent solution accurate to four digits. The equation $x = e^x/4$ can be rewritten in the form $x = \ln(4x)$. Show that this has one solution for $0 < x < 1$ and another for $2 < x < 3$. Show that the method of Simple Iteration for this equation converges for one of the solutions and diverges for the other. Find the convergent solution accurate to four digits.

b) If the error e_n at the nth stage of the Newton-Raphson method for solving $f(x) = 0$ is small enough, show that the error at the $(n+1)$th stage is proportional to e_n^2 provided that at the root, $x = \alpha$, $f'(\alpha) \neq 0$ and $f''(\alpha) \neq 0$. Show that if $f'(\alpha) \neq 0$ but $f''(\alpha) = 0$ and $f'''(\alpha) \neq 0$, e_{n+1} is proportional to e_n^3.

[20 marks]

3. Verify that the matrices A, L and U given below satisfy $A = L.U$, provided that a, b, c, d, e and f have particular values. Find these values of a, b, c, d, e and f.

$$A = \begin{pmatrix} 1 & -1 & 4 & 6 \\ 3 & 0 & 33 & 24 \\ 5 & 1 & 67 & 67 \\ 4 & 17 & 178 & 148 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ a & 1 & 0 & 0 \\ 5 & b & 1 & 0 \\ 4 & 7 & c & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 1 & d & 4 & 6 \\ 0 & 3 & e & 6 \\ 0 & 0 & 5 & f \\ 0 & 0 & 0 & 7 \end{pmatrix}.$$

Find two matrices D and M such that D is a diagonal matrix and M has its leading diagonal elements all equal to 1 such that $U = D.M$.

Find the inverses of the matrices L and U and show that the inverse of A is given by:

$$A^{-1} = \begin{pmatrix} 459/5 & -208/105 & -782/35 & 47/7 \\ 318/5 & -166/105 & -544/35 & 33/7 \\ -49/5 & 11/35 & 82/35 & -5/7 \\ 2 & -1/7 & -3/7 & 1/7 \end{pmatrix}.$$

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. 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?

For the case when A and $\mathbf{b}$ are given by

$$A = \begin{pmatrix} 210.3 & 12.0 & 8.7 \\ 7.3 & 150.7 & 8.5 \\ 6.6 & 4.7 & 110.4 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 25.5 \\ 18.7 \\ 12.4 \end{pmatrix},$$

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 10^4.

[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 four distinct eigenvalues. Show that these eigenvalues are all real and find the intervals within which they lie.

$$A = \begin{pmatrix} 105.9 & 32.3 & -7.7 & 3.2 \\ 32.3 & 275.5 & -4.5 & 4.3 \\ -7.7 & -4.5 & 17.7 & 1.9 \\ 3.2 & 4.3 & 1.9 & -35.5 \end{pmatrix}$$

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 Gershgorin'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) The two point Lagrange Interpolation formula with remainder term for a function $f(x)$ is:

$$f(x) = \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1) + \frac{(x - x_0)(x - x_1)}{2} f''(\xi).$$

Use this formula to derive the Trapezoidal Rule for the approximation to the integral

$$\int_0^h f(x)\,dx = \frac{h}{2}(f(0) + f(h)) - \frac{h^3}{12} f''(\xi).$$

What is the highest order polynomial which can be integrated exactly using the Trapezoidal Rule?

Describe the composite Trapezoidal Rule for n strips of width h. Estimate the number of strips needed to evaluate the integral $\int_1^{1.6} dx/x$ with an absolute error of less than 0.00001. What would the absolute error be if the integral was evaluated with twice the number of strips?

b) Write down the three point Lagrange interpolation formula for a function $f(x)$ whose values at a, b and c are $f(a)$, $f(b)$ and $f(c)$ respectively. Integrate this expression from -1 to 1 to obtain the quadrature formula:

$$\int_{-1}^{1} f(x)\,dx = \alpha f(a) + \beta f(b) + \gamma f(c).$$

Find the values of α, β, γ in terms of a, b and c.

Write down the values of α, β, γ in the particular case when $a = -\sqrt{3/5}$, $b = 0$ and $c = \sqrt{3/5}$. Show that with these values of a, b and c the quadrature formula is exact for $f(x) = x^n$ for $n = 0$, $n = 1$, $n = 2$, $n = 3$, $n = 4$ and $n = 5$.

[20 marks]