This question paper consists of 7 printed pages, each of which is identified by the Code Number COMP242101.

A non-programmable calculator may be used.

A formulae sheet is provided.

School of Computing

January 2018

COMP2421

Numerical Computation

Answer *ALL FOUR* questions

Time allowed: 2 hours

 TURN OVER FOR FORMULA SHEET AND QUESTIONS

Formula Sheet

- For a normalised floating point number system with t digits in base β the machine precision (or unit roundoff) is given by: $\frac{1}{2}\beta^{1-t}$.
- The algorithm for Gaussian elimination with pivoting proceeds as follows:

```
Before eliminating entries in column j:
   find the entry in column j, below the diagonal, of maximum magnitude;
   if this entry is larger in magnitude than the diagonal entry then
      swap its row with row j.
Eliminate column j.
```

- The general form of an LU factorization for a 4×4 matrix $\mathbf{A}$ is:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix}$$

- Jacobi iteration for each row, i, of the $n \times n$ linear system $A\underline{x} = \underline{b}$ is given by:

$$x_i^{(k+1)} = x_i^{(k)} + \frac{1}{A_{ii}} \left(b_i - \sum_{j=1}^{n} A_{ij} x_j^{(k)} \right) .$$

- Gauss-Seidel iteration for each row, i, of the $n \times n$ linear system $A\underline{x} = \underline{b}$ is given by:

$$x_i^{(k+1)} = x_i^{(k)} + \frac{1}{A_{ii}} \left(b_i - \sum_{j=1}^{i-1} A_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} A_{ij} x_j^{(k)} \right) .$$

- SOR iteration updates each Gauss-Seidel iterate $x_i^{(k+1)}$ as follows:

$$x_i^{(k+1)} = (1-w)x_i^{(k)} + w x_i^{(k+1)} , \qquad \text{for some } w \in (0,2) .$$

- The derivative of the function f(t) is defined as: $f'(t) = \lim_{\mathrm{d}t \to 0} \frac{f(t+\mathrm{d}t) - f(t)}{\mathrm{d}t}$.
- Each step of Euler's method for the solution of an ordinary differential equation initial value problem

$$y'(t) = f(t,y) \quad y(t_0) = y_0 ,$$

takes the following form (where $\mathrm{d}t$ is the step size):

```
y[k+1] = y[k] + dt * f(t[k],y[k])
t[k+1] = t[k] + dt
```

- Each step of the midpoint rule for the solution of an ordinary differential equation initial value problem
$$y'(t) = f(t,y) \qquad y(t_0) = y_0 \ ,$$
takes the following form (where $\mathrm{d}t$ is the step size):

```
yhalf= y[k] + 0.5 * dt * f(t[k],y[k])
thalf= t[k] + 0.5 * dt
y[k+1] = y[k] + dt * f(thalf,yhalf)
t[k+1] = t[k] + dt
```

- Each iteration of Newton's method for solving a nonlinear equation $f(x) = 0$ is given by:
$$x_{i+1} \ = \ x_i - \frac{f(x_i)}{f'(x_i)} \ .$$

- Each iteration of the secant method for solving a nonlinear equation $f(x) = 0$ is given by:
$$x_{i+1} \ = \ x_i - f(x_i)\frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})} \ .$$

- The Newton interpolating polynomial, passing through $n+1$ points (whose t values are $t_0, t_1, \ldots t_n$), is given by:
$$\begin{aligned} p_n(t) \ &= \ c_0 \ + \ c_1(t-t_0) \ + \ c_2(t-t_0)(t-t_1) \ + \ldots \ + \ c_n(t-t_0)(t-t_1)(t-t_2)\ldots(t-t_{n-1}) \\ &= \ c_0 + (t-t_0)(c_1 + (t-t_1)(c_2 + (t-t_2)(c_3 + \ldots))) \ . \end{aligned}$$

Question 1

(a) Consider the *normalised* number system defined by $(\beta, t, L, U) = (10, 2, -3, 3)$, where numbers are represented in normalised floating point form (*i.e.* $b_1 \neq 0$)

$$fl(x) \;=\; \pm .b_1 b_2 b_3 \ldots b_{t-1} b_t \times \beta^e \qquad \text{where} \quad L \leq e \leq U\,.$$

(i) Estimate the unit roundoff, eps

(ii) Give the floating point representations $fl(x)$ and $fl(y)$ of the numbers $x = 4/9$ and $y = 20/3$ in this number system.

(iii) Write down the largest and smallest positive numbers which can be represented in this number system.

(iv) How many different numbers can be represented by this number system?

(v) Calculate the absolute and relative errors in the representation of $\pi \approx 3.14159265$ in this number system, giving your answers to 4 significant figures.

[8 marks]

(b) Consider the following three matrices:

$$A = \begin{pmatrix} 2/5 & 1/5 \\ 2/3 & 1/10 \end{pmatrix}, \quad B = \begin{pmatrix} 0.1 & 1.0 \end{pmatrix}, \quad C = \begin{pmatrix} 0.1 & 0.1 \end{pmatrix}.$$

(i) Show that in the number system defined in Part (a) above, it is possible to have

$$fl\left(A(B^{\mathrm{T}} + C^{\mathrm{T}})\right) \;\neq\; fl\left(AB^{\mathrm{T}} + AC^{\mathrm{T}}\right)\,.$$

(ii) Based on the observation in (i), state a practical rule for minimising rounding errors when summing over lists of numbers in floating point arithmetic.

[5 marks]

[question 1 total: 13 marks]

Question 2

(a) Consider the system of linear equations given by

$$\begin{aligned} x_1 &= -x_2 \ , \\ 200\,x_3 &= 200 \ , \\ x_3 &= 4 - 3\,x_4 \ , \\ 100\,x_2 + 100\,x_3 &= 100 \ . \end{aligned}$$

(i) Write the above system of equations in matrix form.
(ii) Use Gaussian Elimination with row pivoting to solve this system of equations.
(iii) Explain why row pivoting is important in each instance for the above example.

[10 marks]

(b) Consider three systems of linear equations

- $A\underline{x} = \underline{u}$, where $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$,
- $B\underline{x} = \underline{v}$, where $B = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$, and
- $B\underline{x} = \underline{w}$.

(i) State which direct method would be most efficient for solving all three problems and briefly justify your answer.
(ii) Write down a Pivot matrix P such that $PA = B$.

[2 marks]

(c) Consider the system of linear equations given by

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 4 \end{pmatrix} .$$

(i) Use two Jacobi iterations to solve this system with an initial guess of $\underline{x}^{(0)} = (1,1,1,1)^{\mathrm{T}}$.
(ii) For each iterate $k = 1$ and $k = 2$ calculate $\underline{r}^{(k)} = \underline{x}^{(k)} - \underline{x}^{(k-1)}$, and the Euclidean norm $||\underline{r}^{(k)}||$.
(iii) In indirect solvers of linear systems of equations, the quantity $||\underline{r}^{(k)}||$ is monitored. State briefly the purpose this quantity serves (i.e., how it is used in the algorithm).
(iv) Rewrite the matrix above (of Part (c)) in sparse form.
(v) State the reduction in storage and computational cost for the use of sparse representations of matrices when using indirect solvers such as Jacobi or Gauss-Seidel. Assume the matrix in question contains at most $\mathcal{O}(\alpha n)$ nonzero entries, where $\alpha \ll n$ is a constant.

[10 marks]

[question 2 total: 22 marks]

Question 3

(a) Consider the nonlinear equation: $x^3 - x^2 - 1 = 0$.

(i) Take one iteration of the secant method, starting with initial iterates of $x_0 = 1.4$ and $x_1 = 1.5$ to estimate a solution of the above equation.

(ii) Take one iteration of the Newton method to solve this problem using the initial iterate $x_0 = 1.5$. (You may use the fact that, for this problem, $f'(x) = 3x^2 - 2x$.)

(iii) State the errors in $f(x)$ of your estimated answers from (i) and (ii).

(iv) Using the example from (ii) and assuming that Newton's method has nearly converged after a single iteration (so that both $f(x)$ and $(x - x*)$ are small, where $x*$ denotes the exact solution), estimate the error in x.

(v) State briefly why specifying a tolerance in $f(x)$ in root finding algorithms does not guarantee that the solution is close to the root.

(vi) State briefly why specifying a tolerance in x in root finding algorithms can lead to numerical problems.

(vii) How many iterations of the bisection method would be needed to obtain a tolerance in x of 0.003, given an initial interval of $[1.4, 1.5]$? Explain your answer.

(viii) State one advantage of the Newton method over the bisection algorithm.

[14 marks]

(b) Suppose that the population of a town p is measured at 10 year intervals (time t) to be

$$\begin{aligned} t\ (\text{years}) &= [0, 10, 20, 30] \\ p\ (\text{thousands}) &= [34, 30, 36, 40]\ , \end{aligned}$$

(i) Use Newton interpolation to find a polynomial of degree at most three which passes through each of these data points.

(ii) State one reason why cubic spline interpolation is often used instead of Newton interpolation in visualization software.

[6 marks]

[question 3 total: 20 marks]

Question 4

(a) Consider the ordinary differential equation initial value problem given by

$$y'(t) \;=\; y^2 + t \qquad \text{with} \qquad y(0) = 1\,,$$

where $y'(t)$ denotes the time derivative of y.

(i) Approximate the solution of this problem using two steps of Euler's method using a value of $\mathrm{d}t = 0.1$.

(ii) Approximate the solution of this problem using two steps of the midpoint method using a value of $\mathrm{d}t = 0.2$.

(iii) How would you expect the error in the computed solution at time $t = 0.2$ to scale with $\mathrm{d}t$ for each of these computational algorithms (as $\mathrm{d}t$ is decreased)?

(iv) In general terms (regardless of the specific numerical integration method), how would you expect the error to behave after a long time $t \gg \mathrm{d}t$, for a 'small' choice of time step $\mathrm{d}t$ as compared to a significantly larger choice of $\mathrm{d}t$?

(v) State how you would check that the choice of $\mathrm{d}t$ is appropriate in the above example.

[16 marks]

(b) Consider the following system of differential equations:

$$\begin{pmatrix} y_1' \\ y_2' \\ y_3' \end{pmatrix} = \begin{pmatrix} -1 & 0 & 2 \\ 2 & -1 & 4 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}.$$

Use one step of the Midpoint method with $\mathrm{d}t = 0.2$ to estimate the solution of this problem at $t = 0.2$, subject to the initial condition:

$$\underline{y}(0) = (1, -1, 0.5)^{\mathrm{T}}\,.$$

[4 marks]

[question 4 total: 20 marks]

[grand total: 75 marks]