UNIVERSITY OF DUBLIN
TRINITY COLLEGE

Faculty of Engineering, Mathematics and Science

School of Computer Science & Statistics

Integrated Computer Science **Trinity Term**
Junior Sophister

Computational Mathematics

Fergal Shevlin

Friday May 3rd 2013 **Luce Upper** **09:30–11:30h**

Instructions to Candidates:

Answer **both** questions in **Part A** and **two** out of four in **Part B.** All questions carry equal marks.

Suggestion: Take 20 minutes to read all six questions. This leaves 100 minutes for answering questions worth 100 marks in total. So if a part of a question is worth five marks, spend five minutes answering it.

Materials permitted for this examination:
Log tables—available from the invigilators.
Graph paper—available from the invigilators.
Non-programmable calculator—indicate make and model.

Part A

Question 1. (i) How many non-unique, non-normalised, numbers can be represented in a floating-point system defined by parameters β, s, m, M?

[5 marks]

(ii) How many unique, normalised, numbers can be represented in a floating-point system defined by parameters above? Hint: it is proportional in some way to β^{s-1} because no number other than zero itself can start with zero.

[8 marks]

(iii) Enumerate all the non-negative, non-unique, non-normalised, numbers in the floating-point system defined by parameters $\beta = 4, s = 2, m = -1, M = 1$.

[8 marks]

(iv) Convert the numbers enumerated above into a floating-point system with $\beta = 10, s = 3, m = -1, M = 1$. Comment on their distribution and some consequences for computation.

[4 marks]

Question 2. Wave motion can be approximated as a second order differential equation $\partial^2 U(t, x)/\partial t^2 = C^2\, \partial^2 U(t, x)/\partial x^2$ where $U(t, x)$ is the height of the wave at position x at time t and C is a constant representing propagation speed.

For the purposes of simulation over the temporal interval $[T_1, T_2]$ and the spatial interval $[X_1, X_2]$, assume the following initial and boundary conditions are given: $U(T_1, x)$; $\partial U(T_1, x)/\partial t$; $U(t, X_1)$; $U(t, X_2)$.

(i) Show how a second order central difference approximation is used to find an expression for $U(t + \Delta t, x)$ which can be used in a program to simulate wave motion at a time Δt after the current time—where the current time is *not* the start time of the simulation.

[10 marks]

(ii) Show how a first order central difference approximation is used to modify the above expression when the current time is the start time of the simulation.

[10 marks]

(iii) A problem arises when $C^2 \Delta t^2/\Delta x^2 > 1$. How can this problem be avoided in a simulation program?

[5 marks]

Part B

Question 3. (i) For root-finding with the iterative method of bisection, use a relative termination criterion to derive an efficient iteration limit N. Use machine epsilon ϵ where $1 + \epsilon > 1$ and a, b as the first and last points of the initial interval.

[13 marks]

(ii) Use both the Simple Iterative method and the Newton-Raphson method to find a positive root of the function $f(x) = x^3 - 30x^2 + 2552$ in the vicinity of $x_0 = 11.87$, with a termination tolerance of 1×10^{-4}.

[12 marks]

Question 4. (i) Give a worked example of a rounding error arising in the addition of two numbers from the floating-point system defined by parameters $\beta = 10, s = 6, m = -1, M = 1$.

[6 marks]

(ii) Give a worked example of an error arising in a floating-point system such that $(a + b)/2 \neq a + (b - a)/2$.

[6 marks]

(iii) For the number $x = 3.526437$, make a table comprising of six rows of three columns: $\tilde{x}$, an approximation of x with A digits; the relative error of x with A digits; A, the number of digits, $6 \ldots 1$.

[5 marks]

(iv) Derive an expression for the relative error of subtraction of two floating point numbers. The expression should show the error term as clearly as possible.

[8 marks]

Question 5. (i) Use the composite trapeziodal rule to numerically integrate $\int_0^1 e^x dx$ with intervals $h_0 = 1, h_1 = \frac{h_0}{2}, h_2 = \frac{h_1}{2}$. Note the true solution is $e - 1$.

[10 marks]

(ii) Combine the above estimates using Richardson's deferred approach to the limit with h^2 extrapolation.

[10 marks]

(iii) Why is iterative computation so often required to approximate the solutions of mathematical problems arising in science and engineering?

[5 marks]

Question 6. (i) Show how a system of linear equations can be written in matrix form as $\mathbf{Ax} = \mathbf{b}$. Derive the expression for the Moore-Penrose pseudo-inverse of $\mathbf{A}$. In what circumstance is it appropriate to use the pseudo-inverse to solve a system of linear equations?

[9 marks]

(ii) Describe Cholesky's reduction to factorise matrix $\mathbf{A}$ into a pair of lower and upper triangular matrices $\mathbf{L}$ and $\mathbf{U}$.

[9 marks]

(iii) Show how $\mathbf{L}$ and $\mathbf{U}$ can be used to solve for $\mathbf{x}$. Explain one particular advantage of using Cholesky's reduction in a computer program.

[7 marks]