Duration: ***50 minutes***

Aids Allowed: **A calculator.**
A single 8.5 by 11 inch aid sheet, written on one side only.
The aid sheet must be a handwritten original copy. No photocopies.

Student Number: |_|_|_|_|_|_|_|_|_|_|

Family Name: ______________________________

First Name: ______________________________

---

*Do **not** turn this page until you have received the signal to start.*
Please fill out the identification section above.

This test consists of 5 questions on 6 pages (including this one). *When you receive the signal to start, please make sure that your copy of the test is complete.*

This test is double-sided.

1: ______/ 6

2: ______/ 6

3: ______/ 6

4: ______/ 6

5: ______/ 6

TOTAL: ______/30

Good Luck

Question 1. [6 MARKS]

Part (a) [2 MARKS]

In a floating-point number system, is the product of two floating-point numbers usually exactly representable in the floating-point system? Explain.

Part (b) [2 MARKS]

Give examples of floating-point arithmetic operations that would produce each of the exceptional values `Inf` and `NaN`.

Would produce `Inf`:

1/0

Would produce `NaN`:

(4-4)/(2-2)

Part (c) [2 MARKS]

In what circumstances does *cancellation* occur in a floating-point system?

Subtraction for two close numbers

Question 2. [6 MARKS]

Part (a) [4 MARKS]

If $\mathbf{A}$ is an $n \times n$ matrix, $\underline{x}$ is an $n \times 1$ (column) vector and $\underline{y}$ is a $1 \times n$ (row) vector, which of the following computations requires less work? Explain.

1. $\mathbf{B} = (\underline{x} * \underline{y}) * \mathbf{A}$
2. $\mathbf{B} = \underline{x} * (\underline{y} * \mathbf{A})$

*The second requires less work . Because if we calculate x*y which gives us an nXn matrix. And if you multiply two nXn matrix takes O(n^3) flops.*

Part (b) [2 MARKS]

Give an example of a 3×3 matrix $\mathbf{A}$, other than the identity matrix, such that $\text{cond}(\mathbf{A}) = 1$. Justify your response.

permutation of identity matrix

Question 3. [6 MARKS]

The dot product of two n-vectors $\underline{x}$ and $\underline{y}$ may be defined using the formula

$$\underline{x} \cdot \underline{y} = \sum_{i=1}^{n} x_i * \ y_i.$$

The formula suggests use of the algorithm:

```
dotProd = x(1) * y(1)
for i = 2:length(x)
    dotProd = dotProd + x(i) * y(i)
end for
```

This algorithm does not always compute an accurate result when implemented using floating-point arithmetic. Discuss the numerical difficulties that might arise and propose alternative approaches that may make the algorithm's result more accurate.

*First, compute all x(i)*y(i) then sort them in ascending order. Add the small number first then bigger number.*

Question 4. [6 marks]

Let $\mathbf{A} = \begin{bmatrix} 1 & 3 & 5 \\ 4 & 8 & 4 \\ 2 & 6 & 8 \end{bmatrix}$, $\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$, $\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 0.5 & 1 \end{bmatrix}$, $\mathbf{U} = \begin{bmatrix} 4 & 8 & 4 \\ 0 & 2 & 6 \\ 0 & 0 & 1 \end{bmatrix}$ and $\underline{b} = \begin{bmatrix} 4 \\ 4 \\ 6 \end{bmatrix}$.

Part (a) [2 marks]

Show that $\mathbf{PA} = \mathbf{LU}$.

simple mulipication

Part (b) [4 marks]

Use the $\mathbf{PA} = \mathbf{LU}$ factorization to solve the problem $\mathbf{A}\underline{x} = \underline{b}$ for $\underline{x}$. Please show all steps in your solution.

Ux = y, Ly = Pb

Question 5. [6 MARKS]

The "floor" of a real number x is an integer valued function of x, usually denoted $\lfloor x \rfloor$, and defined as

$$\lfloor x \rfloor \equiv \text{the largest integer} \leq x.$$

Suppose that you have a function with prototype `int mfloor ( float x ).` Given a normalized number x in the IEEE single-precision floating-point number system, `mfloor` returns an approximation to $\lfloor x \rfloor$.

The following program segment attempts to print out $\lfloor 2^k e \rfloor$ for k from 1 to 29.

```
float x;
int j,k;
x = (float) exp( 1.0 );
for ( k = 1; k < 30; k ++ ) {
    x = 2.0 * x;
    j = mfloor( x );
    printf("k = %2d  j = %10d \n", k, j);
}
```

Explain why the program will most likely print out a wrong value for $\lfloor 2^k e \rfloor$, for some k.

Note: The `exp` function computes the exponential of `x`, e^x, where the irrational number e is the base of natural logarithms. The `float` variables implement the IEEE single precision floating-point number system which has parameters $\beta = 2$, $p = 24$, $L = -126$, and $U = 127$. Assume that `int` variables are 32 bits in size.