Question 1. [8 marks]

Circle either "True" or "False" for each of the below statements.

1. False — We need a well-conditioned algorithm in order to obtain an accurate result.
2. False — The truncation error is larger if we use a floating-point system with a smaller precision.
3. True — In the floating-point system $F(\beta = 2, p = 23, L = -100, U = 100)$, the underflow level is less than the machine precision.
4. True — Floating-point multiplication is commutative.
5. False — If an $n \times n$ matrix A is invertible, then it can be written as a product $A = LU$ where L is lower-triangular and U is upper-triangular.
6. True — The number of solutions to the system $A\mathbf{x} = \mathbf{b}$ where A is $n \times n$ can be determined without knowing the right-hand side vector $\mathbf{b}$.
7. False — A 2×2 matrix A is ill-conditioned if its determinant is small.
8. True — The conditioning of a system $A\mathbf{x} = \mathbf{b}$ where A is an $m \times n$ matrix with $m > n$ is worse when the angle between $\mathbf{b}$ and $\text{span}(A)$ is large.

Question 2. [5 MARKS]

Answer the following questions with at most 1-2 sentences.

Part (a) [2 MARKS]

Recall that a problem is well-posed if a solution exists, the solution is unique, and the solution depends continuously on the problem data. Why is the continuity condition necessary for a numerical problem that we wish to solve using a computer?

Solution: Continuity condition is necessary because the problem is usually represented using floating-point numbers, so the problem representation will not be exact. Only if the problem is continuous will our computed solution be close to the exact solution.

Grading: One point for mentioning floating-point numbers. One point for detailed explanation about continuity.

Part (b) [2 MARKS]

Is it possible for floating-point division to overflow? Justify your answer.

Solution: Yes, by dividing a number with an exponent close to U by a number with an exponent close to L (or just something negative).

Grading: One point for "yes". One point for justification or example.

Part (c) [1 MARK]

Show that the pseudo-inverse of an invertible $n \times n$ matrix A is A^{-1}.

Solution: $A^{+} = A^{-1}$ because

$$A^{+} = (A^T A)^{-1} A^T$$
$$A^{+} A = (A^T A)^{-1} A^T A = I$$

Grading: No part marks except for very minor typos.

Question 3. [5 marks]

Consider the floating-point system $F(\beta = 5, p = 3, L = -5, U = 5)$, where chopping is used for rounding.

Part (a) [1 mark]

What is the representation of the decimal number 12.6 in this system? (What are the values of the mantissa and exponent?)

Solution: The values of the mantissa is 2.23 and the exponent is 1.

$$\begin{aligned} 12.6 &= 2 \cdot 5 + 2 \cdot 5^0 + 3 \cdot 5^{-1} \\ &= (2 + 2\frac{1}{5} + 3\frac{1}{5^2}) \times 5^1 \end{aligned}$$

Grading: Half point if only exponent the is wrong. Otherwise no part marks.

Part (b) [1 mark]

Perform floating-point addition on these two floating-point values: 4.31×5^3 and 2.44×5^2, where the mantissa here contains digits in base $\beta = 5$.

Solution:

$$\begin{array}{ll} 04.31 & \times 5^3 \\ +00.244 & \times 5^3 \\ =10.104 & \times 5^3 \\ =1.01 & \times 5^4 \end{array}$$

Grading: Part mark for minor typos only.

Part (c) [2 marks]

Show that the relative error in the above computation is below ϵ_{mach}. Show all your work.

Solution and Grading:

- one point for computing $\epsilon_{\text{mach}} = \beta^{1-3} = 0.04$
- one point for computing the relative error $\frac{1.01 - 1.0104}{1.0104} = -0.00396$

Question 4. [4 marks]

Perform one step of Gauss Elimination **with pivoting** on the matrix A'' to elimiate below the third diagonal. The first two steps of Gauss Elimination has already been done for you.

Write down the permutation matrix P_3, the elementary matrix M_3, and the new value $A''' = M_3 P_3 A''$.

$$A'' = \begin{bmatrix} 1 & 2 & 5 & 0 & 7 \\ 0 & 2 & 0 & 1 & 3 \\ 0 & 0 & 3 & 0 & 1 \\ 0 & 0 & -4 & 1 & 2 \\ 0 & 0 & 2 & 1 & 3 \end{bmatrix}$$

Solution: First, apply the permutation matrix to swap rows 3 and 4.

$$P_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \; P_3 A'' = \begin{bmatrix} 1 & 2 & 5 & 0 & 7 \\ 0 & 2 & 0 & 1 & 3 \\ 0 & 0 & -4 & 1 & 2 \\ 0 & 0 & 3 & 0 & 1 \\ 0 & 0 & 2 & 1 & 3 \end{bmatrix}$$

Now the matrix M_3 is below, and when applied the new matrix becomes:

$$M_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & 1 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 1 \end{bmatrix} \; A''' = \begin{bmatrix} 1 & 2 & 5 & 0 & 7 \\ 0 & 2 & 0 & 1 & 3 \\ 0 & 0 & -4 & 1 & 2 \\ 0 & 0 & 0 & 0.75 & 2.5 \\ 0 & 0 & 0 & 1.5 & 5 \end{bmatrix}$$

Grading:

- one point for P_3 (no part marks?)
- one point for M_3 (half point for getting the sign right)
- two points for $A^{(\prime\prime\prime)}$

Question 5. [6 marks]

Part (a) [3 marks]

Given an $n \times n$ non-singular matrix A and a second matrix B, describe an efficient algorithm to compute $A^{-1}B$.

- First, compute the LU factorization of $PA = LU$
- Then, for each column $\mathbf{b}_k$ of B, compute $x_k = A^{-1}\mathbf{b}_k$ using forward and backward substitution
- The solution matrix hs the columns x_k from the previous step

Grading:

- one point: Computes the LU factorization only once,
- one point: Specifies use of forward/backward substitution
- one point: Separating the columns of B, and treating them as problems $A\mathbf{x} = \mathbf{b}$

Part (b) [3 marks]

Perform Cholesky Factorization on this matrix.

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 5 & 1 \\ 1 & 1 & 3 \end{bmatrix}$$

Solution:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix}$$

Grading: Half point per entry.

Question 6. [4 MARKS]

Consider the function $f(\mathbf{x}) = ||A\mathbf{x}||_2$, where $\mathbf{x}$ is a $n \times 1$ vector, A is an $n \times n$ invertible matrix, and $f(\mathbf{x})$ is a real number representing the 2-norm of the vector $A\mathbf{x}$. We'll omit the subscript in $|| \cdot ||_2$ to keep the notation clean, but all norms in this question are 2-norms.

In this question, we will show that the condition number of f is $\text{cond}(A)$, so that the relative error $\left|\frac{f(\mathbf{x}+\Delta\mathbf{x})-f(\mathbf{x})}{f(\mathbf{x})}\right|$ is bounded above by $\text{cond}(A)\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||}$.

To that end, let $\mathbf{y} = A\mathbf{x}$ so that $f(\mathbf{x}) = ||\mathbf{y}||$. Suppose that there is a perturbation $\Delta\mathbf{x}$ in $\mathbf{x}$. Let $\Delta\mathbf{y} = A(\mathbf{x} + \Delta\mathbf{x}) - A\mathbf{x}$

Part (a) [1 MARK]

Show that $||\Delta\mathbf{y}|| \leq ||A||||\Delta\mathbf{x}||$, justifying your steps.

Solution: First, $\Delta\mathbf{y} = A(\mathbf{x} + \Delta\mathbf{x}) - A\mathbf{x} = A\Delta\mathbf{x}$

So, using the matrix norm property that $||A\Delta\mathbf{x}|| \leq ||A||||\Delta\mathbf{x}||$, we have $||\Delta\mathbf{y}|| = ||A\Delta\mathbf{x}|| \leq ||A||||\Delta\mathbf{x}||$

Part (b) [1 MARK]

Show that $||\mathbf{x}|| \leq ||A^{-1}||||\mathbf{y}||$, justifying your steps.

Solution: Since $y = Ax$ and A is invertible, we can write $x = A^{-1}y$. Using the matrix norm property that $||A^{-1}y|| \leq ||A^{-1}||||y||$, we have $||\mathbf{x}|| = ||A^{-1}y|| \leq ||A^{-1}||||y||$

Part (c) [2 MARKS]

Show that $\frac{||\Delta\mathbf{y}||}{||\mathbf{y}||} \leq \text{cond}(A)\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||}$, so that $\left|\frac{f(\mathbf{x}+\Delta\mathbf{x})-f(\mathbf{x})}{f(\mathbf{x})}\right| \leq \text{cond}(A)\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||}$.

Solution:

Since $||\Delta\mathbf{y}|| \leq ||A||||\Delta\mathbf{x}||$ and $||\mathbf{x}|| \leq ||A^{-1}||||\mathbf{y}||$, we have that

$$\frac{||\Delta\mathbf{y}||}{||A^{-1}||||\mathbf{y}||} \leq \frac{||A||||\Delta\mathbf{x}||}{||\mathbf{x}||}$$

$$\frac{||\Delta\mathbf{y}||}{||\mathbf{y}||} \leq ||A||||A^{-1}||\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||}$$

$$\frac{||\Delta\mathbf{y}||}{||\mathbf{y}||} \leq cond(A)\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||}$$