

Calculators are allowed; write computational results with precision of 5 significant digits, unless stated otherwise.

1. (20 points) Consider a floating-point binary number system. Apart from the sign bit and the leading normalizing bit, two bits are used to represent the fraction (mantissa) and the exponents range between -7 and +8.

(a) Complete the table:

number (dec)	sign (bin)	mantissa (bin)	exp (dec)
$1.5 = 0.75 \cdot 2^1$	0	0.(1)10	1
$-5 = 0.625 \cdot 2^3$	1	0.(1)01	3
$0.3125 = 0.625 \cdot 2^{-1}$	0	0.(1)01	-1
smallest positive number = $0.5 \cdot 2^{-7}$	0	0.(1)01	-7
largest positive number = $0.875 \cdot 2^8$	0	0.(1)11	8

(b) What is the result of addition  $1.2345 + 0.2543$  in this system? Explain.

*The numbers 1.2345 and 0.2543 are represented (approximately) by binary numbers  $1.01 = 0.101 \cdot 2^1$  and  $0.01 = 0.100 \cdot 2^{-1}$  in this system.*

*There sum  $1.01 + 0.01 = 1.10 = 0.110 \cdot 2^1$  is a number in our system, so the result of the addition is 1.5*

(c) Compute the round-off error in (b).

*The error = computed - exact, i.e:  $1.5 - 1.4888 = 0.0112$ .*

2. (20 points) Let  $f(x) = x^3 + 3x^2 - 4 = (x - 1)(x + 2)^2$

(a) Complete the table:

root	Newton iterator	$x_0$	$x_1$	$x_2$	$x_3$	conv. rate
$x = 1$	$x_{i+1} = x_i - \frac{x_i^3 + 3x_i - 4}{3x_i^2 + 3}$	1.1	1.0061	1.0000	1.0000	quadratic
$x = -2$	$x_{i+1} = x_i - \frac{x_i^3 + 3x_i - 4}{3x_i^2 + 3}$	-2.2	-2.2	-2.1030	-2.0264	linear
$x = -2$ (part (b))	$x_{i+1} = x_i - 2\frac{x_i^3 + 3x_i - 4}{3x_i^2 + 3}$	-2.2	-2.0061	-2.0000	-2.0000	quadratic

(b) What can be done to improve convergence around  $x = -2$ ? Fill in the last row of the table above.

*Assuming the multiplicity of this solution is 2 (which is the case here), multiply the correction part of the Newton iterator by 2.*

3. (20 points) Assume there are  $x_i$  trillion rabbits in the world in the beginning of year  $i$ . Each year one half of the population dies and  $1/x_i$  trillion bunnies are born.

(a) Write a formula for  $x_{i+1}$  in terms of  $x_i$ .

$$x_{i+1} = \frac{x_i}{2} + \frac{1}{x_i}$$

(b) If there are 1 trillion rabbits in year 2006, what will be the population size in 2007, 2008, 2009?

*The sequence is  $x_{2006} = 1$ ,  $x_{2007} = 1.5$ ,  $x_{2008} = 1.4166$ ,  $x_{2009} = 1.4142$ .*

(c) Does the sequence  $x_i$  converge? Explain why (why not).

*Yes. The iterator above is of form  $x = g(x)$ , where  $g(x) = x/2 + 1/x$ . Compute  $g'(x) = 1/2 - 1/x^2$ : The fixed-point iteration sequence converges since  $|g'(x)| < 1$  for an interval containing the beginning of the sequence.*

4. (20 points) Let  $0 < a < 0.1$ , consider the matrix

$$A = \begin{bmatrix} a & 0 & 4 & 1 \\ 2 & 4 & 6 & 2 \\ 1 & 2 & 4 & 0 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

(a) Perform Gaussian reduction with pivoting for better numerical stability.

*Permute rows 1 and 2, since the entry (2,1) is the largest in the first column. Reduce by the first row now.*

$$A = \begin{bmatrix} 2 & 4 & 6 & 2 \\ (a/2) & -2a & 4 - 3a & 1 - a \\ (1/2) & 0 & 1 & -1 \\ (1/2) & 0 & 0 & 3 \end{bmatrix}$$

(b) If  $PA = LU$  is an LU-decomposition of  $A$ , then

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ a/2 & 1 & 0 & 0 \\ 1/2 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1 \end{bmatrix} \quad U = \begin{bmatrix} 2 & 4 & 6 & 2 \\ 0 & -2a & 4 - 3a & 1 - a \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

(c) Compute the determinant of  $A$ .

*The following holds:  $\det(P) \det(A) = \det(L) \det(U)$ . Therefore,  $-\det(A) = 2 \cdot (-2a) \cdot 1 \cdot 3 = -12a$ , so  $\det(A) = 12a$ .*

5. (20 points) Consider the (matrix/vector) norm  $\|A\| = \|A\|_\infty$ .

(a) Let  $A = \begin{bmatrix} 1 & -2.1 \\ -2 & 4 \end{bmatrix}$ , compute its inverse  $A^{-1} = (1/(-0.2)) \begin{bmatrix} 4 & 2.1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} -20 & -10.5 \\ -10 & -5 \end{bmatrix}$

(b) Find  $\|A\|$ ,  $\|A^{-1}\|$ , and the condition number  $\kappa(A)$ .

$$\|A\| = 6, \|A^{-1}\| = 30.5, \kappa(A) = \|A\| \cdot \|A^{-1}\| = 183$$

(c) Let  $b = [-0.1, 0]^T$ , solve  $Ax = b$

$$x = [2, 1]^T$$

(d) Given an approximate solution  $\tilde{x} = [0, 0]^T$ , compute:

the residual  $r = b - A\tilde{x} = [-0.1, 0]^T$  and its norm  $\|r\| = 0.1$

the norm of the absolute error  $\|e\| = \|x - \tilde{x}\| = 2$

(e) What are the lower and the upper bounds for the relative error in terms of  $\kappa(A)$  and  $\|r\|/\|b\|$ ? Make sure that these bounds are consistent with computations in part (d).

*The bounds are:*

$$\frac{1}{\kappa(A)} \cdot \frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}$$

*They are verified in our case:*

$$\frac{1}{183} \leq 1 \leq 183$$