

## Math 417, Homework 1

1. The floating point representation of a real number is  $x = \pm(0.a_1a_2 \dots a_n)_B \cdot B^e$ , where  $a_1 \neq 0$ ,  $-M \leq e \leq M$ . Suppose that  $B = 2$ ,  $n = 6$ ,  $M = 8$ .

Find the smallest and largest positive floating point numbers that can be represented. Give the answers in decimal form.

2. Consider the system of linear equations

$$x_1 + 2x_2 + 3x_3 = 6$$

$$-x_1 + x_2 + 2x_3 = 2$$

$$4x_1 + x_3 = 5$$

(a) Write the system in the form  $(A|b)$ . Solve by reduction to triangular form and back substitution.

(b) Show that  $\det A = a_{11}^{(1)} a_{22}^{(2)} a_{33}^{(3)}$

3. (a) Given

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{pmatrix}, \quad P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Find  $\tilde{E}_1$  such that  $P_2 E_1 = \tilde{E}_1 P_2$ . (This is used in showing that if  $A$  is nonsingular, there exists a permutation matrix  $P$  such that  $PA$  has an  $LU$  factorization.)

(b) In general, let  $P_j^i$  denote the permutation matrix which interchanges row  $i$  and row  $j$ . Let  $E_k$  be an elementary row operation matrix used in Gaussian elimination. Show that if  $i > j > k$ , then  $P_j^i E_k = \tilde{E}_k P_j^i$ , where  $\tilde{E}_k$  has the same form as  $E_k$ . What is the difference between  $E_k$  and  $\tilde{E}_k$ ?

4. Given

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}.$$

(a) Find a permutation matrix  $P$  such that  $PA$  has an  $LU$  factorization. Give the  $L$ ,  $U$  factors of  $PA$ .

(b) Solve  $PAx = Pb$  by forward elimination ( $Ly = Pb$ ) and back substitution ( $Ux = y$ ).

5. Given

$$A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ 2 \\ 4 \end{pmatrix}.$$

(a) Show that  $A$  is positive definite.

(b) Find the  $LU$  factorization of  $A$ . Solve  $Ax = b$  by forward elimination ( $Ly = b$ ) and backward substitution ( $Ux = y$ ).