Scientific Computing Homework 2

Michael Nelson

Problem 1

Exercise 1. Consider the following problems

1. Compute the 64 bit floating point format of the number $x = 71.625$ (give sign, mantissa, exponent in binary notation, you can use "..." to denote large number of zeros).
2. What is the next representable number bigger than x? Derive this from your answer in part 1, and give an approximate value in decimal.
3. What is the smallest integer n that is not representable in 64 bit floating point format, so $fl(n) \neq n$. What is $fl(n)$?

Solution 1. 1. Note that

$$71.625 = 2^6 + 2^2 + 2^1 + 2^0 + 2^{-1} + 2^{-3},$$

so the binary representation of 71.625 is 1000111.101. In particular if e denotes the exponent, then $e = 1023 + 6 = 1029$. The binary representation of 1029 is 10000000101. Thus the 64 bit floating point format of 71.625 is given in the table below:

Sign	0
Exponent	10000000101
Mantisa	0001111010000000000000000000000000000000000000000000000

2. The next representable number bigger than x is obtained by adding a 1 at the very end of the Mantisa. Thus the next number is given in the table below in 64 bit floating point format:

Sign	0
Exponent	10000000101
Mantisa	0001111010000000000000000000000000000000000000000000001

In decimal format, this number is approximately

$$71.625 + 2^{-52+6} \approx 71.625000000000014210854715202003717422485351562 5$$

3. The largest positive number that may be represented using 64 bits is given in the table below:

Sign	0
Exponent	01111111111
Mantisa	1111111111111111111111111111111111111111111111111111

This number is given by

$$\left(1 + \sum_{i=1}^{52} 2^{-i}\right) 2^{1023}.$$

Problem 2

Exercise 2. Consider the Euclidean norm

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$$

computed using floating point arithmetic.

1. Explain how you would implement this sum to minimize roundoff error in the computation. Give an example where the result is more accurate than the obvious implementation.
2. Give an example where the obvious implementation can create an overflow and discuss how one could avoid this problem.

Solution 2.

Problem 3

Exercise 3. Implement Gauss Elimination with partial pivoting (for each column, swap two rows so that the largest absolute value is on the diagonal before eliminating the entries below). You can test your work on

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 2 \\ 2 & 2 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 6 \\ 1 \end{pmatrix}.$$

Submit: GaussPartialPivoting.m

Solution 3.

Problem 4

Exercise 4. Let $\mathcal{L}$ be the set of invertible $n \times n$ lower triangular matrices. Show:

1. $L_1 L_2 \in \mathcal{L}$ for $L_1, L_2 \in \mathcal{L}$.
2. $L^{-1} \in \mathcal{L}$ if $L \in \mathcal{L}$.
3. A lower triangular matrix L is invertible if and only if the diagonal has no zero entry.
4. The product $A = LDU$ is unique if it exists (L is lower triangular, U is upper triangular, D is diagonal, L and U have ones on the diagonal).

Solution 4. 1. Write $L_1 = (a_{ij})$ and $L_2 = (b_{ij})$. Since L_1 and L_2 are lower triangular, we have $a_{ij} = 0 = b_{ij}$ whenever $i < j$. Now the entry in the ith row and jth column of $L_1 L_2$ is given by

$$\sum_{k=1}^{n} a_{ik} b_{kj}. \tag{1}$$

Now asumme that $i < j$. Then if $k < j$, then $b_{kj} = 0$, and if $k \geq j$, then $i < k$ which implies $a_{ik} = 0$. In either case, we see that each term in (**??**) is zero whenever $i < j$. Thus the (i, j) entry of $L_1 L_2$ is zero whenever $i < j$, which implies $L_1 L_2$ is lower triangular. To see that $L_1 L_2$ is also invertible, note that

$$\det(L_1 L_2) = \det(L_1) \det(L_2) \neq 0.$$

3. Write $L = (a_{ij})$. We prove the contrapositive: L is not invertible if and only if all diagonal entries are zero. Equivalently, we prove $\det L = 0$ if and only if all diagonal entries are zero. From the Leibniz formula of the determinant, we have

$$\begin{aligned} \det L &= \sum_{\sigma \in S_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)} \right) \\ &= \prod_{i=1}^{n} a_{ii}, \end{aligned}$$

where we used the fact that if $\sigma \in S_n$ is not the identity permutation, then there exists an i such that $i < \sigma(i)$, which implies $a_{i\sigma(i)} = 0$ since L is lower triangular; hence $\prod_{i=1}^{n} a_{i\sigma(i)} = 0$. Now it is clear that $\det L \neq 0$ if and only if $a_{ii} \neq 0$ for all i.