Homework 02

Math 8600 Timo Heister, heister@clemson.edu

1. (a) 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).
   (b) What is the next representable number bigger than x? Derive this from your answer in part a) and give an approximate value in decimal.
   (c) What is the smallest integer n that is not representable in 64 bit floating point format, so $fl(n) \neq n$. What is $fl(n)$?

2. Consider the Euclidean norm
$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$$
computed using floating point arithmetic.
   (a) 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.
   (b) Give an example where the obvious implementation can create an overflow and discuss how one could avoid this problem.

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`

4. Let $\mathcal{L}$ be the set of invertible n by n lower triangular matrices. Show:
   (a) $L_1 \cdot L_2 \in \mathcal{L}$ for $L_1, L_2 \in \mathcal{L}$.
   (b) $L^{-1} \in \mathcal{L}$ if $L \in \mathcal{L}$.
   (c) A lower triangular matrix L is invertible if and only if the diagonal has no zero entry.
   (d) The product $A = LDU$ is unique if it exists (L is upper triangular, U is lower triangular, D is diagonal, L and U have ones on the diagonal).