Homework 1 – Math/CS 5485
Due September 19

1. (Sherman-Morrison-Woodbury formula) Let \( X, Y \in \mathbb{R}^{n \times p} \) and \( A \in \mathbb{R}^{n \times n} \). Suppose \( A \) is invertible and define
\[
W = I + Y^t A^{-1} X.
\]
   (a) Prove that \( A + XY^t \) is invertible if and only if \( W \) is invertible.
   (b) Provided that \( W \) is invertible, prove that
\[
(A + XY^t)^{-1} = A^{-1} - A^{-1} X W^{-1} Y^t A^{-1}.
\]
   (c) Suppose that a particular \( n \times n \) matrix \( \tilde{A} \) is given and any system of linear equations having \( \tilde{A} \) as a coefficient matrix can be efficiently solved. Explain how one could use this fact to solve the system
\[
Bx = b \quad \text{where} \quad B \text{ is an } (n+1) \times (n+1) \text{ matrix having } \tilde{A} \text{ as a leading principal submatrix.}
\]

2. Between an adjacent pair of nonzero IEEE single precision real numbers how many IEEE double precision numbers are there?

3. In a binary floating point system having arithmetic and rounding done in accordance with the IEEE standard with \( u \) denoting the unit rounding error, show that the computed result of
\[
fl \left( \left( \frac{1}{u} \right) + 1 \right) = \frac{1}{u} = 2^i
\]
but
\[
fl \left( \left( \frac{1}{u} \right) - 1 \right) = \frac{1}{u} - 1 = 1.11\ldots1 \times 2^{i-1} \neq \frac{1}{u}
\]

4. In a binary floating point system with an underflow threshold smaller than \(-3\) and unit rounding error \( u \), show that the computed result of
\[
fl(\left[ \frac{2}{3} - \frac{1}{2} \right] \times 3 - \frac{1}{2}) = \begin{cases} 
  u & \text{if the mantissa has an even number of bits} \\
  -u & \text{if the mantissa has an odd number of bits}
\end{cases}
\]
and that a computed value of 0 is never obtained provided that floating point arithmetic and rounding is done in accordance with the IEEE standard. (Hint: The (infinite) binary expansion for \( 2/3 \) is \( 1.01010101\ldots \times 2^{-1} \))
5. If \(fl(AB)\) denotes the floating point result of the traditional inner product oriented algorithm for calculating matrix-matrix products, prove

\[
fl(AB) = AB + E \quad \text{with} \quad |E| \leq nu\|A\|\|B\| + O(u^2)
\]

(the inequality is to be interpreted elementwise) where \(u\) denotes the unit rounding error.

6. Let \(f(x) = \sqrt{1 + x^2} - 1\).

(a) Explain the difficulty in computing \(f(x)\) for small values of \(|x|\) and show how it can be circumvented.

(b) Compute the condition number of \(f\), \(\text{cond}_f(x)\) and discuss the conditioning of \(f\) for small values of \(|x|\).

(c) Reconcile these observations.

7. The \(n^{th}\) power of a (positive) floating point number \(x\) can be computed either by

- repeated multiplication by \(x\), \(n\) times, or by
- \(x^n = e^{n \log(x)}\).

Derive bounds in each case for the relative error induced by floating point arithmetic assuming that \(\log\) and \(\exp\) are accurately evaluated (producing the nearest floating point numbers to the true result). Neglect higher order powers of \(u\). Based on these bounds, which approach is preferable from the point of view of accuracy and how does this depend on the magnitude of \(x\)?