Math 781: Homework Assignment # 2

due Thursday, 9/11/14

Reading 2.2

Problems §2.1: 1, 4, 9, 13, 14, 16; Problems Plus 1 and 2.

**Plus 1.** Given a general floating point system

\[ \sigma \cdot (0.a_1 \cdots a_n)\beta \cdot \beta^m \]

with \( L \leq m \leq U \) and \( \beta > 1 \). Find the greatest and smallest positive numbers and the unit roundoff.

**Plus 2.** (Challenge !!!) In a typical floating point number system a non-zero number \( x \) is stored in the form

\[ x = \sigma \cdot (a_1 a_2 \cdots a_t)\beta \cdot \beta^e, \]

where \( \sigma = +1 \) or \( -1 \), \( 0 \leq a_i \leq \beta - 1 \), \( t = 53 \), \( \beta = 2 \), and \(-1021 \leq e \leq 1024\).

(a) Which of the following are numbers in this typical floating point number system ? Explain.

\[ 10, \quad 1 + 2^{-53}, \quad 1 - 10^{-53}, \quad 2^{1024} \]

(b) The following short MATLAB program is run using this typical floating point arithmetic. Is it an infinite loop ? Does it generate a floating point underflow ? Approximately what is the final value of \( E \) and explain its significance. Explain.

\[
\begin{align*}
E &= 1 \\
\text{while} \ 1 + E > 1 \\
\quad E &= E/2; \\
\text{end} \\
E &= 2*E
\end{align*}
\]