Worksheet

10/30/2019

Name: ______________________________

1. Consider a floating-point system described as follows. A number x is represented approximately as $x \approx \sigma \cdot \bar{x} \cdot 2^e$ where

   - If $1 \leq E \leq 14$ then

   $$\begin{aligned} \sigma &= \begin{cases} 1 & \text{if} \quad x \geq 0, \\ -1 & \text{if} \quad x < 0, \end{cases} \\ e &= E - 7, \\ \bar{x} &= (1.a_1a_2a_3)_2 \quad \text{(rounding to truncate)} \end{aligned}$$

   - If $E = 0$ then $e = -6$ and $\bar{x} = (0.a_1a_2a_3)_2$ (rounding to truncate).
   - If $E = 15$ then the bit sequence represents $\pm\infty$ (depending on the sign σ).

   (a) Represent the number 2.8 in this format.

   (b) Let $x = -(1.001)_2 \times 2^1$ and $y = (1.010)_2 \times 2^2$. Perform the operation xy in this floating-point format.

2. Consider the function $f(x) = xe^x$.

   (a) Find the degree n Taylor polynomial of f about $x_0 = 0$. Hint: use the Taylor expansion of e^x about 0.

   (b) Suppose we want to approximate xe^x by the polynomial $p_n(x)$ found above. For what values of n can we guarantee that the error of this approximation is at most $\epsilon = 10^{-6}$ for any $1 \leq x \leq 2$?