CS 323E # Homework 2 **Due: Tue, Sept. 21, 2004**

For questions 1–6, assume that a floating point number is represented as $\pm .b_1 b_2 \cdots b_t * \beta^k$. Use the following rules:

Rules for Conversion of Real Numbers to Floating Point

1. Zero is converted to zero.
2. For a non-zero value, write it as $\pm .b_1 b_2 \cdots b_t b_{t+1} \cdots * \beta^k$, with $b_1 \neq 0$ and at least $t+1$ base-β digits expressed.
3. If using truncating conversion, drop digits $b_{t+1} \cdots$. If rounding, drop digits $b_{t+1} \cdots$ if $b_{t+1} < \beta/2$. Otherwise, if $b_{t+1} \geq \beta/2$, increase digit b_t (and if this generates a carry, keep increasing digits until the carry is resolved).
4. Check the exponent. If $k > U$, the number overflows (and has no converted value). If $k < L$, the number underflows (and most systems convert the number to zero).

Rules for Floating Point Arithmetic

5. Convert all real values in an expression to floating point.
6. In the appropriate order, do every arithmetic operation exactly and then convert the result to floating point before doing any subsequent operation. Stop if any result overflows or underflows (unless underflows are converted to zeros).

Floating Point Arithmetic Exercises

Assume that $\beta = 10, t = 3, L = -4, U = 5$ for all the following problems:

1. (5 points) Assume that the arithmetic is *truncating*.
   A. What is $fl(.00003)$?
   B. What is $fl(9.356)$?
   C. What is $fl(99996.)$?
   D. What are $fl((955. + .89) + .51)$ and $fl(955. + (.89 + .51))$?
   E. What are $fl(2.34 * (5.67 + 8.90))$ and $fl((2.34 * 5.67) + (2.34 * 8.90))$?
2. (5 points) Repeat #1, assuming that the arithmetic is *rounding*.
3. (5 points) Determine the following: a) the largest floating point number, b) the smallest positive floating point number, c) the largest floating point number smaller than one, and d) the smallest floating point number greater than one.
4. (10 points) $\frac{a+b}{2}$ is a number that lies in the interval $[a, b]$. With *truncating* arithmetic, find an example where $fl(\frac{a+b}{2})$ is strictly less than both a and b (where both a and b are floating point numbers). Find such an example for *rounding* arithmetic. Note that an alternative formula to compute $\frac{a+b}{2}$ is $a + \frac{b-a}{2}$. Using the latter formula, what results do you get with the examples you found above?