AMSC/CMSC 460 Dr. Wolfe ASSIGNMENT #1 Due June 11, 2004

1. Ex.1.35, p.47, *Numerical Computing with MATLAB* . Modify the programs by inserting a counter that will count the lines of output. Do not hand in any output. Just answer the questions.

2. Ex.1.38, p.48, *Numerical Computing with MATLAB*

3. Ex.1.39, p.48, *Numerical Computing with MATLAB*

4. Suppose a computer carries three decimal digits and rounds. If x and y are machine numbers, define the machine version of addition $x \oplus y$ to be the result of adding x and y and rounding to three digits. For example

$$49.3 \oplus 57.4 = 107.$$

Define machine multiplication $x \otimes y$ similarly. For example,

$$1.23 \otimes 4.86 = 5.98.$$

Constuct examples to show that, in general, the following statements are <u>not</u> true:
(a) $(x \oplus y) \oplus z = x \oplus (y \oplus z)$,
(b) $(x \otimes y) \otimes z = x \otimes (y \otimes z)$,
(c) $x \otimes (y \oplus z) = (x \otimes y) \oplus (x \otimes z)$.

5. For $\alpha = 0.8717$ and $\beta = 0.8719$ calculate the midpoint m of the interval $[\alpha, \beta]$ by using the formula $m = (\alpha + \beta)/2$. First use four-digit decimal chopped arithmetic, then four digit rounded arithmetic. How reasonable are the answers ? Find another formula for the midpoint and use four-digit decimal (rounded or chopped) arithmetic to calculate the midpoint of $[0.8717, 0.8719]$. Is your formula better or worse ?

6. Let all numbers given below be correctly rounded to the number of digits shown. For each calculation, determine the smallest interval in which the result, using the true instead of rounded values , must lie.
(a) $1.1062 + 0.947$ (b) $23.46 - 12.751$
(c) $(2.747) \times (6.83)$ (d) $8.473/0.064$

7. Use three-digit arithmetic with rounding to compute the following sums (sum in the given order).

$$\text{(a)} \sum_{k=1}^{6} \frac{1}{3^k} \qquad \text{(b)} \sum_{k=1}^{6} \frac{1}{3^{7-k}}$$

Also, compare the answers with the exact sum. Which is better ?

8. We wish to calculate $f(q)$ for small $q \geq 0$ where

$$f(q) = \frac{1}{q}[1 - (1 - 2q)^{-3/2}].$$

Explain why $f(q)$ cannot be evaluated accurately in finite precision arithmetic when q is small. In the explanation you should assume that $y^{-3/2}$ can be evaluated with a relative error that is bounded by a small multiple of the unit roundoff. Use the binomial series (Taylor series) to show

$$f(q) = -3 - \frac{3 \cdot 5}{2!} q - \frac{3 \cdot 5 \cdot 7}{3!} q^2 - \cdots .$$

Why is this series a better way to evaluate $f(q)$ when q is small?

9. A study of the viscous decay of a line vortex leads to an expression for the velocity

$$u_\theta = \frac{\Gamma_0}{2\pi r}\left(1 - \exp(-\frac{r^2}{4\nu t})\right)$$

at a distance r from the origin at time $t > 0$. Here Γ_0 is the initial circulation and $\nu > 0$ is the kinematic viscosity. For some purposes the behavior of the velocity at distances $r << \sqrt{4\nu t}$ is of particular interest. Why is the form given for the velocity numerically unsatisfactory for such distances? Assuming your computer can calculate $\sinh(x)$ accurately, manipulate the expression into one that can be evaluated in a more accurate way for very small r.

10. What is the nearest floating point number to 64 on a base-2 computer with 5-bit mantissas? Show work.