233 Computational Techniques

Problem Sheet for Tutorial 1

Problem 1

Accuracy of floating point operations. Assuming rounded binary arithmetic, determine *(a)* the largest $\delta > 0$ such that $fl(1 + \delta) = 1$, and *(b)* the number of significant decimal digits, for both single and double precision.

Problem 2

Error propagation in the arithmetic operations. Analyze the propagation of the relative error in each of the four arithmetic operations by comparing the relative error of the result with the sum of the relative errors of the operands, assuming that the operations themselves do not introduce additional loss of accuracy. (Treat multiplication first, it's the easiest!)

Hint: Assume that the operands, x and y, are represented in the machine as $x + \Delta x$ and $y + \Delta y$, where $|\Delta x|$ and $|\Delta y|$ are the absolute errors of x and y. For each of the arithmetic operations, try to find an upper bound first for the absolute error of the result in terms of $|\Delta x|$ and $|\Delta y|$, then a similar relation for the relative errors. Assume that $|\Delta x|/|x|$ and $|\Delta y|/|y|$ are small compared to 1 so that terms involving their product can be neglected. Is the relative error of the result always bounded by the sum of the relative errors of each operand?

If, in your analysis of addition, you allow the operands to be negative, you need not treat subtraction separately.