Numerical Analysis **S. Nanda**
Assignment 1 **Due: 17, Jan 2011 in class**

Please follow the instructions for assignments and homework as given in the course web page. You may discuss the problems and solutions with anyone but the work written up and submitted must be done on your own.

---

1. **Loss of Precision**

   (a) Complete proof of loss of precision theorem. In class we proved that the loss of significant digits in the subtraction of $x - y$ is atleast p, by showing that in normalized form $x - y$ is shifted p digits to the left, that is, p spurious zeros are added to the right which causes the loss of p significant bits. Now prove that at most q significant bits are lost in the subtraction.

   (b) **Use of Loss of Precision Theorem:**
   Assume $x > 0$. In calculating $x - sinx$ how small should x be before we choose to use Taylor Approximation? Assume that we want to lose no more than 1 significant digit.

2. **Machine Epsilon**
   Machine epsilon is the smallest number x greater than zero that is representable on your computer. Or equivalently it is the largest floating point number x such that $1 + x$ cannot be distinguished from 1. Write code to find the machine epsilon of your machine.

3. (a) **Floating point/binary representation**
   Assume 32 bit binary representation of numbers. Assume the first bit is the sign bit s, the next 8 bits are the exponent or characteristic, called c in floating point, and the last 23 bits are the mantissa or fractional part. Call the fractional part f in floating point format. Thus we get a floating point number of the form $(-1)^s 2^{c-127}(1 + f)$. Change the following binary representation to a *normalised* decimal floating point representation:

   0 01000111 10110111001000100000000

   What is the next largest decimal number that can be represented by this binary word?

   Write 0.10 in binary form using 32 bit representation described above. Any observations?

   (b) Suppose decimal machine numbers are of the form
   $\pm 0.d_1d_2d_3d_4 \times 10^n$ with $1 \leq d_1 \leq 9,\ 0 \leq d_i \leq 9,\ if\ i = 2, 3, 4$ and $|n| \leq 15$.

What is he largest value of m for which the binomial coefficient $\binom{m}{k}$ can be represented for all k by the definition without causing overflow. (Burden and Faires exercise set 1.2)

.