Due: February 11th, 2008
Problems to be turned in 5,7

1. Prove the Theorem stated in class for the fixed point iteration method.
2. Manually convert the following numbers to base 2: 5, 21, 35, 64. Check your conversion with the built-in `dec2bin` function.
3. Convert the following numbers to floating point values with eight-bit mantissas: 0.4, 0.5, 1.5
4. OCTAVE has inbuilt variables called `realmax` and `realmin` denoting the largest and the smallest numbers it can store.
   (a) Check that 10*`realmax` generates an overflow while `realmax` + 1 does not. Explain.
   (b) Using OCTAVE command window find the largest n such that $n!$ exceeds `realmax`. Briefly describe how you found it.
5. Consider $f(x) = x^3 - 7$. Find a positive root of $f(x) = 0$, with the help of a calculator. Now
   (a) Write a m-file function `bisection` that takes in as input a(left end point), b (right end point), n the number of iterations and performs the Bisection method for the above function.
   (b) Starting with the interval $[1, 2]$ perform 3 iterations. Compare your answer to the calculator answer.
6. Using (OCTAVE and) Newton's method approximate to within 10^{-4}, the value of x_0 which is the point on the graph of $y = x^2$ that is closest to $(1, 0)$.
7. Find enough terms in the Taylor Series for the function $f(x) = x(1 - \ln(x))$ at $x_0 = 2$, so that the truncation error is fourth-order $O((x - x_0)^4)$). Now once you have done that, procduce the following figure using subplot: The figure needs to arrange 4 plots in a 2 by 2 grid. In the top left, plot the function and its first-order Taylor series. In the top right, plot the function and its second order Taylor series. In the bottom left and right plots, do the same with the third and fourth-order Taylor series respectively. Plot the function as a dashed line, and the Taylor series as a solid line on the x-axis from 0 to 5.

All the programs that you will need to use below are in **errors** directory of the nmm toolbox.

1. Assume a and b are two floating point numbers such that $a < b$. Suppose we decide to halve the interval iteratively for a large number of steps. Describe what you might see in floating point arithmetic vis-a-vis what you will get in real number arithmetic ? Check this intuition for $a = 1$, $b = 2$ using `halfDiff`. Describe the algorithm and its output.
2. (**Cancellation error**) We all know that the classical formula
$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n.$$
Suppose $a_n(1 + \frac{1}{n})^n$ and $E_{a,n} = |e - a_n|$ be the absolute error. Using `eprrox` describe the behaviour of this absolute error as a function of n.
3. (**Truncation Versus Round-off**) Let $x \in \mathbb{R}$, $T_k(x) = \frac{x^k}{k!}$ be the k-th term in the series expansion of e^x and $S_k(x)$ be the partial sum up to the k-th term (i.e. $1 + \sum j = 1^k T_j$). Let $E_{a,k}(x) = |S_k(x) - e^x|$.
   (a) Using `expSeriesPlot` discuss the behaviour of the absolute error for a fixed x as a function of k.
   (b) Let $x = -10$. Eventually round off error will prevent any changes in $S_k(x)$ after some k. Can you find that k ?
4. Read `demotaylor` and decide what `demotaylor(1.6,0.8)` is doing.

Summary

At the end of this chapter you should be able to

1. List the digits used in base two arithmetic.
2. Give a simple explanation of the term "floating point number".
3. Give a definition of "roundoff error".
4. Sketch the floating point number line and label its major features. Explain the expression, "There are holes in the floating point number line".
5. Explain why integer arithmetic is "exact"? and why floating point arithmetic is not "exact".
6. Identify (at least) two important differences between symbolic and numeric computations. Give one example of cancellation error.
7. Give a simple explanation of overflow.
8. Give a simple expression that defines machine precision, ϵ_m. Name the built in variable in OCTAVE that holds the value of ϵ_m.
9. With the value of ϵ_m for computations in OCTAVE, write a simple, but carefully coded `if` statement that determines whether two scalar values are close enough to be considered equal.
10. Write the formulas for computing the relative and absolute errors if α is an exact (scalar) value, and $\hat{\alpha}$ is its floating point approximation.
11. Identify the truncation error of a Taylor series expansion. Use an infinite series to give an example of truncation error and use order notation to express it.
12. Be able to distinguish the effects of roundoff and truncation errors in a computed result,