MA/CS/EGR 537
Numerical Analysis

Problem Set 1

Due: Saturday January 18, 2017

Note: Turn in any code that you used to solve the problems. Your functions/scripts must be commented and output must be formatted nicely. Turn in your assignment in class or through Canvas. All of the following problems are from chapter one of the textbook. Each problem is worth ten points toward your final grade.

1. (Like 1.4E) Consider a miniature binary computer whose floating-point numbers consist of four binary digits(no hidden bit un-normalized) for the mantissa and three binary digits for the exponent (plus a sign bit),i.e. $t = 4, s = 3$. In this case, what is machine precision? What is the largest floating point number? Let

$$x = .625, \quad y = .875$$

Give the binary representations of these numbers in this floating point number system. Perform the following computations in this floating point number system,

$$\begin{array}{llll} (a)x - y & (b)(x-y)^2 & (c)(x-y)^4 & (d)(x-y)^6, \\ (e)x + 2y & (f)(x+2y)^2 & (g)(x+2y)^4 & (h)(x+2y)^6. \end{array}$$

To estimate the absolute and relative error, also perform each computation in double precision. Note which of the computations overflow or underflow. Are the relative errors consistent with what you would expect from what you have learned about error propogation?

2. Consider computing the following sequence in floating point arithmetic

$$s_0 = 0, \quad s_i = s_{i-1} + \frac{1}{i}.$$

In double precision, how many terms in the sequence must be computed before the sequence converges?