Math 4500/6500 Homework #1

This homework assignment covers our notes on Error and Floating Point Representations.

1. Problems

1. Write down the first two terms (and the error term) in the Taylor series expansion of
$$f(h) = \ln(3 - 2h)$$
around $h = 0$. The error term should involve a mysterious unknown constant ξ. There's no way to determine ξ from the information given here; in practice, if you were to evaluate your Taylor expansion at, say, $h =$, you would know that $\xi \in (0, 4)$ and you could get bounds on the error by bounding the derivatives of $f(h)$ on this interval.

2. Consider the function $f(x) = e^{\sin x}$. Write down the first three terms of the Taylor series for $f(x)$ expanded around $x = 0$. Now use term-by-term integration to write down the first three terms of the Taylor series for the integral $F(x) = \int_0^x f(t)\,dt$ of this function.

   Approximate the integral
$$\int_0^{\frac{1}{2}} e^{\sin x}\,dx = F(0.5)$$
by evaluating your Taylor approximation to $F(x)$ at $x = 0.5$. What is the relative error in this approximation to the value of the integral? You'll need to evaluate the integral (somehow) or determine the value with software in order to answer this; I suggest using Wolfram Alpha or *Mathematica* in order to do the integral.

3. Determine the single-precision (32 bit) floating point representation of the decimal numbers 0.5,-0.5, -0.03125. It's ok to check your answers using an online tool, but I want you to do the calculation by hand and show your work.

4. Identify the IEEE-754 floating point numbers corresponding to the bit strings below, or state that they are special cases such as ± 0, NaN or $\pm\,\mathrm{Inf}$. The vertical lines and spaces aren't bits. They are just there to keep you from going insane counting digits. Explain your work.
   (1) $0|0000\,0000|0000\,0000\,0000\,0000\,0000\,000$
   (2) $1|0000\,0000|0000\,0000\,0000\,0000\,0000\,000$
   (3) $0|1111\,1111|0000\,0000\,0000\,0000\,0000\,000$
   (4) $0|1000\,0001|0110\,0000\,0000\,0000\,0000\,000$

5. (Challenge) You work for NASA. The Mars rover *Curiosity* is transmitting scientific data from the surface when a solar flare wipes out an important part of the message. The missing part of the message identifies which data set the rover will send next. Mission control knows that the portion of the message received either comes from the radiation detector, which sends a stream of 32-bit single-precision IEEE-754 floating point numbers, or from one of the cameras, which sends arbitrary strings of bits. You are given parts of several message fragments (written in hexadecimal form):
   (1) 0x02722020
   (2) 0x1a1a1a1a
   (3) 0x7fb02020

(4) 0xcabe6f94

Research "hexadecimal form" and analyze these data fragments. Which of them are likely to have come from the radiation sensor? The $0x$ is a prefix denoting hexadecimal data, not part of the message. This is not a question with a cut-and-dried answer, so I am grading on your ability to reason about Mars, radiation and these numbers, and make a coherent argument.

6. (Challenge) How big is the hole at zero in 32-bit single-precision IEEE-754 floating point arithmetic? (That is, what is the smallest IEEE-754 single-precision number that is larger than zero?)

7. Suppose that $y(x) = \sqrt{x^4 + 4} - 2$ and that you need accurate values of $y(x)$ for x near 0.
   (1) Explain why the formula given for $y(x)$ will yield high relative error.
   (2) Write a function `y[x]` in *Mathematica* (or Wolfram Alpha) and evaluate `y[0.0001]`. This does the computation in double-precision IEEE-754 floating point arithmetic.
   (3) Get the correct answer from *Mathematica* (to 16 digits) by evaluating `N[y[1/1000],16]` and compute the relative error in your first computation.
   (4) Rewrite the formula for $y(x)$ to get a new arithmetic expression for $y(x)$ which can be evaluated with low relative error. Explain why your new arithmetic expression is better.
   (5) Evaluate $y(0.001)$ using your new formula and recompute relative error to show that the relative error in your new formula is smaller.

8. You are developing a new library for a computer system. You have developed a very accurate function for computing $\cos x$. Your boss, an MBA graduate of the Terry College of Business, has never taken Numerical Analysis and suggests that you don't have to waste time developing an accurate function for $\sin x$ since you can use the identity

$$\sin x = \pm\sqrt{1 - \cos^2 x}$$

to compute $\sin x$. Explain tactfully to your boss why this is not a good idea for certain values of x and suggest an alternate way to use the $\cos x$ function to compute $\sin x$ which avoids the problem. Your explanation should involve a specific computation of $\sin x$ using your boss's method with a high relative error, and an alternate computation of $\sin x$ using your method which has a much lower relative error.