Assigned: Wednesday, February 4, 2015
Due: Wednesday, February 11, 2015

MTH 371: Homework 2
Machine Numbers and Floating Point Arithmetic

GENERAL HOMEWORK GUIDELINES:

- On the very first page of your homework, provide your name, date, and homework number.
- Homework will be graded in part on neatness, organization, and completeness of solutions. Multiple pages MUST BE STAPLED.
- Attach all Scilab code, output, and plots to the page *immediately following* each problem.
- Clearly label all plots (title, x-axis, y-axis, legend). Use the "subplot" when needed

1. Write down the IEEE double-precision representation for the following decimal numbers.
   (a) 1.5 using round up.
   (b) 5.1 using round to nearest.
   (c) -5.1 using round towards zero.
   (d) -5.1 using round down.
   (e) 50.2 using round to nearest.

2. Answer each and provide justification.
   (a) What is the gap between 2 and the next larger double-precision number?
   (b) What is the gap between 201 and the next larger double-precision number?
   (c) How many different normalized double-precision numbers are there? Express your answer using powers of two.

3. Describe and algorithm to compare two double-precision floating-point numbers a and b to determine whether $a < b$, $a = b$, or $a > b$ by comparing each of their bits from left to right, stopping as soon as a differing bit is encountered.

4. Consider a very limited system in which numbers are only of the form $\pm 1.b_1b_2b_3 \times 2^E$ and the only exponents are $E = -1, 0, 1$. What is the machine precision ϵ for this system? Assuming subnormal numbers aren't used, what is the smallest representable positive number in this system? Largest representable positive number? Draw all the numbers in this system on a number line. Is this system closed under addition? That is, if you add two numbers in the system, do you always get another which is also?

5. In class, we demonstrated round-off error with the following sequence.

$$x_1 = \frac{1}{10}, \quad x_{k+1} = \begin{cases} 2x_k, & \text{if } x_k \in [0, \frac{1}{2}] \\ 2x_k - 1, & \text{if } x_k \in (\frac{1}{2}, 1] \end{cases}$$

   See D2L for the script demonstrated in class. Explain the behavior of this sequence as seen in Scilab. First, note that as long as the exponent in the binary representation of x_k is less than -1, the new iterate x_{k+1} is formed just by multiplying x_k by 2. How is this done in IEEE double-precision arithmetic? Are there rounding errors? Once the exponent of x_k reaches -1, the new iterate x_{k+1} is formed by multiplying x_k by 2, then subtracting 1. What does this do to the binary representation of the number after renormalization? Based on this observations, explain why starting with any $x_1 \in (0, 1]$, the computer iterates eventually reach 1 and remain there. What is the maximum number of iterations possible and why?

6. In the 7th season episode *Treehouse of Horrors VI* of *The Simpsons*, Homer has a nightmare in which the following equation flies past him:

$$1782^{12} + 1841^{12} = 1922^{12}$$

If this equation were true, this would contradict Fermat's last theorem which states for $n \geq 3$, there do not exist any natural numbers x, y and z such that $x^n + y^n = z^n$. Did Homer dream up a counterexample to Fermat's last theorem?

(a) Compute $\sqrt[12]{1782^{12} + 1841^{12}}$ by typing the following command into Scilab.

```
format('v',10); (1782^12+1841^12)^(1/12)
```

What does Scilab report? Try again by displaying 20 digits instead of 10.

(b) Determine the absolute and relative errors in the approximation $1782^{12} + 1841^{12} \approx 1922^{12}$. Such an example is called a Fermat near miss.

(c) Note that the right-hand side of this equation is even. Use this to prove that the equation cannot hold.

In a later episode *The Wizard of Evergreen Terrace*, Homer writes the equation

$$3987^{12} + 4365^{12} = 4472^{12}.$$

Can you debunk this equation?

7. OPTIONAL challenge problem. In class we plotted the following polynomial in Scilab only to see strange behavior.

$$p(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1$$

Note that $p(x) = (x-1)^7$. Plot this factored version of p. Explain why Scilab has no issue with this factored form, but large error is seen with the original.