Homework 1 Solutions

1. Decimal to Binary conversions:

   (a) First, note that $9 = 2^3 + 1$, so converting 9 to binary gives 1001. Secondly, $\frac{1}{2} = 0.1$, so putting these together,

   $$(9.5)_{10} = (1001.1)_2$$

   (b) First, we see that $\frac{44}{7} = 6\frac{2}{7}$, so again we'll deal with the integer part first. We could use the algorithm in class (or simply solve the base 2 conversion by inspection, since $6 = 2^2 + 2^1$). For practice, here's the algorithm:

   $$\begin{array}{rcl} 6 \div 2 & = & 3 \quad R \quad 0 \\ 3 \div 2 & = & 1 \quad R \quad 1 \\ 1 \div 2 & = & 0 \quad R \quad 1 \end{array} \Rightarrow \quad (6)_{10} = (110)_2$$

   Next, the fractional part:

   $$\begin{array}{rcccl} \frac{2}{7} \times 2 & = & \frac{4}{7} & + & 0 \\ \frac{4}{7} \times 2 & = & \frac{1}{7} & + & 1 \\ \frac{1}{7} \times 2 & = & \frac{2}{7} & + & 0 \end{array}$$

   which we see starts repeating, so

   $$\left(\frac{2}{7}\right)_{10} = (0.\overline{010})_2$$

   Put it all together: $(44/7)_{10} = (110.\overline{010})_2$

2. Binary to Decimal Conversions:

   (a) $(1101.0111)_2 = 2^3 + 2^2 + 2^0 + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = 13\frac{7}{16}$

   (b) For the repeating pattern use our trick of multiplying by a power of two. In this case, let $x = 0.0110101010 10....$ To get two numbers with the same tail, notice that:

   $$4x = 1.1010101010\ldots \qquad 16x = 110.1010101010\ldots$$

   so that:

   $$16x - 4x = (110)_2 - (1)_2 \quad \Rightarrow \quad 12x = 5 \quad \Rightarrow \quad x = \frac{5}{12}$$

   We can check our work:

   $$\begin{array}{rcccl} \frac{5}{12} \times 2 & = & \frac{10}{12} & + & 0 \\ \frac{5}{6} \times 2 & = & \frac{4}{6} & + & 1 \\ \frac{2}{3} \times 2 & = & \frac{1}{3} & + & 1 \\ \frac{1}{3} \times 2 & = & \frac{2}{3} & + & 0 \end{array}$$

   At which point the pattern begins to repeat.

3. Convert 9.4 to floating point binary, and find the error in the approximation.

   From class, we converted 0.4, and we already saw the conversion for 9. Putting these together,

$$(9.4)_{10} = (1001.\overline{0110})_2$$

   so the floating point form is:

$$1.00101100110\ldots01100110\underbrace{1}_{52dbit}\times 2^3$$

   with a "tail" of .11001100... $\times 2^{-52} \times 2^3$ which we can write as $.\overline{0110} \times 2^{-51} \times 2^3$. Writing it like this, we see that this is 0.4×2^{-48}.

   Furthermore, we had to round, so we added $2^{-52} \times 2^3 = 2^{-49}$.

   Putting it all together,

$$fl(9.4) = 9.4 + \text{round} - \text{tail}$$

   or

$$fl(9.4) = 9.4 + 2^{-49} - 0.4 \times 2^{-48} = 9.4 + 2^{-49}(1 - 0.8) = 9.4 + 0.2 \times 2^{-49}$$

   Or, putting this in terms of $\epsilon_{\text{machine}}$,

$$fl(9.4) = 9.4 + 1.6\epsilon_{\text{machine}}$$

4. To answer this question, note our representations for $fl(9.4)$ and $fl(0.4)$ from class. There, we saw that the error in the approximation was $0.1\epsilon_{\text{machine}}$.

   In the algorithm given, we take:

$$fl(9.4) - 9 = 0.4 + 1.6\epsilon$$

   By subtracting $fl(0.4)$, we then get:

$$(0.4 + 1.6\epsilon) - (0.4 + 0.1\epsilon) = 1.5\epsilon$$

   which is the same as: 3×2^{-53}.

   In Matlab:

```
>> format long
>> x=9.4;
>> y=x-9;
>> z=y-0.4
z =
    3.330669073875470e-016

>> 3*2^(-53)
ans =
    3.330669073875470e-016
```

5. We'll denote the fourth degree Taylor polynomial by $T_4(x)$. In this case, using x^{-2} at $a = 1$ we get:

$$T_4(x) = 1 - 2(x-1) + 3(x-1)^2 - 4(x-1)^3 + 5(x-1)^4$$

with a remainder term:

$$R = \frac{-720}{5!}c^{-7}(x-1)^5 = -6c^{-7}(x-1)^5$$

where c is in either $[0.9, 1]$ or $[1, 1.1]$.

If $x = 0.9$, then
$$|(0.9)^{-2} - T_4(0.9)| \approx 6.79 \times 10^{-5}$$
The remainder term has the form: $6 \times 10^{-5} \cdot \frac{1}{c^7}$. Since c^{-7} is strictly decreasing, its maximum value occurs at $c = 0.9$. Therefore, we could make the error bound:

$$\max |R| = 6 \times 10^{-5} \times (0.9)^{-7} \approx 6 \times 10^{-5} \times 2.09$$

If $x = 1.1$, then
$$|(1.1)^{-2} - T_4(1.1)| \approx 5.37 \times 10^{-5}$$
and the remainder term has the form $-6 \times 10^{-5} \cdot \frac{1}{c^7}$. Taking the maximum error (in absolute value), the max should occur where $c = 1$;

$$\max |R| = 6 \times 10^{-5}$$