Solutions to Homework 7, Math 575A Fall 2008

13.1 \(2^{29} - 1\).\(2^{29} = 2^{52}/2^{23} - 1\). The numerator is the number of doubles greater than or equal to 1, and less than 2, while the denominator is the number of singles in that range. The -1 is to remove the single in the range.

13.2 a) \(\beta^i + 1\)

b) \(2^{53} + 1\) for double and \(2^{24} + 1\) for single

c) There are lots of ways to pursue this. One way is to try to calculate \(2^{53} + 1 + k\) for \(k = -3 : 3\) and compare the results by checking equality. Another is to write out the results as binary ieee doubles, and then read them back in as an array of integers or bits, and compare the actual bit pattern. The code fragment below does the latter.

```matlab
>> fid=fopen('bindata','wb');
>> fwrite(fid,2^53+[-8:8],'double')
>> fclose(fid);
>> fid=fopen('bindata','rb');
>> [a,n]=fread(fid,17*8,'uint8');
>> fclose(fid);
```

```
ans =

Columns 1 through 12

248 249 250 251 252 253 254 255 0 0 1 2
255 255 255 255 255 255 255 255 0 0 0 0
255 255 255 255 255 255 255 255 0 0 0 0
255 255 255 255 255 255 255 255 0 0 0 0
255 255 255 255 255 255 255 255 0 0 0 0
63 63 63 63 63 63 63 63 64 64 64 64
67 67 67 67 67 67 67 67 67 67 67 67

Columns 13 through 17

2 2 3 4 4
0 0 0 0 0
0 0 0 0 0
0 0 0 0 0
0 0 0 0 0
64 64 64 64 64
67 67 67 67 67
```
The “numbers” are represented by the columns, and as we see, below $2^{53}$, they are all different, above that however, we get repetitions.

13.3 The figures show that the expanded form of the polynomial leads to lower accuracy near $x = 2$. There are large factors and lots of near cancelation.

14.1 a) True, $|\sin(x)| \leq 1$ for all $x$

b) True, as above, but, not sharp, as $|\sin(x)| \leq x$ for all $x$, and so in particular, as $x \to 0$, $\sin(x) = O(x)$

c) True, but not sharp. It is easiest to see this after changing variables, letting $y = x^{1/100}$. so that $\log(x) = \log(y^{100}) = 100 \log(y)$. for $y \geq e^7$, $|100 \log(y)| \leq y$. (check this at $y = e^7$, and then compare the derivatives $100/y$ and 1 for larger $y$’s. Thus for $x \geq e^{700}$, $\log(x) \geq x^{1/100}$.

d) False, Stirling’s formula gives an asymptotic formula for $n!$ saying that

$$\lim_{n \to \infty} \frac{n!}{\sqrt{2\pi n n^n e^{-n}}} = 1,$$

so $\lim_{n \to \infty} \frac{n!}{n^n e^{-n}} = \lim_{n \to \infty} \frac{\sqrt{2\pi n n^n e^{-n}}}{n^n e^{-n}} = \lim_{n \to \infty} \sqrt{2\pi n} = \infty$.

We do not need to use Stirling’s formula as a “black box”. We can derive the part of it we need (with an explicit constant) if we use some facts about the trapezoid and midpoint rules for approximating integrals:
We take the log of \( n! \), and then bound the sum of the concave up function \( \log k \) using an integral that we can evaluate. The integral \( \int_1^n \log x \, dx \) is bounded between the estimate from the trapezoid rule and that for the midpoint rule. Since the function is concave down, the midpoint estimate gives an upper bound for the integral (leading to a lower bound on the sum). Thus we have

\[
\log n! = \sum_{k=1}^{n} \log k \\
\geq \int_{1/2}^{n+1/2} \log x \, dx \\
= \left( n + \frac{1}{2} \right) \log \left( n + \frac{1}{2} \right) - \left( n + \frac{1}{2} \right) - \left( \frac{1}{2} \log \left( \frac{1}{2} \right) - \frac{1}{2} \right) \\
= \left( n + \frac{1}{2} \right) \log(n) - \left( n + \frac{1}{2} \right).
\]

Exponentiating this inequality we get

\[
n! \geq \sqrt{n/e} \left( \frac{n}{e} \right)^n.
\]

This is enough to prove the part of the Stirling formula that we need.

e) True, the units are irrelevant, since the constant is arbitrary.

f) True. This is a consequence of (13.5), where the constant is \( \pi \).

g) False. (13.5) tells us that \( \text{fl}(n\pi) - n\pi = \epsilon n\pi \) for some \( \epsilon \) with \( |\epsilon| \leq \epsilon_m \). That means that the constant should be proportional to \( n \). It is not uniform. This is not a proof that it cannot be bounded independent of \( n \), it is just an indication of why we do not expect it to be true.

To actually prove this, you can use that the set of representable floating point numbers is self-similar with scalings in powers of 2, and is a subset of the rational numbers. Since \( \pi \) is not rational, \( \text{fl}(\pi - p_i) \neq 0 \). Call that number \( \delta \). Then the self-similarity guarantees that \( |\text{fl}(2^k \pi) - 2^k \pi| = 2^k |\text{fl}(\pi) - p_i| = 2^k \delta \). If we had that \( \text{fl}(n\pi) - n\pi = O(\epsilon_m) \), then there would be some number \( C \) such that \( |\text{fl}(n\pi) - n\pi| \leq C \epsilon_m \). If we pick \( k \) sufficiently large \( (k > \log_2(C\epsilon_m/\delta)) \), we will have that \( |\text{fl}(2^k \pi) - 2^k \pi| = 2^k \delta > C \epsilon_m \). This is a contradiction. Thus \( \text{fl}(n\pi) - n\pi \neq O(\epsilon_m) \).

14.2 a) Let \( f_j = 1 + O(\epsilon_m) \), for \( j = 1, 2 \). Then there are \( C_j \) with \( |f_j - 1| \leq C_j \epsilon_m \) for sufficiently small values of \( \epsilon_m \). \( f_1 f_2 - 1 = (f_1 - 1) + (f_2 - 1) + (f_1 - 1)(f_2 - 1) \), so \( |f_1 f_2 - 1| \leq (C_1 + C_2) \epsilon_m + C_1 C_2 \epsilon_m^2 \). We can choose \( C_3 = C_1 C_2 + C_1 C_2 \epsilon_m \), and if we require \( \epsilon_m < 1 \), then we have \( |f_1 f_2 - 1| \leq C_3 \epsilon_m \) for all sufficiently small \( \epsilon_m \). \( f_1 f_2 = 1 + O(\epsilon_m) \).

b) Let \( f = 1 + O(\epsilon_m) \), then there is a \( C \) so that for sufficiently small \( \epsilon \), \( |f - 1| \leq C \epsilon_m \). Require that \( \epsilon < 1/C \), so that \( |f - 1| < 1/2 \) then \( 1/f = 1 - (f - 1) + (f - 1)^2 - (f - 1)^3 + \ldots \),
the convergent infinite geometric series gives the inverse of \( f \). So \( 1/f = 1 - (f - 1) + (f - 1)^2 - (f - 1)^3 + \ldots \), then \( 1/f - 1 = -(f - 1)(1 - (f - 1) + (f - 1)^2 - (f - 1)^3 + \ldots) = -(f - 1)(1/f) \). Since \( |f - 1| < 1/2 \), we know that \( f > 1/2 \), so that \( 1/f < 2 \). Thus we have that \( |1/f - 1| < |f - 1|/2 \leq 2C\epsilon_m \), i.e. \( 1/f = 1 + O(\epsilon_m) \).