PROBLEMS

1. A real number x is represented approximately by 0.6032, and we are told that the relative error is 0.1 %. What is x ? Note: There are two answers.
   Hint : Recall that % relative error is $100 \times |x - \tilde{x}| / |x|$

2. What is the relative error involved in rounding 4.9997 to 5.000 ?
   Hint : Rounding approximates 4.9997 to 5.000, thus $x, \tilde{x}$ in $|x - \tilde{x}| / |x|$ follows.

3. How many decimals must be taken along π and e in order to compute (a) $1.3134 \cdot \pi$ and (b) $0.3761 \cdot e$ to three correct decimals?
   Hint : One way is to generate a sequence $x_1 = \alpha \times i.d_1$, $x_2 = \alpha \times i.d_1 d_2$, ... with increasing number of correct decimals d_1, d_2, ... in π (or e) where α if the fixed number and i is the integer part, until the new term of the converging sequence agrees with the previous term to three correct decimals.

4. Suppose that p^* approximates p to three significant digits. Find the largest interval in which p^* can lie if p is (a) 900, (b) 90, (c) 2.7182, (d) 151.63.
   Hint : x is said to approximate $\tilde{x}$ to at least t-digits when $|x - \tilde{x}| / |x| < \frac{1}{2} \times 10^{1-t}$. Thus the interval follows.

5. Perform the following computations (i) exactly, (ii) using three-digit chopping arithmetic, (iii) using three-digit rounding arithmetic. Then determine any loss in significant digits, assuming that the given numbers are exact. (a) $14.1 + 0.0981$, (b) 0.0218×179, (c) $(164. + 0.913) - (143. + 21.0)$, (d) $(164. - 143.) + (0.913 - 21.0)$.
   Hint : Recall that loss of significant digits occurs when two nearly equal numbers are subtracted. Further, relative error between two numbers when compared againts $\frac{1}{2} \times 10^{1-t}$ is the way to determine the number of digits agreement.

6. Consider the following values of p and p^*. What is (i) the absolute error, (ii) the relative error in approximating p by p^*, and to how many (iii) decimal digits, (iv) significant digits does p^* approximate p? (a) $p = \pi$, $p^* = 3.1$, (b) $p = 1/3$, $p^* = 0.333$, (c) $p = \pi/1000$, $p^* = 0.0031$, (d) $p = 100/3$, $p^* = 33.3$.
   Hint : Recall that relative (absolute) error between two numbers when compared againts $\frac{1}{2} \times 10^{1-t}$ ($\frac{1}{2} \times 10^{-t}$) yields the number of digits (decimals) agreement.

7. Count the number of multiplications and additions involved in evaIuating a polynomial using nested multiplication. Compare with the work needed when the powers of x are calculated by $x^i = x\, x^{i-1}$ and subsequently multiplied by a_i. Note : In order to efficiently

evaluate a polynomial $p(x) = a_0 + a_1x + a_2x^2 + ... + a_{n-1}x^{n-1} + a_nx^n$, we group the terms in a nested multiplication

$$p(x) = a_0 + x(a_1 + x(a_2 + ... + x(a_{n-1} + x(a_n))....)).$$

This is illustrated by the following MATLAB code, where a polynomial $p(x) = \sum_{i=0}^{N-1} a_i x^i$ of degree $N = 10^6$ for a certain value of x is computed by using the two methods:

```
N = 1000000+1; a = [1:N]; x = 1;
tic % initialize the timer
    p = sum(a.*x.^[N-1:-1:0]);
p, toc % measure the time
tic, pn=a(1);
    for i = 2:N
        pn = pn*x + a(i); % nested multiplication
    end
pn, toc
```

8. (a) Evaluate the polynomial $p(x) = x^3 - 5x^2 + 6x + 0.55$ at $x = 2.73$. Use 3-digit arithmetic with chopping. Evalute the relative error. (b) Repeat (a) but express $p(x) = ((x-5)x+6)x + 0.55$. Evaluate the percent relative error and compare with part (a).

9. Plot a seventh-degree polynomial by typing the following statements into the MATLAB command window:

```
X=0.988:.0001:1.012;
y = x.^7-7*x.^6+21*x.^5-35*x.^4+35*x.^3-21*x.^2+7*x-1;
plot(x,y)
```

The resulting plot does not look anything like a polynomial. It is not smooth. You are seeing roundoff error in action. The y-axis scale factor is tiny, 10^{-14}. The tiny values of y are being computed by taking sums and differences of numbers as large as $35 \cdot 1.012^4$. There is severe subtractive cancellation. Note that y is the expanded form of $y = (x-1)^7$ and the range for the x-axis is carefully chosen to be near $x = 1$. If the values of y are computed instead by

```
y = (x-1).^7;
```

then a smooth (but very flat) plot results.

10. (a) Show that the polynomial nesting technique can also be applied to the evaluation of $f(x) = 1.01e^{4x} - 4.62e^{3x} - 3.11e^{2x} + 12.2e^x - 1.99$. (b) Use 3-digit rounding arithmetic, the assumption that $e^{1.53} = 4.62$, and the fact that $e^{nx} = (e^x)^n$ to evaluate $f(1.53)$ as given in part (a). (c) Redo the calculation in part (b) by first nesting the calculations. (d) Compare the approximations in parts (b) and (c) to the true three-digit result $f(1.53) = -7.61$.

11. Recall that the derivative of a function f at a point x is defined by the equation

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

A computer has the capacity of imitating the limit operation by using a sequence of numbers h such as $h = 4^{-1}, 4^{-2}, 4^{-3}, \ldots, 4^{-n}, \ldots$ for they certainly approach zero rapidly.

The following MATLAB code is to compute $f'(x)$ at the point $x = 0.5$ with $f(x) = \sin(x)$.

```
clear all, clf
x=0.5; h=1; n=5; % Try n = 5, 10, 15, and 20
for i=1:n
    h=h/4; H(i)=h;
    y=(sin(x+h)-sin(x))/h; % Forward difference formula
%    y=(sin(x+h)-sin(x-h))/2/h; % Central difference formula
    error(i)=abs(cos(x)-y); % Truncation error
end
loglog(H,error), hold on
N=1; % Try N = 1 and 2
loglog(H,H.^(N),'--')
```

Perform this numerical experiment by running the code and interpret the results. Also study the following **central difference formula**:

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h} \text{ as } h \to 0.$$

Hint : Is a possible loss of significant digit phenomena at play here? Where?

12. The Taylor polynomial of degree n for $f(x) = e^x$ is $\sum_{i=0}^{n} (x^i / i!)$. Use the Taylor polynomial of degree nine to find an approximation to e^{-5} by

$$\text{(a) } e^{-5} \approx \sum_{i=0}^{9} \frac{(-5)^i}{i!} = \sum_{i=0}^{9} \frac{(-1)^i 5^i}{i!} \qquad \text{(b) } e^{-5} = \frac{1}{e^5} \approx \frac{1}{\sum_{i=0}^{9} (5^i / i!)}.$$

An approximate value of e^{-5} correct to three digits is 6.74×10^{-3}. Which formula, (a) or (b), gives the most accuracy, and why?

Hint : The rule of thumb is that if you have the choice, always prefer the formula void of possible cancellation (loss of siginificant digits) errors.

13. In the computer, it can happen that $a + x = a$ when $x \neq 0$. Explain why. Describe the set of n for which $1 + 10^{-n} = 1$ in your computer. The following MATLAB code illustrates the phenomenon.

```
a = 1; x = a;
while a + x > a
    x = x/10;
end, x
```

Hint : For a given precision, any number $|x| \le \text{eps}$ satisfy $1 + x = 1$ where eps is the so-called machine zero (type `eps` in Matlab command window). How is it related to the number of digits carried by the floating point representation in the machine? What if $a \neq 1$? Can you put $a + x = a$ into form $1 + y = 1$ for some y?

14. Establish that the recursion formula $y_n + 5y_{n-1} = \frac{1}{n}$ represents the integral

$$y_n = \int_0^1 \frac{x^n}{x+5}\,dx \quad \text{for} \quad n = 0,1,25 \ .$$

a) Compute the terms $y_1, y_2, y_3, y_4, y_5, \ldots$ starting with $y_0 = \ln(6/5) \approx 0.182$ in a three-decimal computing environment. Do you observe anything strange as we theoretically expect that $y_n > 0$ and $y_1 > y_2 > y_3 > y_4 > y_5 \ldots$.

b) Now try the recursion formula in the other direction $y_{n-1} = \frac{1}{5n} - \frac{1}{5} y_n$. Approximate a starting value by setting $y_{10} \approx y_9$, thus getting $y_9 \approx \frac{1}{60} \approx 0.017$ to compute $y_8, y_7, \ldots, y_0$.

Explain.

Hint : Forward, 5 is a multiplier and backward, 5 is a divisor. How does this effect rounding error accumulation?

15. It is known that

$$\pi^2 = 6 \lim_{n \to \infty} s_n, \qquad s_n = \sum_{k=1}^{n} \frac{1}{k^2}.$$

Compute the respective errors in using forward and backward summation to compute s_n for $n = 10^3$ in a four-digit environment using the following MATLAB code:

```
digits(4),
s = vpa(0);
for n = 0:1e3 % Use n = 1e3:-1:0 for backward summation
   s = vpa(s+1/n/n);
end
```

Interpret the results.

Hint : The correct way is to allow small positive numbers to add up to something before they encounter large numbers.

16. Consider the following two formulas:

$$f_1(x) = \sqrt{x}\left(\sqrt{x+1} - \sqrt{x}\right) \quad \text{and} \quad f_2(x) = \frac{\sqrt{x}}{\sqrt{x+1} + \sqrt{x}}.$$

These are theoretically equivalent, hence we expect them to give exactly the same value. How would you explain the result obtained by running the following MATLAB program to compute the values of the two formulas?

Hint : Is a possible loss of significant digit phenomena at play here? Where?

```
clear all
f1 = inline('sqrt(x)*(sqrt(x + 1) - sqrt(x))','x');
f2 = inline('sqrt(x)/(sqrt(x + 1) + sqrt(x))','x');
x = 1; format long e
for k = 1:15
  fprintf('At x=%15.0f, f1=%20.18f, f2=%20.18f \n',x,f1(x),f2(x));
  x = 10*x;
end
```