Chapter 1 - Taylor Polynomials

Problem 1.1 - 2b

Recall that, $P_n(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + ... + \frac{(x-a)^n}{n!}f^{(n)}(a)$

$f(x) = \sin x, a = \frac{\pi}{4}$

$f'(x) = \cos x, \quad f'(\frac{\pi}{4}) = \cos(\frac{\pi}{4}) = \frac{\sqrt{2}}{2}$

$f''(x) = -\sin x, \quad f''(\frac{\pi}{4}) = -\sin(\frac{\pi}{4}) = \frac{\sqrt{2}}{2}$

linear: $n = 1 \rightarrow P_1(x) = \sin(\frac{\pi}{4}) + (x - \frac{\pi}{4})\cos(\frac{\pi}{4}) = \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}(x - \frac{\pi}{4})$

quadratic: $n = 2 \rightarrow P_2(x) = p_1(x) + \frac{(x-\frac{\pi}{4})^2}{2!}(-\sin(\frac{\pi}{4})) =$
$\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}(x - \frac{\pi}{4}) - \frac{\sqrt{2}}{4}(x - \frac{\pi}{4})^2$

Problem 1.1 - 2c

$f(x) = e^{\cos x}, \quad a = 0$

$f'(x) = -\sin x e^{\cos x}, \quad f'(0) = -\sin(0)e^{\cos(0)} = 0$

$f''(x) = -\cos x e^{\cos x} + \sin^2 x e^{\cos x}, \quad f''(0) = -\cos(0)e^{\cos(0)} + \sin^2(0)e^{\cos(0)} = -e$

linear: $n = 1 \rightarrow P_1(x) = e^{cos(0)} + (x-0)f'(0) = e$

quadratic: $n = 2 \rightarrow P_2(x) = p_1(x) + \frac{(x-0)^2}{2!}f''(0) = e - e\frac{x^2}{2}$

Problem 1.1 - 7

$f(x) = \sin x, \quad a = 0, \quad -\frac{\pi}{2} \leq x \leq \frac{\pi}{2}$

$p_1(x) = \sin(0) + (x-0)\cos(0) = x$

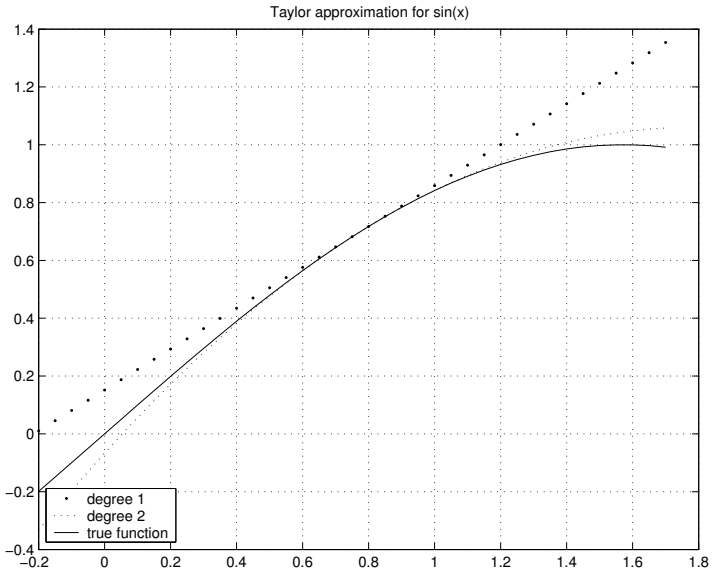

Figure 1: Problem 1.1-2b

$$p_3(x) = x + \frac{(x-0)^2}{2}(-\sin(0)) + \frac{(x-0)^3}{6}(-\cos(0)) = x - \frac{x^3}{6}$$

$$p_5(x) = x - \frac{x^3}{6} + \frac{(x-0)^4}{24}(\sin(0)) + \frac{(x-0)^5}{120}(\cos(0)) = x - \frac{x^3}{6} + \frac{x^5}{120}$$

Note 1: The larger the n the more accurate the approximation.
Note 2: $\sin x$ is an odd function and so is $p_n(x), n \geq 0$.

Problem 1.1 - 8

$f(x) = e^x, n = 1, 2, 3, x \in [-1, 2]$

$$\begin{aligned} p_n(x:a) &= e^a \sum_{j=0}^{n} \frac{(x-a)^j}{j!} \\ p_1(x:0) &= 1 + x \\ p_2(x:0) &= 1 + x + \frac{x^2}{2} \\ p_3(x:0) &= 1 + x + \frac{x^2}{2} + \frac{x^3}{6} \end{aligned}$$

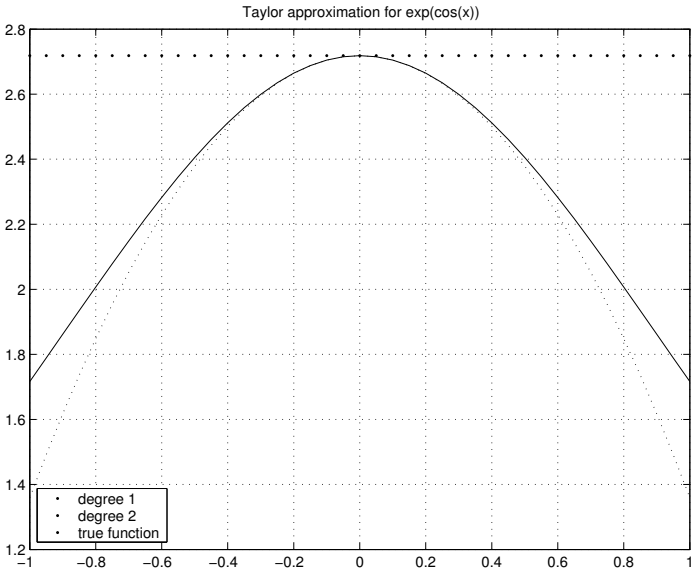

Figure 2: Problem 1.1-2c

$$
\begin{aligned}
p_1(x:1) &= e + (x-1)e = ex \\
p_2(x:1) &= e + (x-1)e + e\frac{x^2}{2} \\
p_3(x:1) &= e + (x-1)e + e\frac{(x-1)^2}{2} + e\frac{(x-1)^3}{6}
\end{aligned}
$$

Problem 1.2 - 5

$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - ... + (-1)^{n-1}\frac{x^{2n-1}}{(2n-1)!} + (-1)^n\frac{x^{2n+1}}{(2n+1)!}$

$|\sin x - p_{2n-1}(x)| \leq 0.001, -\frac{\pi}{2} \leq x \leq \frac{\pi}{2}$

$-0.001 \leq R_{2n-1}(x) \leq 0.001$

$-0.001 \leq \frac{x^{2n+1}}{(2n+1)!}\cos c \leq 0.001$

If we try different values for n, we see the error tolerance is satisfied for $n \geq 4$.This error is bounded by

x	$p_1(x)$	$p_3(x)$	$p_5(x)$	$\sin x$
$-\frac{\pi}{2}$	-1.5707963	-0.9248322	-1.0045249	-1.0
-1.0	-1.0	-0.8333333	-0.8416667	-0.8414710
-0.5	-0.5	-0.4791667	-0.4794271	-0.4794255
-0.1	-0.1	-0.0998333	-0.0998334	-0.0998334
0.0	0.0	0.0000000	0.0000000	0.0000000
0.1	0.1	0.0998333	0.0998334	0.0998334
0.5	0.5	0.4791667	0.4794271	0.4794255
1.0	1.0	0.8333333	0.8416667	0.8414710
$\frac{\pi}{2}$	1.5707963	0.9248322	1.0045249	1.0

Table 1: Problem 1.1-7

x	$p_1(x;0)$	$p_2(x;0)$	$p_3(x;0)$	e^x
-1.0	0	0.5	0.3333333	0.3678794
-0.5	0.5	0.625	0.6041666	0.6065306
-0.1	0.9	0.95	0.9498333	0.9048374
0.0	1.0	1.0	1.0000000	1.0000000
0.1	1.1	1.15	1.1501667	1.1051709
0.5	1.5	1.625	1.6458333	1.6487213
1.0	2.0	2.5	2.6666667	2.7182818
1.5	2.5	3.725	4.2875	4.4816891
2.0	3.0	5.0	6.3333333	7.3890561

Table 2: Problem 1.1-8a

$$\frac{(\frac{\pi}{2})^9}{9!} = 0.000160 \leq 0.001$$

while for $n = 3$ the error is bounded by

$$\frac{(\frac{\pi}{2})^7}{7!} = 0.00243 \geq 0.001$$

At $x = \frac{\pi}{2}$

$$p_7(\frac{\pi}{2}) = (\frac{\pi}{2}) - (\frac{(\frac{\pi}{2})^3}{3!}) + (\frac{(\frac{\pi}{2})^5}{5!}) - (\frac{(\frac{\pi}{2})^7}{7!}) \approx 0.999843$$

The error = (the true value) - (the Taylor approximation)

$$= \sin\frac{\pi}{2} - p_7(\frac{\pi}{2}) \approx 1 - 0.99843 = 0.000157 < 0.001$$

x	$p_1(x;1)$	$p_2(x;1)$	$p_3(x;1)$	e^x
-1.0	-2.7182818	2.7182818	-0.9060939	0.3678794
-0.5	-1.3591409	1.6989261	0.1698926	0.6065306
-0.1	-0.2718281	1.3727323	0.7697269	0.9048374
0.0	0.0	1.3591409	0.9060939	1.0000000
0.1	0.2718281	1.3727323	1.0424611	1.1051709
0.5	1.3591409	1.6989261	1.6422953	1.6487213
1.0	2.7182818	2.7182818	2.7182818	2.7182818
1.5	4.0774227	4.4172080	4.4738388	4.4816891
2.0	5.4365637	6.7957046	7.2487515	7.3890561

Table 3: Problem 1.1-8b

This verifies our result.

Problem 1.2 - 8

$$e^x = 1 + x + \tfrac{x^2}{2!} + ... + \tfrac{x^n}{n!} + \underbrace{\frac{x^{n+1}}{(n+1)!}e^c}_{\text{the error}}$$

$$|e^x - p_n(x)| \leq 10^{-5}, -1 \leq x \leq 1$$

$$-10^{-5} \leq R_n(x) \leq 10^{-5}$$

$$-10^{-5} \leq \tfrac{x^{n+1}}{(n+1)!}e^c \leq 10^{-5}$$

To find the bound, we try the case which x=1. Choose n such that $\frac{e}{(n+1)!} \leq 10^{-5}$, which requires $n \geq 8$.

Problem 1.3 - 9

$$e^x = 1 - \tfrac{x^3}{3!} + \tfrac{x^6}{6!} - \tfrac{x^9}{9!} + \tfrac{x^{12}}{12!} - \tfrac{x^{15}}{15!}$$

Regarding three ways of evaluating the polynomials, mentioned in the text,
1) evaluating each term independently

$$2 + 5 + 8 + 11 + 14 = 40$$

2) computing each power of x by multiplying it with the preceding power of x,

$$2 + 2 + 2 + 2 + 2 = 10$$

3) nested multiplication (the efficient answer)

$$1 + x^3(-\tfrac{1}{3!} + x^3(\tfrac{1}{6!} + x^3(-\tfrac{1}{9!} + x^3(\tfrac{1}{12!} + \tfrac{x^3}{15!}))))$$
$$2 + 1 + 1 + 1 + 1 = 6$$

It takes two multiplications to evaluate x^3. Assuming x^3 and all the coefficients to be stored, there are four multiplication in the above expression. Then for any x, six multiplications are required to evaluate $p(x)$.

Problem 1.3 - 10

$f(x) = 2e^{4x} - e^{3x} + 5e^x + 1$

To evaluate efficiently, put z instead of the term e^x
then the $f(x)$ will be,

$$f(x) = 1 + z(5 + z^2(1 + 2z))$$

which needs three multiplications.

Problem 1.3 - 11

$f(x) = e^x, \quad x \in [-1, 1]$

We know that,

$$e^x - p_n(x) = \tfrac{x^{n+1}}{(n+1)!}e^c$$

In order to find the degree of Taylor polynomial, by which the error is below 10^{-7},
we consider the case $c = 1$. So,

$$|e^x - p_n(x)| \leq 10^{-7}$$

$$\tfrac{e}{(n+1)!}e^c \leq 10^{-7}$$

which yields $n \geq 10$. So the approximation will be:

$$p_n(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ... + \frac{x^{10}}{10!}$$

Now we use MATLAB programing to see the result and compare it to true value of e^x, The table below, shows the result.

x	$p_{10}(x)$	e^x	$p_{10}(x) - e^x$
-1.0	0.36787946428571	0.36787944117144	2.3114e-008
-0.9	0.40656966705094	0.40656965974060	7.3103e-009
-0.8	0.44932896613407	0.44932896411722	2.0168e-009
-0.7	0.49658530425936	0.49658530379141	4.6795e-010
-0.6	0.54881163618057	0.54881163609403	8.6545e-011
-0.5	0.60653065972438	0.60653065971263	1.1742e-011
-0.4	0.67032004603666	0.67032004603564	1.0167e-012
-0.3	0.74081822068176	0.74081822068172	4.3077e-014
-0.2	0.81873075307798	0.81873075307798	4.4409e-016
-0.1	0.90483741803596	0.90483741803596	0
0.0	1.00000000000000	1.00000000000000	0
0.1	1.10517091807565	1.10517091807565	-4.4409e-016
0.2	1.22140275816017	1.22140275816017	-6.6613e-016
0.3	1.34985880757596	1.34985880757600	-4.5741e-014
0.4	1.49182469764018	1.49182469764127	-1.0869e-012
0.5	1.64872127068737	1.64872127070013	-1.2763e-011
0.6	1.82211880029486	1.82211880039051	-9.5652e-011
0.70	2.01375270694458	2.01375270747048	-5.2590e-010
0.8	2.22554092618768	2.22554092849247	-2.3048e-009
0.9	2.45960310266210	2.45960311115695	-8.4948e-009
1.0	2.71828180114638	2.71828182845905	-2.7313e-008

Table 4: Problem 1.3-11

As can be seen, the errors at each point do not exceed 10^{-7}.

Chapter 2 - Error and Computer Arithmetic

Problem 2.1 - 1acd

Storage of IEEE double precision floating-point format,

$$\underbrace{b_1}_{\sigma}\underbrace{b_2b_3...b_{12}}_{E}\underbrace{b_{13}b_{14}...b_{64}}_{\bar{x}}$$

In MATLAB, use the command
`> format hex`
to see the hexadecimal format (base16). Hexadecimal digits with binary equivalents are:

$$\begin{aligned}
1 &= 0001\\
2 &= 0010\\
3 &= 0011\\
4 &= 0100\\
5 &= 0101\\
6 &= 0110\\
7 &= 0111\\
8 &= 1000\\
9 &= 1001\\
a &= 1010\\
b &= 1011\\
c &= 1100\\
d &= 1101\\
e &= 1110\\
f &= 1111
\end{aligned}$$

After finding the IEEE double precision representation, replace each digit of the binary format with its hex equivalent to get the double precision format as 8 hex bytes:

$$\begin{aligned}
[8]_{10} &\equiv [4020000000000000]_{16} \equiv [0100000000100000\cdots0000]_2\\
[1.5]_{10} &\equiv [3ff8000000000000]_{16} \equiv [0011111111111000\cdots0000]_2\\
[0.5]_{10} &\equiv [3fe0000000000000]_{16} \equiv [0011111111100000\cdots0000]_2
\end{aligned}$$

Note that the relationships are not equalities, since the double precision format is a coded version of the mathematical number it represents.

Problem 2.2 - 1a

True value: x_T, approximated value: x_A
Error: $x_T - x_A$
Rel: $\frac{x_T - x_A}{x_T}$

Problem 2.2 - 1b

Error $= x_T - x_A = 0.028254 - 0.028271 = -0.000017$ Rel $= \frac{x_T - x_A}{x_T} = \frac{-0.000017}{0.028254} \approx -0.000602$
$x_A = 0.028271$ has three digits of accuracy relative to $x_T = 0.028254$. The error is less than 5 units in the fourth digit.

Problem 2.2 - 1d

Error $= x_T - x_A = \sqrt{2} - 1.414 \approx 1.414214 - 1.414 = 0.000214$
Rel $= \frac{x_T - x_A}{x_T} = \frac{0.000214}{1.414214} \approx 0.000151$
$x_A = 1.414$ has four significant digits relative to $\sqrt{2}$.

Problem 2.2 - 1e

Error $= x_T - x_A = \log 2 - 0.7 \approx 0.693147 - 0.7 = -0.006853$
Rel $= \frac{x_T - x_A}{x_T} = \frac{-0.006853}{0.693147} \approx -0.009887$
$x_A = 0.7$ has only 1 significant digit relative to log.

Problem 2.2 - 5

Considering the problem of *loss of significant digits*, sometimes reaaranging the functions might help, to avoid for example the subtraction of two values that are very close to each other.

Problem 2.2 - 5b

Note that $\log \frac{a}{b} = \log a - \log b$

$$\log(x+1) - \log x = \log \tfrac{x+1}{x} \quad or \quad \log(1+\tfrac{1}{x})$$

Problem 2.2 - 5d

Note that $(a^3 - 1) = (a-1)(a^2 + a + 1)$

$$\text{So,} \quad a - 1 = (a^{\frac{1}{3}} - 1)(a^2 + a + 1)$$

$$\sqrt[3]{1+x} - 1 = [(1+x)^{\frac{1}{3}} - 1]\frac{(1+x)^{\frac{2}{3}} + (1+x)^{\frac{1}{3}} + 1}{(1+x)^{\frac{2}{3}} + (1+x)^{\frac{1}{3}} + 1}$$

$$= \frac{1+x-1}{(1+x)^{\frac{2}{3}} + (1+x)^{\frac{1}{3}} + 1} = \frac{x}{(1+x)^{\frac{2}{3}} + (1+x)^{\frac{1}{3}} + 1}$$

In this way, if x is much smaller relative to one, the error of loss of signifacant digit is avoided.

Problem 2.2 - 5e

Note that $(a^2 - b^2) = (a-b)(a+b) \quad \text{or}(a-b) = (\sqrt{a} - \sqrt{b})(\sqrt{a} + \sqrt{b})$
or $(\sqrt{a} - \sqrt{b}) = \frac{(a-b)}{(\sqrt{a}+\sqrt{b})}$ So, we can say:

$$\frac{\sqrt{4+x}-2}{x} = \frac{(4+x)-4}{x(\sqrt{4+x}+2)}$$

Problem 2.2 - 6

In this problem, rather than rearranging the functions, we try to use the Taylor approximation to avoid the error of the loss of significant digit. Consider that the following formulas are in the case that x is near 0.

Problem 2.2 - 6a

$\frac{e^x-1}{x}$

$$e^x \approx 1 + x + \frac{x^2}{2!} + ... + \frac{x^n}{n!}$$
$$\frac{e^x-1}{x} \approx \frac{1+x+\frac{x^2}{2!}+...+\frac{x^n}{n!}-1}{x}$$
$$= \frac{x(1+\frac{x}{2!}+...+\frac{x^{n-1}}{n!})}{x}$$
$$= 1 + \frac{x}{2!} + ... + \frac{x^{n-1}}{n!} \quad n \geq 1$$

Problem 2.2 - 6d

$\frac{\log(1-x)+xe^{x/2}}{x^3}$, if $y = \log u$, then $y\prime = \frac{u\prime}{u}$

$\log(1-x)\prime = \frac{1}{x-1}$ $log(1-x)\prime\prime = \frac{-1}{(x-1)^2}$

$\log(1-x) = \log(1-0) + (x-0)\frac{1}{0-1} + \frac{(x-0)^2}{2!}\frac{-1}{(0-1)^2} + ...$

$\log(1-x) \approx -x - \frac{x^2}{2} - \frac{x^3}{3} - ... - \frac{x^{n+1}}{n+1}$

and for $xe^{x/2}$, replace the term $e^{x/2}$ with its Taylor approximation,

$xe^{x/2} \approx x(1 + (x/2)) + \frac{(x/2)^2}{2!} + ... + \frac{(x/2)^n}{n!})$

$\frac{\log(1-x)+xe^{x/2}}{x^3} \approx \frac{(-x-\frac{x^2}{2}-\frac{x^3}{3}-...-\frac{x^{n+1}}{n+1})+(x+\frac{x^2}{2}+...+\frac{x^{n+1}}{2^n n!})}{x^3}$

linear and quadratic terms get removed,

$= \frac{x^3(\frac{-1}{3}+\frac{1}{2^2.2!})+x^4(\frac{-1}{4}+\frac{1}{2^3.3!})+...}{x^3}$

$= [\frac{-1}{3} + \frac{1}{2^2.2!}] + [\frac{-1}{4} + \frac{1}{2^3.3!}]x + ... + [\frac{-1}{n+1} + \frac{1}{2^n.n!}]x^{n-2}, n \geq 2$

Problem 2.2 - 6g

$\frac{x-\sin x}{\tan x}$,

$\sin x \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} - ... + (-1)^{n-1}\frac{x^{2n-1}}{(2n-1)!}$

$\cos x \approx 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - ... + (-1)^{n}\frac{x^{2n}}{2n!}$

$\frac{x-\sin x}{\tan x} = \frac{(x-\sin x)\cos x}{\sin x}$

$\frac{(\frac{x^3}{3!}-\frac{x^5}{5!}+...+(-1)^{n-1}\frac{x^{2n-1}}{(2n-1)!})(1-\frac{x^2}{2!}+\frac{x^4}{4!}-...+(-1)^{n}\frac{x^{2n}}{2n!})}{x-\frac{x^3}{3!}+\frac{x^5}{5!}-...+(-1)^{n-1}\frac{x^{2n-1}}{(2n-1)!}}$

We can factor out one x from the numerator and the denominator,

$\frac{(\frac{x^2}{3!}-\frac{x^4}{5!}+...+(-1)^{n-1}\frac{x^{2n-2}}{(2n-1)!})(1-\frac{x^2}{2!}+\frac{x^4}{4!}-...+(-1)^{n}\frac{x^{2n}}{2n!})}{1-\frac{x^2}{3!}+\frac{x^4}{5!}-...+(-1)^{n-1}\frac{x^{2n-2}}{(2n-1)!}}$

From the Taylor approximation, $\frac{1}{1-x} \approx 1 + x, \quad x << 1$

$(\frac{x^2}{3!} - \frac{x^4}{5!} + ... + (-1)^{n-1}\frac{x^{2n-2}}{(2n-1)!})(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - ... + (-1)^{n}\frac{x^{2n}}{2n!})(1 + \frac{x^2}{3!} - \frac{x^4}{5!} + ... - (-1)^{n-1}\frac{x^{2n-2}}{(2n-1)!}$

$\approx \frac{x^2}{6} - \frac{23}{360}x^4 - \frac{11}{15120}x^6 - \frac{143}{604800}x^8 - \frac{361}{17107200}x^{10}$

Problem 2.2 - 9

We know that, in $ax^2 + bx + c = 0$, x equals to $\frac{-b\pm\sqrt{b^2-4ac}}{2a}$ and if b is an even value, we can re-write the x as $\frac{-b1\pm\sqrt{b1^2-ac}}{a}$, where $b1$ is $b/2$.

Here, we have $x^2 - 26x + 1 = 0$, so:

$$x = \frac{13\pm\sqrt{13^2-1}}{1}$$

Looking at this equation, we see how the error of loss of siginificant digit might happen.

Problem 2.2 - 9a

$x_1 = 13 + \sqrt{168} \approx 13.000 + 12.961 = 25.961$
while,

$x_2 = 13 - \sqrt{168} \approx 13.000 - 12.961 = 0.039000$

Three digits were lost in the computation of x_2. 0.039000 contains only two significant digits.

Problem 2.2 - 9b

$x_2 = (13 - \sqrt{168})\frac{13+\sqrt{168}}{13-\sqrt{168}} = \frac{169-168}{13+\sqrt{168}} \approx \frac{1.000}{25.961} = 0.38519$
which is correct in all five digits.

Problem 2.2 - 10

In this problem, we repeat Problem 9 for the equation $x^2 - 40x + 1 = 0$, using $\sqrt{399} \approx 19.975$.

Problem 2.2 - 10a

$x_1 = 20\sqrt{399} \approx 20.000 + 19.975 = 39.975$
while,
$x_2 = 20 - \sqrt{399} \approx 20.000 - 19.975 = 0.025000$
x_1 is accurate in all five digits. Though three digits were lost in x_2.

Problem 2.2 - 10b

$$x_2 = 20 - \sqrt{399} = \frac{1}{20+\sqrt{399}} \approx \frac{1.0000}{39.975} = 0.025016$$

which is correct in all five digits.

Problem 2.2 - 14

$$\cos(x) \approx 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - ... + (-1)^n \frac{x^{2n}}{2n!}$$

Using the above approximation to evaluate the $\cos(2\pi) = 1$

$$\cos(2\pi) \approx 1 - \frac{2\pi^2}{2!} + \frac{2\pi^4}{4!} - ... + (-1)^n \frac{2\pi^{2n}}{2n!}$$

Considering n as large as 10, try to calculate each term in the above equation individually.

n	$(-1)^n(2\pi)^{2n}/(2n)!$
0	1.000
1	-19.74
2	64.94
3	-85.64
4	60.24
5	-26.43
6	7.904
7	-1.714
8	0.2820
9	-0.03638
10	0.003780

Table 5: Problem 2.2-14

The exact sum of these 11 terms rounded to four digits, is 0.9894; so the error is 0.01106. This exceeds what we know theoretically of the bound of error, which is $(2\pi)/22! = 0.000323$. The inaccuracy is due to the loss-of-significance error. The large initial terms cancel one another to give a small sum.

Problem 2.2 - 19

Recall that IEEE double precision arithmetic has the form of:

$$\underbrace{b_1}_{\sigma}\ \underbrace{b_2 b_3 ... b_{12}}_{E}\ \underbrace{b_{13} b_{14} ... b_{64}}_{\bar{x}}$$

(1) Underflow: In normalized floating-point format, the smallest nonzero positive number is:

$$m = 2^{-1022}$$
$$\text{(in decimal form)} \quad \approx 2.225 \times 10^{-308}$$

since the bits in mantisaa are zero and so are the bits in exponent, except the least significant one $b_1 2$. So if any value is below this limit, it will be set to zero(the error of underflow).

$$x^{10} < m$$
$$|x| < \sqrt[10]{m} \approx 1.717 \times 10^{-31}$$
$$-1.717 \times 10^{-31} < x < 1.717 \times 10^{-31}$$

(2) Overflow: The largest positive value in this format will be the case when all bits in mantisaa are 1 and so are the bits in exponent.

$$M = [(2^{53} - 1).2^{-52}].2^{1023}$$

So, any value above M, will be set to M.

Considering $x^(10)$,$x^{10} = M$
We can take $\log_{10}$ from both sides, so

$\log_{10} x^{10} = \log_{10}([(2^{53} - 1).2^{-52}]2^{1023} \log_{10}((2^{53} - 1).2^{971})$

$10log_{10}x = log_{10}((2^{53} - 1).2^{971})$

Then $x \approx 6.6907 \times 10^{30}$.

Supplementary problems: IEEE floating point representation

Find the IEEE 754 standard single precision floating point values representing the decimal numbers: $666.7523, 3.1415926535, -3.0E127, 3.0E - 127$.

Here is a detailed explanation by Professor Stephen Birkett.

STEP 1:
Work first with the non-decimal part of the number:

$$666 = 1010011010$$

which can be found following the old Grade 2 method of subtracting powers of 2 from the number.

STEP 2:
Now work with the decimal part:

$$0.7523 = 0.11000000100101101011101110 01001...$$

which can be found by subtracting powers of 1/2 (starting with $(1/2)^1$ not $(1/2)^0$ since the 0 power is already included in the binary for the non-decimal part of the number). At this point you can decide how many significand bits will be required for the decimal part because you already know how many are used up for the non-decimal part of the number. So keep enough digits according to whether single or double precision is required.

STEP 3:
Combine your answers in steps 1 and 2 to get:

$$1010011010.11000000100101101011101110 01001....$$

and move the decimal to normalize the number:

$$1.01001101011000000100101101011101110 01001....$$

noting that the exponent is 9 positions in binary. Also note the arrangement in groups of 4 bits, which is useful to keep track of these sorts of numbers. [It also reflects the hexadecimal represenetation of course.]

STEP 4:
Write the significand bits, keeping in mind that the first 1 is dropped because its the assumed J bit:

$$01001101011000000100101 1...$$

and round to the appropriate number of bits required (e.g. 23 in single precision):

$$01001101011000000100110$$

STEP 5:
Calculate the exponent bits. In single precision you bias the exponent by adding $[127]_{10} = [01111111]_2$ to the value you found in step 3 ($9 = 00001001$):

$$01111111 + 00001001 = 10001000$$

This is the biased exponent for single precision representation.

32 BIT SINGLE PRECISION ANSWER:
23 significand bits: 01001101011000000100110
8 exponent bits: 10001000
1 sign bit: 0

The method for double precision keeps more significand bits and uses 1023 to bias the exponent. Otherwise the technique for converting is identical.

(a) 666.7523

Single Precision (32 bits):
Status: Normal
Sign bit (bit 31): 0, (0 for +ve; 1 for ve)
Exponent Field (bit 30-23): 10001000 ,(dec. value: 136-127 = 9)
Significand (bit 22-0): 1. 01001101011000000100110
decimal value of the significand: 1.3022506
Decimal Equivalent: 666.75232, (i.e. reversion to decimal)

Double Precision (64 bits):
Status: Normal
Sign bit (bit 63): 0 (0 for +ve; 1 for ve)
Exponent Field (bit 62-52): 10000001000 (dec. value: 1032-1023 = 9)
Significand (bit 22-0): 1.0100110101100000010010110101110111001100011000111111
decimal value of the significand: 1.3022505859375000
Decimal Equivalent: 666.75230000000000 (i.e. reversion to decimal)

(b) 3.1415926535

Single Precision (32 bits):
Status: Normal
Sign bit (bit 31): 0, (0 for +ve; 1 for ve)
Exponent Field (bit 30-23): 10000000 (dec. value: 128-127 = 1)
Significand (bit 22-0): 1. 10010010000111111011011
decimal value of the significand: 1.5707964
Decimal Equivalent: 3.1415927 (i.e. after reversion to decimal)

Double Precision (64 bits):
Status: Normal
Sign bit (bit 63): 0, (0 for +ve; 1 for ve)
Exponent Field (bit 62-52): 10000000000 (dec. value: 1024-1023 = 1)

Significand (bit 22-0): 1. 10010010000111111011010101000100000100010111010001 00
decimal value of the significand: 1.5707963267500000
Decimal Equivalent: 3.1415926535000000 (i.e. after reversion to decimal)

(c) 3.0 E 127

Single Precision (32 bits):
Status: Overflow
Sign bit (bit 31): 1, (0 for +ve; 1 for ve)
Exponent Field (bit 30-23): 11111111 (dec. value: 255-127 = 128)
Significand (bit 22-0): 0. 00000000000000000000
Decimal Equivalent: -Infinity (i.e. after reversion to decimal)

Double Precision (64 bits):
Status: Normal
Sign bit (bit 63): 1, (0 for +ve; 1 for ve)
Exponent Field (bit 62-52): 10110100110 (dec. value: 1446-1023 = 423)
Significand (bit 22-0): 1. 0110001010001011110111110111110100110101011000111100
decimal value of the significand: 1.3849467926678604)
Decimal Equivalent: -3.0000000000000000E+127 (reversion to decimal)

(d) +3.0 E 127

Single Precision (32 bits):
Status: Overflow
Sign bit (bit 31): 0, (0 for +ve; 1 for ve)
Exponent Field (bit 30-23): 11111111 (dec. value: 255-127 = 128)
Significand (bit 22-0): 0. 00000000000000000000
Decimal Equivalent: +Infinity (i.e. after reversion to decimal)

Double Precision (64 bits):
Status: Normal
Sign bit (bit 63): 0, (0 for +ve; 1 for ve)
Exponent Field (bit 62-52): 10110100110 (dec. value: 1446-1023 = 423)
Significand (bit 22-0): 1. 0110001010001011110111110111110100110101011000111100
decimal value of the significand: 1.3849467926678604)
Decimal Equivalent: 3.0000000000000000E+127 (reversion to decimal)