MATH 128A: Homework #1

Professor John Strain
Assignment: 1-18

Eric Chuu

June 30, 2016

Problem 1 (BF 1.1.5)

Show that the graph of $f(x) = x^3 + 2x + k$ crosses the x-axis exactly once, regardless of the constant k.

Solution
We first show that the graph of $f(x)$ crosses the x-axis. We consider the domain of f. For $x < 0$, $x^3 < 0$, so $x^3 + 2x + k < 2x + k$, which is negative if $x < -\frac{1}{2}k$. For $x > 0$, $x^3 > 0$, so $x^3 + 2x + k > 2x + k$, which is positive if $x > -\frac{1}{2}$. Since f is also continuous, then by the intermediate value theorem, there exists $c \in \mathbb{R}$ such that $f(c) = 0$, so the graph of $f(x)$ intersects the x-axis. Suppose now that there exists $c' \neq c$ such that $f(c') = 0$. Then $f(c) = f(c') = 0$. Then by Rolle's theorem, there exists d such that $f'(d) = 0$. However, if we take the derivative of f, we see that $f'(x) = 3x^2 + 2 > 0$ for all $x \in \mathbb{R}$. This contradiction establishes uniqueness of the point c. □

Problem 2 (BF 1.1.9)

Find the second Taylor Polynomial $P_2(x)$ for the function $f(x) = e^x \cos x$ about $x_0 = 0$.
(a) Use $P_2(0.5)$ to approximate $f(0.5)$. Find an upper bound for the error $|f(0.5) - P_2(0.5)|$ using the error formula, and compare it to the actual error.
(b) Find a bound for the error $|f(x) - P_2(x)|$ in using $P_2(x)$ to approximate $f(x)$ on the interval $[0, 1]$.
(c) Approximate $\int_0^1 f(x)dx$ using $\int_0^1 P_2(x)dx$.
(d) Find an upper bound for the error in (c) using $\int_0^1 |R_2(x)|$, and compare the bound to the actual error.

Solution
Since f is continuous and infinitely differentiable, we can use Taylor's theorem to find $P_2(x)$ for the function $f(x) = e^x \cos x$.

$$P_2(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2.$$

We calculate the derivatives and evaluate them at $x_0 = 0$.

$$f(x) = e^x, f(0) = 1$$
$$f'(x) = e^x - e^x \sin x, f'(0) = 1$$
$$f''(x) = -2e^x \sin x, f''(0) = 0$$

Plugging $f(0) = 1, f'(0) = 1, f''(0) = 0$ back into $P_2(x)$, we get

$$P_2(x) = 1 + x.$$

(a) We can use $P_2(x)$ to approximate $f(0.5)$, where $P_2(0.5) = 1.5$. If we consider the error using the error formula,

$$R_2(x) = \frac{f^{(3)}(\xi(x))}{3!}x^3 = \frac{-2e^{\xi(x)}(\sin(\xi(x)) + \cos(\xi(x)))}{3!}x^3 = -\frac{x^3}{3}\left(e^{\xi(x)}(\sin(\xi(x)) + \cos(\xi(x)))\right)$$

To find an upper bound for $R_2(x)$, we find maximize $e^{\xi(x)}(\sin(\xi(x)) + \cos(\xi(x)))$ over $\xi(x) \in [0, 0.5]$. Using a temporary substitution and taking the derivative we get $2e^x \cos x$, which is positive for $x \in [0, 0.5]$, so $e^x(\cos x + \sin x)$ is increasing for $x \in [0, 0.5]$, so the maximum occurs at $x = 0.5$. Then we can bound the error,

$$|R_2(0.5)| \leq \left|\frac{0.5^3}{3}e^{0.5}(\sin(0.5) + \cos(0.5))\right| \approx 0.093.$$

Comparing this to the actual error, $|f(0.5) - P_2(0.5)| = |1.446889 - 1.5| = 0.0531$, which is indeed less than than our calculated bound.

(b) Similar to part (a), we find an upper bound for the error, $R_2(x)$, on the interval $[0, 1]$. We can use our calculations from part (a) to save some time, and we see that

$$|R_2(x)| \leq \left|\frac{x^3}{3}\left(e^{\xi(x)}(\sin(\xi(x)) + \cos(\xi(x)))\right)\right|$$

Note the derivative of the inner quantity, which can be represented as $2e^x \cos x$, is positive for all $x \in [0, 1]$, which means that $e^{\xi(x)}(\sin(\xi(x)) + \cos(\xi(x)))$ is increasing on $[0, 1]$, so it attains its maximum at $x = 1$. Then,

$$|R_2(x)| \leq \left|\frac{1}{3}e^1(\sin(1) + \cos(1))\right| \approx 1.252.$$

(c) We can approximate $\int_0^1 f(x)$ using $\int_0^1 P_2(x)dx$.

$$\int_0^1 P_2(x)dx = \int_0^1 (1 + x)dx = \frac{3}{2}.$$

(d) We can then bound the error from (c) using the upper bound we calculated for $R_2(x)$ from part (b)

$$\int_0^1 |R_2(x)dx| \leq \int_0^1 \left|1.25x^3 dx\right| \approx 0.313.$$

Comparing this to the actual error,

$$\left|\int_0^1 f(x)dx - \int_0^1 P_2(x)dx\right| = |1.378 - 1.5| = 0.122 < 0.313$$

so the upper bound we found holds. □

Problem 3 (BF 1.2.15)

Use the 64-bit long real format to find the decimal equivalent of the following floating-point machine numbers.

Solution
(a)

$$\begin{aligned} s &= 0 \\ c &= 2^{10} + 2^3 + 2^1 = 1024 + 8 + 2 = 1034 \\ f &= \left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^7 + \left(\frac{1}{2}\right)^8 \end{aligned}$$

Then the decimal equivalent is given by

$$(-1)^s \cdot 2^{c-1023}(1+f) = 2^{11}\left(1 + \frac{1}{2} + \frac{1}{16} + \frac{1}{128} + \frac{1}{256}\right) = 3224.$$

(b)

$$\begin{aligned} s &= 1 \\ c &= 2^{10} + 2^3 + 2^1 = 1024 + 8 + 2 = 1034 \\ f &= \left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^7 + \left(\frac{1}{2}\right)^8 \end{aligned}$$

Then the decimal equivalent is given by

$$(-1)^s \cdot 2^{c-1023}(1+f) = (-1) \cdot 2^{11}\left(1 + \frac{1}{2} + \frac{1}{16} + \frac{1}{128} + \frac{1}{256}\right) = -3224.$$

(c)

$$\begin{aligned} s &= 0 \\ c &= 1024 - 1 = 1023 \\ f &= \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^7 + \left(\frac{1}{2}\right)^8 \end{aligned}$$

Then the decimal equivalent is given by

$$(-1)^s \cdot 2^{c-1023}(1+f) = 2^0 \cdot \left(1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{128} + \frac{1}{256}\right) = 1.32421875.$$

(d)

$$\begin{aligned} s &= 0 \\ c &= 1024 - 1 = 1023 \\ f &= \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^7 + \left(\frac{1}{2}\right)^8 + \left(\frac{1}{2}\right)^{52} \end{aligned}$$

Then the decimal equivalent is given by

$$\begin{aligned} (-1)^s \cdot 2^{c-1023}(1+f) &= 2^0 \cdot \left(1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{128} + \frac{1}{256} + \frac{1}{2^{52}}\right) \\ &= 1.32421875000000022204460492503130808472633361816406 25. \end{aligned}$$

□

Problem 4 (BF 1.2.21)

(a) Show that the polynomial nesting technique can be applied in the evaluation of

$$f(x) = 1.01e^{4x} - 4.62e^{3x} - 3.11e^{2x} + 12.2e^{x} - 1.99.$$

(b) Use three-digit rounding arithmetic, $e^{1.53} = 4.62$, 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 part (b) and (c) to the true three-digit result $f(1.53) = -7.61$.

Solution
(a) We can factor e^x out of the first 4 terms of $f(x)$ and continue to do so to get a nested representation,

$$f(x) = -1.99 + e^x(12.2 - e^x(3.11 + e^x(4.62 - 1.01e^x)))$$

(b) Calculating each of the exponentials using three-digit rounding arithmetic, we get

$$\begin{aligned} e^{1.53} &= 4.62 \\ e^{2\cdot 1.53} &= (e^{1.53})^2 = 21.3 \\ e^{3\cdot 1.53} &= (e^{1.53})^3 = 98.6 \\ e^{4\cdot 1.53} &= (e^{1.53})^4 = 456 \end{aligned}$$

Using these to evaluate $f(1.53)$, we get

$$\begin{aligned} f(1.53) &= 1.01(456) - 4.62(98.6) - 2.11(21.3) = 12.2(4.62) - 1.99 \\ &= 461 - 456 - 66.2 + 56.4 - 1.99 = -6.79 \end{aligned}$$

(c) If we first nest the calculations, we get

$$\begin{aligned} f(1.53) &= -1.99 + 4.62(12.2 - 4.62(3.11 + 4.62(4.62 - 1.01 \cdot 4.62))) \\ &= -1.99 + 4.62(12.2 - 4.62(3.11 + 4.62(4.62 - 4.67))) \\ &= -1.99 + 4.62(12.2 - 4.62(3.11 + 4.62(-0.05))) \\ &= -1.99 + 4.62(12.2 - 4.62(3.11 - 0.231)) \\ &= -1.99 + 4.62(12.2 - 4.62(2.88)) \\ &= -1.99 + 4.62(12.2 - 13.3) \\ &= -1.99 + 4.62(-1.10) \\ &= -1.99 - 5.08 \\ &= -7.07 \end{aligned}$$

(d) The answer we get from nesting the calculations is a more accurate than the answer we get from the calculation from part (b). The relative error from part from parts (b) and (c), denoted $\varepsilon_b, \varepsilon_c$ are as follows:

$$\begin{aligned} \varepsilon_b &:= \left| \frac{-7.61 + 6.79}{-7.61} \right| = 0.11 \\ \varepsilon_c &:= \left| \frac{-7.61 + 7.07}{-7.61} \right| = 0.07 \end{aligned}$$

and we see that the relative error of the answer from part (c) is smaller than the relative error of the answer from part (b). □

Problem 5 (BF 1.3.3)

The Maclaurin series for the arctangent function converges for $-1 < x \leq 1$ is given by

$$\arctan x = \lim_{n\to\infty} P_n(x) = \lim_{n\to\infty} \sum_{i=1}^{n} (-1)^{i+1} \frac{x^{2i-1}}{2i-1}$$

(a) Use the fact that $\tan \pi/4 = 1$ to determine the number of n terms of the series that need to be summed to ensure that $|4P_n(1) - \pi| < 10^{-3}$.
(b) How many terms of the series would we need to sum to obtain accuracy within 10^{-10}.

Solution
(a) Since $\tan \pi/4 = 1$, then for $x = 1$, $\arctan x \to \pi/4$, so

$$\pi = 4 \lim_{n\to\infty} \sum_{i=1}^{n} (-1)^{i+1} \frac{1}{2i-1}$$

To bound the error by 10^{-3}, we consider

$$|4P_n(x) - \pi| = \left| 4(-1)^{n+1} \frac{x^{2n-1}}{2n-1} \right|$$

For $x = 1$, the error is maximized, so we see that

$$|4P_n(1) - \pi| \leq \frac{4}{2n-1} < 10^{-3}$$

This occurs when $\frac{2n-1}{4} > 1000 \Rightarrow n > 2000$, so we need to sum at least 2000 terms to achieve the desired accuracy.
(b) To achieve accuracy within 10^{-10}, we just evaluate the last step from part (a):

$$|4P_n(1) - \pi| \leq \frac{4}{2n-1} < 10^{-3}$$

which occurs when $\frac{2n-1}{4} > 1000 \Rightarrow n > 20,000,000,000$ terms. □

Problem 6 (BF 1.3.5)

Another formula for computing π can be deduced from the identity $\pi/4 = 4 \arctan \frac{1}{5} - \arctan \frac{1}{239}$. Determine the number of terms that must be summed to ensure approximation to π within 10^{-3}

Solution
From the identity, we get

$$\pi = 16 \sum_{i=1}^{\infty} (-1)^{i+1} \frac{(1/5)^{2i-1}}{2i-1} - 4 \sum_{i=1}^{\infty} (-1)^{i+1} \frac{(1/239)^{2i-1}}{2i-1}$$

Since the first term in the difference is significantly larger than the second, we only need to bound the first one above by 10^{-3}.

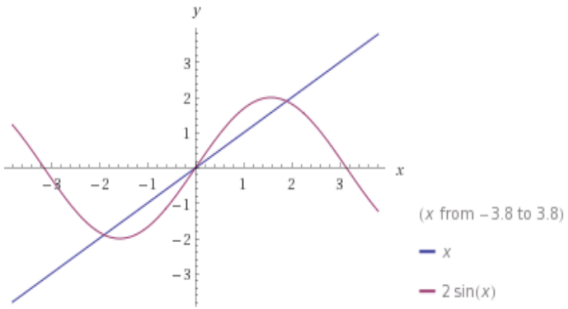
$$16 \sum_{i=1}^{\infty} (-1)^{i+1} \frac{(1/5)^{2i-1}}{2i-1} < 10^{-3} \Rightarrow \frac{16}{5^{2i-1}(2i-1)} < 10^{-3} \Rightarrow 5^{2i-1}(2i-1) > 16000.$$

If $i = 3$, then $5^{2\cdot 3-1}(2 \cdot 3 - 1) = 15625 < 16000$, but $i = 4$ would give a number surpassing 16000, so we must sum 3 terms to achieve desired accuracy. □

Problem 7 (BF 2.1.7)

(a) Sketch the graphs of $y = x$ and $y = 2\sin x$
(b) Use the Bisection method to find an approximation to within 10^{-5} to the first positive value of x with $x = 2\sin x$.

Solution
(a)



(b) The following matlab program finds the root within 10^{-5} to the first positive value of x with $x = 2\sin x$ using the bisection method. The result was: $x = 1.89549255$, as shown in the comments below the program.

```
function bisection()
    low = 1.5;                % f(1.5) < 0
    high = 2;                 % f(2) > 0
    mid = (low + high) / 2;
    tol = 10^(-5);
    f = fun(mid);
    iter = 0;
    while abs(f / min(low,high)) > tol
        if f < 0
            low = mid;
        else
            high = mid;
        end
        mid = (low + high) / 2;
        f = fun(mid);
        iter = iter + 1;
    end
    root = mid;
    fprintf('root: %.8f, f = %.3f, iter = %d', root, f, iter);

    function f = fun(x)
        f = x - 2*sin(x);
%{ OUTPUT:
>> bisection
root: 1.89549255, f = -0.000, iter = 14
%}
```

□

Problem 8 (BF 2.1.17)

Let (p_n) be the sequence defined by

$$p_n = \sum_{k=1}^{n} \frac{1}{k}.$$

Show that (p_n) diverges even though $\lim_{n\to\infty}(p_n - p_{n-1}) = 0$.

Solution
We first show that $\lim_{n\to\infty}(p_n - p_{n-1}) = 0$. Since $p_{n-1} = \sum_{k=1}^{n-1} \frac{1}{k}$, the difference yields just the extra term, $(p_n - p_{n-1}) = \frac{1}{n}$. Then taking limits, we see that $\lim_{n\to\infty}(p_n - p_{n-1}) = \lim_{n\to\infty} \frac{1}{n} = 0$. Suppose, now, that (p_n) converges, say $p_n \to p$. Then,

$$p = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} \dots$$

Since $\frac{1}{3} > \frac{1}{4}$, $\frac{1}{5} > \frac{1}{6}$, ..., then if we replace the terms on the left with the terms of the right in the harmonic series, we see that

$$\begin{aligned} p &\geq 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{6} + \frac{1}{6} \dots + \dots \\ &= 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{6} + \frac{1}{6}\right) \dots + \dots \\ &= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{3} + \dots \\ &= \frac{1}{2} + 1 + \frac{1}{2} + \frac{1}{3} + \dots \\ &= \frac{1}{2} + p, \end{aligned}$$

which is a contradiction. Thus, (p_n) diverges. □

Problem 9

Show that the equation $x = (\ln x)^x$ has at least one solution x in the interval $[\pi, 2\pi]$.

Solution
The equation $x = (\ln x)^x$ has a solution $x \in [\pi, 2\pi]$ if and only if $x - (\ln x)^x = 0$. Define $f(x) = x - (\ln x)^x$, where f is a continuous function defined on $[\pi, 2\pi]$. Evaluating f at the endpoints of this interval, we see that

$$\begin{aligned} f(2\pi) &\approx -42.64 < 0 \\ f(\pi) &\approx 1.61 > 0, \end{aligned}$$

so by the intermediate value theorem, there exists $x \in [\pi, 2\pi]$ such that $f(x) = 0$, and since $f(x) = x - (\ln x)^x$, then it follows that $x = (\ln x)^x$. □

Problem 10

Use the Taylor expansion

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

and the binomial theorem

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b_{n-k}$$

to prove that $e^{a+b} = e^a e^b$.

Solution
Expanding the right hand side first, we get

$$e^a e^b = \sum_{n=0}^{\infty} \frac{a^n}{n!} \sum_{n=0}^{\infty} \frac{b^n}{n!}. \tag{1}$$

We can use the Taylor expansion and the binomial theorem to expand the left hand side,

$$e^{a+b} = \sum_{n=0}^{\infty} \frac{(a+b)^n}{n!} = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \tag{2}$$

Recall that $\binom{n}{k} = \frac{n!}{(n-k)!k!}$, so

$$\sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{a^k b^{n-k}}{k!(n-k)!} \tag{3}$$

$$= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{a^k b^{n-k}}{k!(n-k)!} \tag{4}$$

$$= \sum_{k=0}^{\infty} \frac{a^k}{k!} \sum_{n=k}^{\infty} \frac{b^{n-k}}{(n-k)!} \tag{5}$$

Note that in (4), we used Fubini's Theorem to rewrite the summation. If we now let $m = n - k$, then we can rewrite (5) as

$$\sum_{k=0}^{\infty} \frac{a^k}{k!} \sum_{n=k}^{\infty} \frac{b^{n-k}}{(n-k)!} = \sum_{k=0}^{\infty} \frac{a^k}{k!} \sum_{m=0}^{\infty} \frac{b^m}{m!}, \tag{6}$$

which corresponds to the expanded form of the right hand side, as shown in (1), and we are done. □

Problem 11

Estimate the location of n and size $x^n/n!$ of the largest term of the series in problem 10 for any $x > 0$.

Solution
Recall that Stirling's approximation gives us

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Then we can rewrite the terms of the Taylor series as

$$f(n) := \frac{x^n}{n!} \approx \frac{x^n}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}$$

To maximize this, we take the derivative with respect to n, and we see that

$$f'(n) = -2n\log(x) + 2n\log(n) + 1.$$

If we then consider $n = x$, we see that

$$f'(x) = -2x\log(x) + 2x\log(x) + 1 = 1 > 0.$$

If we then consider $n = x - 1$, then we get

$$f'(x-1) = 2(x-1)\log(x-1) - 2(x-1)\log(x) + 1,$$

but for $x = 2$, we see that $f'(2-1) = 2\log(1) - 2\log(2) + 1 \approx -0.4 < 0$, so the root exists between $n = x$ and $n = x - 1$. However, since we are trying to find the location n of the largest term, it suffices to take $n = \lfloor x \rfloor$ as the position of the largest term for any $x > 0$. Since $n \approx x$ at the position of the largest term of the series, then the size of the of the largest term is approximately $x^n/n! \approx x^x/x!$. □

Problem 12

Fix integer $n \geq 1$, n points x_i with $|x_i| \leq 1$, n points y_j with $|y_j| \leq 1$, n coefficients f_j, and n coefficients g_j.
(a) Fix integer $k \geq 0$. Design an algorithm for evaluating

$$f(x) = \sum_{j=1}^{n} f_j (x y_j)^k$$

at n points x_i, in $O(n)$ operations.
(b) Find a degree-8 polynomial $P(x)$ with

$$|P(x) - \cos(x)| \leq 10^{-6}$$

on the interval $|x| \leq 1$.
(c) Design an algorithm for approximating

$$g(x) = \sum_{j=1}^{n} g_j \cos(x y_j)$$

at n points x_i in $O(n)$ operations with absolute error bounded by $10^{-6} \sum_{j=1}^{n} |g_j|$.

Solution

(a) Since x does not depend on on j in the definition of $f(x)$, we can rewrite this as

$$f(x) = x^k \sum_{j=1}^{n} f_j (y_j)^k$$

By pulling the x^k out of the summation, we can define $F := \sum_{j=1}^{n} f_j (y_j)^k$, where calculating F would have $O(n)$ floating point operations. Then, calculating $f(x_i) = F x_i^k$ would have $O(1)$ floating point operations for $1 \leq i \leq n$, so in total, this algorithm would have $O(n)$ floating point operations.

(b) We consider the leading five terms of the Taylor series expansion for $\cos(x)$:

$$P(x) := 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!}$$

and we use Sterling's approximation to check that the remainder is bounded above by the error shown above on the interval $|x| \leq 1$:

$$|R| = \left| \frac{x^9}{9!} \sin(\xi) \right| \leq \frac{1}{9!} = \left(\frac{e}{9} \right)^9 \cdot \frac{1}{\sqrt{2\pi 9}} \approx 10^{-6}$$

(c) From part (b), we found that we could approximate $\cos(x)$ with our choice of $P(x)$. If we use this approximation in conjunction with our algorithm from part (a), we see that we can approximate $g(x_i)$ with $\hat{g}(x_i) = \sum_{j=1}^{n} g_j P(x_i y_j)$ for $1 \leq i \leq n$. Then we can calculate the absolute error,

$$|g(x_i) - \hat{g}(x_i)| = \left| \sum_{j=1}^{n} g_j \left[\cos(x_i y_j) - P(x_i y_j) \right] \right| \leq \left| \sum_{j=1}^{n} g_j \cdot 10^{-6} \right| \leq 10^{-6} \sum_{j=1}^{n} |g_j|$$

By using $P(x)$ to approximate $\cos(x)$, we're able to use the same trick as we used in part (a), and by the same reasoning in part (a), this would only take $O(n)$ operations, and we have a sufficient algorithm. $\square$

Problem 13

Design an algorithm to evaluate

$$f(x) = \frac{e^x - 1 - x}{x^2}$$

in IEEE double precision arithmetic, to 12-digit accuracy for all machine numbers $|x| \leq 1$.

Solution
We first evaluate this using floating point arithmetic:

$$\begin{aligned} fl\left(\frac{e^x - 1 - x}{x^2}\right) &= \frac{((e^x(1+\delta_1) - 1)(1+\delta_2) - x)(1+\delta_3)}{x^2(1+\delta_4)}(1+\delta_5) \\ &= \frac{(e^x(1+5\delta_1') - 1(1+4\delta_2') - x(1+3\delta_3'))}{x^2} \\ &= \frac{e^x - 1 - x}{x^2} + \frac{e^x \cdot 5\delta_1 - 1 \cdot 4\delta_2 - x^3 \cdot \delta_3}{x^2} \end{aligned}$$

Then we want to bound the relative error by 10^{-12}, so

$$\texttt{rel\_error} \leq \frac{\frac{22\epsilon}{x^2}}{\frac{e^x-1-x}{x^2}} \leq \frac{22\epsilon/x^2}{1/2} \approx \frac{44\epsilon}{x^2} \leq 10^{-12}$$

This holds when $|x| \geq 0.3$. Thus, we can use the Taylor series expansion for e^x to continue evaluating $f(x)$,

$$\begin{aligned} fl\left(\frac{e^x - 1 - x}{x^2}\right) &= \frac{e^x - 1 - x}{x^2} + \frac{O(22\epsilon)}{x^2} \\ &= \frac{1}{2!} + \frac{1}{3!}x + \ldots + \frac{1}{n!}x^{n-2} + R_n, \end{aligned}$$

Then we consider the number of terms we need to sum to get the desired error

$$|R_n| \leq \frac{e^\xi}{(n+1)!}|x|^{n-1} \leq \frac{e}{(n+1)!} \leq e\left(\frac{e}{n+1}\right)^{n+1} \leq 10^{-12}$$

so for $n \approx 20$, we get the accuracy we want, and if we begin the summation from the right starting with the highest degree term, we get less error, so even though the rightmost term is involved in about 60 operations, the values are small, so the error is less significant. □

Problem 14

Write a Matlab program which tabulates the relative error in Stirling's approximation

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

for $1 \leq n \leq 10$.

Solution

```
% calculates and tabulates the sterling approximation
% for integers 1 to n
function error = sterling_approx(n)
    error_matrix = zeros(n, 1); % n x 1 matrix to store the errors
    for i = 1:n
        approx  = sqrt(2 * pi * i) * (i / exp(1)) ^ i;
        fprintf('approximation for %d!: %.4f', i, approx);
        fprintf('%s', 10);
        actual  = factorial(i);
        rel_err = abs(approx - actual) / actual;
        error_matrix(i,1) = rel_err;
    end
    error = error_matrix;


% calculates the factorial of the argument passed in
function f = factorial(n)
    result = 1;
    for i = n:-1:1
        result = result * i;
    end
    fprintf('%d! is: %d', n , result);
    fprintf('%s', 10);
    f = result;
```

The output when we call the `sterling_approx` function with argument $n = 10$ is shown on the following page. The final answer shows the relative errors for $n = 1, 2, \ldots, 10$. During each iteration, Sterling's approximation for each value of n is printed, along with the actual value of $n!$.

Output of `sterling_approx(10)`

```
%{
The output of the code is as follows:

>> sterling_approx(10)
approximation for 1!: 0.9221
1! is: 1
approximation for 2!: 1.9190
2! is: 2
approximation for 3!: 5.8362
3! is: 6
approximation for 4!: 23.5062
4! is: 24
approximation for 5!: 118.0192
5! is: 120
approximation for 6!: 710.0782
6! is: 720
approximation for 7!: 4980.3958
7! is: 5040
approximation for 8!: 39902.3955
8! is: 40320
approximation for 9!: 359536.8728
9! is: 362880
approximation for 10!: 3598695.6187
10! is: 3628800

ans = 0.0779
      0.0405
      0.0273
      0.0206
      0.0165
      0.0138
      0.0118
      0.0104
      0.0092
      0.0083
%}
```

Problem 15

The Fibonacci numbers f_n are defined by

$$f_{n+2} = f_{n+1} + f_n \tag{7}$$

with $f_0 = 0$ and $f_1 = 1$.
(a) Show that

$$\frac{f_{n+1}}{f_n} \to \varphi = \frac{1+\sqrt{5}}{2}$$

as $n \to \infty$ (b) Determine the rate of convergence of $\frac{f_{n+1}}{f_n} \to \varphi$

Solution
Taking the equation in (7) and dividing through by f_n, we get

$$\frac{f_{n+1}}{f_n} 1 + \frac{f_{n-1}}{f_n}.$$

Setting $\varphi_n = \frac{f_n}{f_{n-1}}$, we can rewrite this as

$$\varphi_{n+1} = 1 + \frac{1}{\varphi_n} \tag{8}$$

Suppose now that $\varphi_n \to \varphi$, then (8) becomes

$$\varphi = 1 + \frac{1}{\varphi},$$

so if we consider

$$|\varphi_{n+1} - \varphi| = \left| \frac{1}{\varphi_n} - \frac{1}{\varphi} \right| = \left| \frac{\varphi - \varphi_n}{\varphi_n \varphi} \right| \leq \frac{|\varphi - \varphi_n|}{1.6} \tag{9}$$

The last inequality holds because $\varphi_n \geq 1$ for all n, and $\varphi = \frac{1+\sqrt{5}}{6} \approx 1.6$. Thus, we can conclude that $|\varphi_{n+1} - \varphi| \to 0$ for n large enough. In other words, $\frac{f_{n+1}}{f_n} \to \varphi$ for n large enough and by the last inequality in (9), we know that the rate of convergence is $O\left(\left(\frac{1}{\varphi^2}\right)^n\right) = O\left(\left(\frac{1}{1.6^2}\right)^n\right) = O\left(2.6^{-n}\right)$. □

Problem 16

Show that the floating point arithmetic sums

$$s_n = \sum_{k=1}^{n} \frac{1}{k^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{n^2}$$

with accuracy $\mathcal{O}(n)\varepsilon$ from left to right, while summing from right to left gives accuracy $\mathcal{O}(\log n)\varepsilon$.

Solution
Summing from the right to left using floating point arithmetic yields

$$\begin{aligned} fl(s_n) &= 1 \cdot (1+\delta_n)\frac{1}{2^2} \cdot (1+2\delta_{n-1}) + \ldots + \frac{1}{(n-1)^2} \cdot (1+n\delta_2) + \frac{1}{n^2} \cdot (1+(n+1)\delta_1) \\ &= 1 + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{n^2} + 1 \cdot \delta_n + \frac{1}{2^2} \cdot 2\delta_{n-1} + \ldots + \frac{1}{n^2} \cdot (n+1)\delta_n \\ &= 1 + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{n^2} + O\left(\sum_{k=1}^{n} \frac{1}{k}\right)\delta \\ &\leq 1 + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{n^2} + O(\log n)\delta \end{aligned}$$

for $\delta \leq \varepsilon$, which gives us accuracy of $O(\log n)\varepsilon$. If we sum from the left to right, the $n\delta$ is multiplied to the first term, which gives us less accuracy,

$$fl(s_n) = 1 \cdot (1 + (n+2)\delta_1) + \frac{1}{2^2}(1 + (n+1)\delta_2) + \ldots + \frac{1}{n^2}(1 + 3\delta_n),$$

so the error is given by

$$\sum_{k=1}^{n} \frac{n+3-k}{k^2}\varepsilon = (n+3)\sum_{k=1}^{n} \frac{1}{k^2}\varepsilon - \sum_{k=1}^{n} \frac{1}{k}\varepsilon = (n+3)\sum_{k=1}^{n} \frac{1}{k^2}\varepsilon - (\log n)\varepsilon = O(n)\varepsilon$$

We can thus conclude that summing from right to left and giving more of the δ's to the smaller terms results in better accuracy than summing from left to right. □

Problem 17

Suppose a, b are floating point numbers with $a < b$. Show that

$$a \leq fl\left(\frac{a+b}{2}\right) \leq b, \tag{10}$$

in IEEE standard floating point arithmetic.

Solution
Note that if $a, b \in \mathbb{R}$, if we treat taking the floating point numbers of both numbers as a function, then this function is monotonic. That is, if $a \leq b$, then $fl(a) \leq fl(b)$. Since $a < b$, we clearly have $a \leq \frac{a+b}{2}$. Since a is a floating point number, then $a = fl(a) \leq fl\left(\frac{a+b}{2}\right)$, by monotonicity. Similarly, since $a < b$, we also have the inequality $\frac{a+b}{2} \leq b$. Since b is a floating point number, we look at their floating point representations, and we see that $fl\left(\frac{a+b}{2}\right) \leq fl(b) = b$, again by monotonicity. Putting the inequalities together, we get

$$a \leq fl\left(\frac{a+b}{2}\right) \leq b,$$

which is exactly the inequality in (10). □

Problem 18

Figure out exactly what sequence of intervals is produced by bisection for solving $x = 0$ with initial interval $[a_0, b_0] = [-1, 2]$. How many steps will it take to get maximum accuracy in IEEE standard floating point arithmetic.

Solution
We consider the first few intervals produced by bisection:

$$\begin{aligned} 0 &: [-1, 2] \\ 1 &: \left[-1.\frac{1}{2}\right] \\ 2 &: \left[-\frac{1}{4}, \frac{1}{2}\right] \\ 3 &: \left[-\frac{1}{4}, -\frac{1}{8}\right] \end{aligned}$$

We see that the upper limit of the interval goes down by a factor of 4 every other iteration, and the lower limit of the interval goes down by a factor of 4 whenever the upper limit stays the same. This iterative process ends when the sum of the upper and lower limits is 0, which happens either when the interval is $\left[-2^{-1074}, 0\right]$ or $\left[0, 2^{-1074}\right]$, which takes approximately 1074 steps. □