1.) How would you perform the following calculations to avoid cancellation?

i. Evaluate $\sqrt{x+1}-1$ for $x \simeq 0$.
First, let's rewrite our expression as

$$(\sqrt{x+1}-1)\cdot\frac{\sqrt{x+1}+1}{\sqrt{x+1}+1}=\frac{x}{\sqrt{x+1}+1}.$$

Using this new expression to compute $\sqrt{x+1}-1$ is favorable because it avoids cancellation error for $x \simeq 0$ by removing the similar term subtraction in the numerator and replaces it with safe addition in the denominator.

ii. Evaluate $\sin(x)-\sin(y)$ for $x \simeq y$.
Using a trig identity, we can rewrite our expression as

$$\sin(x)-\sin(y)=2\sin\left(\frac{x-y}{2}\right)\cos\left(\frac{x+y}{2}\right).$$

This expression still has subtraction between $x$ and $y$ but now our error produced by the subtraction is much more stable. The relative error of $\sin(x)$ and $\sin(y)$ is higher than $fl(x)$ and $fl(y)$ and so the subtraction between $\sin(x)$ and $\sin(y)$ amplifies the relative error of sin. When we use our rewritten expression, $x-y$ has a smaller relative error than the difference between the sines and so our overall calculation reduces cancellation error.

iii. Evaluate $\frac{1-\cos(x)}{\sin(x)}$ for $x \simeq 0$.
We can rewrite the expression above using a conjugate and a trig identity as follow:

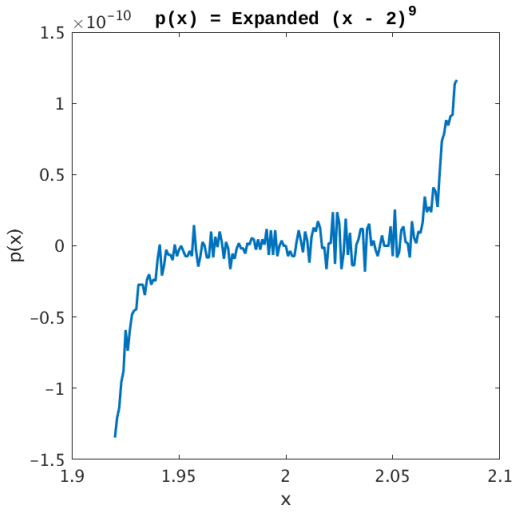
$$\begin{aligned}\frac{1-\cos(x)}{\sin(x)}\cdot\frac{1+\cos(x)}{1+\cos(x)}&=\frac{1-\cos^2(x)}{\sin(x)(1+\cos(x))}\\&=\frac{\sin^2(x)}{\sin(x)(1+\cos(x))}\\&=\frac{\sin(x)}{1+\cos(x)}.\end{aligned}$$

The rewritten expression no longer contains the cancellation in the numerator introduced by the $1-\cos(x)$ for $x \simeq 0$ and instead has a simple and accurate $\sin(x)$ in the numerator. In the denominator, our new expression has an accurate addition that will not shoot to 0 when $x \simeq 0$.

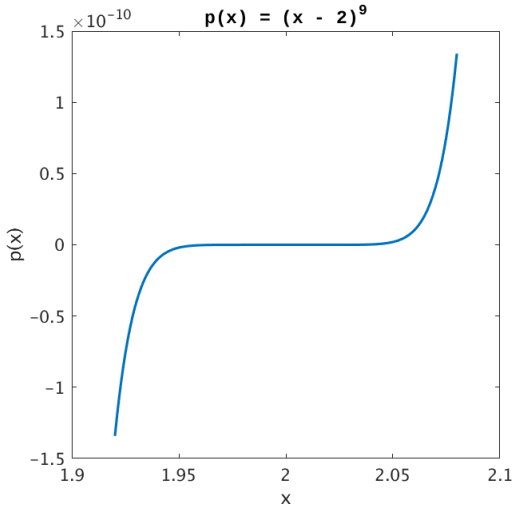
2.) Consider the polynomial

$$p(x) = (x-2)^9 = x^9 - 18x^8 + 144x^7 - 672x^6 + 2016x^5 - 4032x^4 + 5376x^3 - 4608x^2 + 2304x - 512.$$

i. Plot $p(x)$ for $x = 1.920, 1.921, 1.922, \ldots, 2.080$ (i.e. $x = [1.920 : 0.001 : 2.080]$) evaluating $p$ via its coefficients.




ii. Produce the same plot again, now evaluating $p$ via the expression $(x-2)^9$.




iii. What is the difference? What is causing the discrepancy? Which plot is correct?

The plot in part i is considerably more noisy than the plot in part ii. The difference in the results are due to the extra cancellation and rounding error caused by the abundance of addition and subtraction of small magnitude numbers with relatively large magnitude numbers. The plot in part ii is the more correct plot because it doesn't suffer from the increased error imposed by the expanded polynomial coefficients.

3.) **Cancellation of terms.** Consider computing $y = x_1 - x_2$ with $\tilde{x}_1 = x_1 + \Delta x_1$ and $\tilde{x}_2 = x_2 + \Delta x_2$ being approximations to the exact values. If the operation $x_1 - x_2$ is carried out exactly we have $\tilde{y} = y + \underbrace{(\Delta x_1 - \Delta x_2)}_{\Delta y}$.

i. Find upper bounds on the absolute error $|\Delta y|$ and the relative error $|\Delta y| / |y|$, when is the relative error large?
**Absolute error upper bound:**

$$\begin{aligned} |\Delta y| &= |y - \tilde{y}| \\ &= |x_1 - x_2 - (x_1 + \Delta x_1 - (x_2 + \Delta x_2))| \\ &= |\Delta x_1 - \Delta x_2| \\ &\leq |\Delta x_1| + |\Delta x_2| \,. \end{aligned}$$

**Relative error upper bound:** Using our absolute error upper bound, we have

$$\begin{aligned} \frac{|\Delta y|}{y} &\leq \frac{|\Delta x_1| + |\Delta x_2|}{|y|} \\ &= \frac{|\Delta x_1| + |\Delta x_2|}{|x_1 - x_2|} \end{aligned}$$

From this error bound, we can see that we would have large relative errors when the denominator $|x_1 - x_2|$ is small or when $x_1 \simeq x_2$.

ii. First manipulate $\cos(x+\delta) - \cos(x)$ into an expression without subtraction. Then, plot the difference between your expression and $\cos(x + \delta) - \cos(x)$ for $\delta = 10^{-16}, 10^{-15}, \ldots, 10^{0}$.
We can remove subtraction from the expression above by rewriting it as

$$\begin{aligned} \cos(x + \delta) - \cos(x) &= -2 \sin\left(\frac{x + \delta + x}{2}\right) \sin\left(\frac{x + \delta - x}{2}\right) \\ &= -2 \sin\left(\frac{2x + \delta}{2}\right) \sin\left(\frac{\delta}{2}\right) \end{aligned}$$

With our rewritten expression, we can generate the difference plot below:

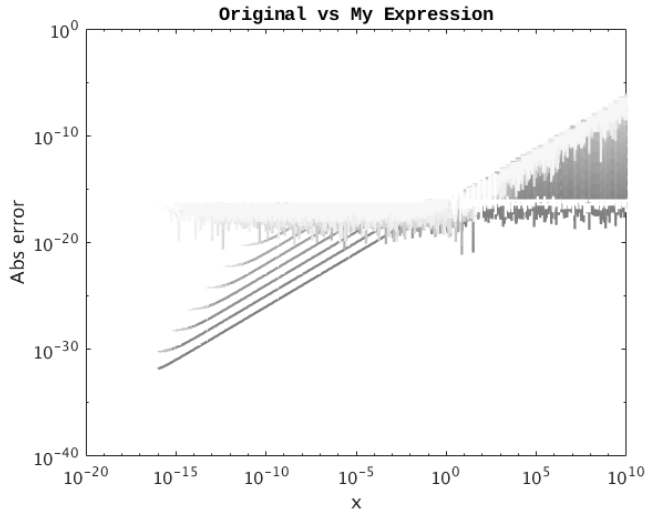
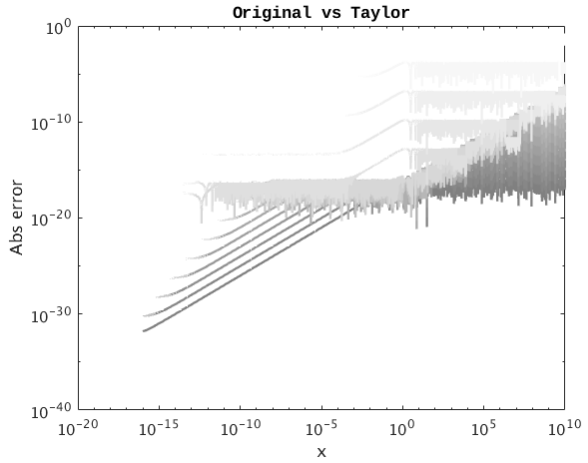
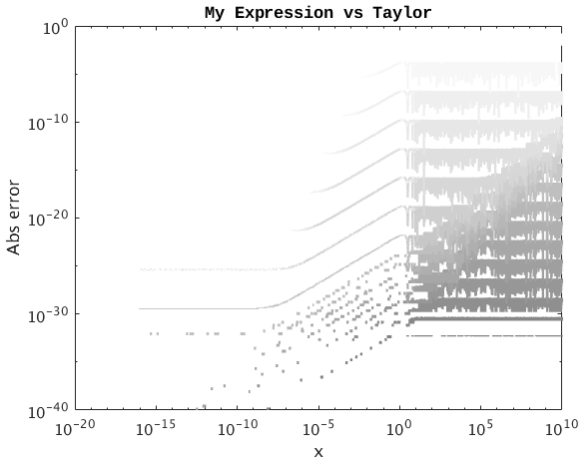



Figure 1: Absolute error between the original difference function and my expression. Darker lines = smaller values of $\delta$ starting at $\delta = 10^{-16}$ up to $\delta = 1$ for the whitest line.

iii. Taylor expansion yields $f(x + \delta) - f(x) = \delta f'(x) + \frac{\delta^2}{2!} f''(\xi), \xi \in [x, x + \delta]$. Use this expression to approximate $\cos(x + \delta) - \cos(x)$ for the same values of $\delta$ as in (ii). For what values of $\delta$ is each method better?




For really small values of $\delta$, the Taylor expansion is the most accurate but my expression is also very accurate. The accuracy of the Taylor expansion when $\delta$ is small makes sense because Taylor expansions are most accurate near their expansion center. As $\delta$ gets closer to 1, the original expression and my expression both perform accurately with little difference in there calculations. The Taylor expansion does not perform well for $\delta$ close to 1 because we don't have enough terms to extend the accuracy that far from the center. With that being said, my expression has the most reliable computations of the three expressions.

From the plots above, we can see that as we increase $\delta$, the Taylor approximation decreases in accuracy as a whole uniformly.

4.) Show that $(1+x)^n = 1 + nx + o(x)$ as $x \to 0$ where $n \in \mathbb{Z}$.

Consider

$$\begin{aligned}
\lim_{x\to 0} \frac{(1+x)^n - (1+nx)}{x} &= \lim_{x\to 0} \left( \frac{(1+x)^n - 1)}{x} - n \right) \\
&= \lim_{x\to 0} \left( \frac{(1+x)^n - 1)}{x} \right) - \lim_{x\to 0} n \\
&= \lim_{x\to 0} \left( \frac{(1+x)^n - 1)}{x} \right) - n \\
&= \lim_{x\to 0} \left( \frac{n(1+x)^{n-1})}{1} \right) - n \\
&= n - n = 0.
\end{aligned}$$

Therefore

$$(1+x)^n - (1+nx) = o(n)$$

or in other words

$$(1+x)^n = 1 + nx + o(n).$$

5.) Show that $x\sin(\sqrt{x}) = O(x^{3/2})$ as $x \to 0^+$.

Suppose we $x > 0$. Then

$$\begin{aligned}
\lim_{x\to 0^+} \left| \frac{x\sin(\sqrt{x})}{x^{3/2}} \right| &= \left| \lim_{x\to 0^+} \frac{\frac{1}{2}\sqrt{x}\cos(\sqrt{x}) + \sin(\sqrt{x})}{\frac{3}{2}x^{1/2}} \right| \\
&= \lim_{x\to 0^+} \frac{1}{3}\cos(\sqrt{x}) + \lim_{x\to 0^+} \frac{3\sin(\sqrt{x})}{2x^{1/2}} \\
&= 1 + \lim_{x\to 0^+} \frac{\frac{3}{2}x^{-1/2}\cos(\sqrt{x})}{x^{-1/2}} \\
&= 1 + 0 \\
&\leq 1.
\end{aligned}$$

Therefore, $x\sin(\sqrt{x}) = O(x^{3/2})$ as $x \to 0^+$.

6.) The function $f(x) = (x-5)^9$ has a root at $x = 5$ and is monotonically increasing (decreasing) for $x > 5$ ($x < 5$) and should thus be a suitable candidate for your function above. Set $a = 4.8$ and $b = 5.3$ and $tol = 1e-4$ and use bisection with:

i. $f(x) = (x-5)^9$.
Using my bisection code converges in 13 iterations to a root of $x = 5.0000280761718763$.

ii. The expanded version of $(x-5)^9$.
Using my bisection code on the expanded polynomial converges in 13 iterations to a root of $x = 5.1221740722656248$.

iii. Explain what is happening.
From question 2, we can see that a fully expanded polynomial may not always be the best idea when using floating point numbers. The loss of accuracy when using the bisection method in part (ii) can be attributed to the expanded polynomial. The expanded polynomial adds and subtracts relatively large and relatively small numbers which causes cancellation and rounding error to run rampant causing loss in significant digits. Thus the best accuracy we can hope for in the expanded polynomial is much less than the compact polynomial.

---

```
% Bisection routine for computing roots of functions
% Author: Caleb Jacobs
% Date last modified: 02-09-2021

format longE

% Function defintions
f1 = @(x) (x - 5).^9;
f2 = @(x) -1953125 + 3515625*x - 2812500*x.^2 + 1312500*x.^3 - ...
          393750*x.^4 + 78750*x.^5 - 10500*x.^6 + ...
          900*x.^7 - 45*x.^8 + x.^9;

% Set parameters
a    = 4.8;
b    = 5.31;
tol = 1e-4;
maxIts = 20;

% Find root with bisection
bisect(a, b, f1, tol, maxIts);
bisect(a, b, f2, tol, maxIts);

% Bisection method definition
function c = bisect(a, b, f, tol,  maxIts)
    % Check if root is guaranteed in the interval
    if f(a) * f(b) > 0
        fprintf('Root not guarnteed, exiting!\n\n')
        return
    end

    for i = 1:maxIts
        c = (b + a) / 2;                    % Midpoint of interval

        if f(c) == 0 || (b - c) < tol       % Check convergence criteria
            fprintf('Its used = %d\n',i)    % Display number of iterations
            fprintf('x = %.16f\n\n', c)     % Display root information
            return
        end

        if f(c) * f(a) >= 0                 % Update searching interval
            a = c;
        else
            b = c;
        end
    end

```

```
47      fprintf('Convergence could not happen to desired tolerance\n\n')
48  end
```