SECTION 1.2: ARITHMETIC COMPUTER ERROR

Even though the real line is continuous, computer arithmetic is being computed by a machine that has finite storage and memory.

IEEE double precision format describes how numbers are stored in a computer.
floating point (double precision) is constructed of 8 byte (or 64 bit) words

sign [1-bit]	exponent [11 bits]	mantissa (or fraction) [52 bits]

underflow - a number is too small to be represented so it is shown as 0
overflow - a number is too big to be represented so it can cause a program to error out (NaN)

Any real number can be put into normalized decimal format:

$$y = \pm 0.d_1 d_2 d_3 ... d_k d_{k+1} d_{k+2} ... \times 10^n$$

where $1 \leq d_1 \leq 9$.

There are two methods used to terminate the mantissa:

1.) chopping after the k^{th} digit

2.) rounding the k^{th} digit based on the $k+1$ digit

- If $d_{k+1} \geq 5$, $d_k = d_k + 1$.
- If $d_{k+1} < 5$, $d_k = d_k$.

Example 1.2.1: Given $\pi = 3.141592654$, determine the five-digit value of π using chopping and rounding.

Suppose that p^* is an approximation to an actual value p.

- actual error is $p - p^*$
- absolute error is $|p - p^*|$
- relative error is $\dfrac{|p - p^*|}{|p|}$ assuming $p \neq 0$

Example 1.2.2: Given that $p = 8!$ is approximated by $p^* = 40000$, compute the actual error, absolute error, and the relative error.

The approximation p^* approximates p to t significant digits if the relative error is less than or equal to 5×10^{-t}. In the above example,

$$\frac{300}{8!} = 0.0079365079 \leq 0.05 = 5 \times 10^{-2}$$

So 40000 approximates 8! to two significant digits.

Example 1.2.3: Perform the following calculations on $\frac{1}{3} - \frac{3}{11} + \frac{3}{20}$ using:

(a) exact arithmetic

(b) 3-digit chopping arithmetic

(c) 3-digit rounding arithmetic

Error can be reduced by:

- avoid adding and subtracting nearly identical numbers
- avoid adding and subtracting really large numbers
- minimize the number of operations performed

The quadratic formula states that the solutions to the quadratic equation $ax^2 + bx + c = 0$ are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

An alternate form of the quadratic formula can be derived as

$$x = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}$$

Example 1.2.4: Solve $x^2 + 62.1x + 1 = 0$ using 4-digit chopping.

(a) Use the original quadratic formula

(b) Use the alternative quadratic formula

(c) The approximate answers are $x_1 \approx -62.08389276$ and $x_2 \approx -0.0161072374$. Which answers for x_1 and x_2 were closest to the actual answers? Why?

Example 1.2.5: Evaluate $f(4.3)$ using 4-digit rounding given

$$f(x) = x^3 - 4.5x^2 + 2x + 0.1$$

then compare the answer to the exact value of 5.002.

Nested arithmetic is a method of rearranging calculations in an attempt to lessen the number of operations performed thus reducing round-off error.

Example 1.2.6: Evaluate $f(4.3)$ using 4-digit rounding and nesting given

$$f(x) = x^3 - 4.5x^2 + 2x + 0.1$$

then compare the answer to the exact value of 5.002.

Taylor's Theorem: Suppose $f \in C^n[a,b]$. $f^{(n+1)}$ exists on $[a,b]$, and $x_0 \in [a,b]$. For every $x \in [a,b]$, there exists a number $\mathcal{E}_x$ between x_0 and x with

$$f(x) = P_n(x) + R_n(x)$$

where the n^{th} degree Taylor polynomial of f about x_0 is

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

and the remainder term (or truncation error) is

$$R_n(x) = \frac{f^{(n+1)}(\mathcal{E}_x)}{(n+1)!}(x - x_0)^{n+1}$$

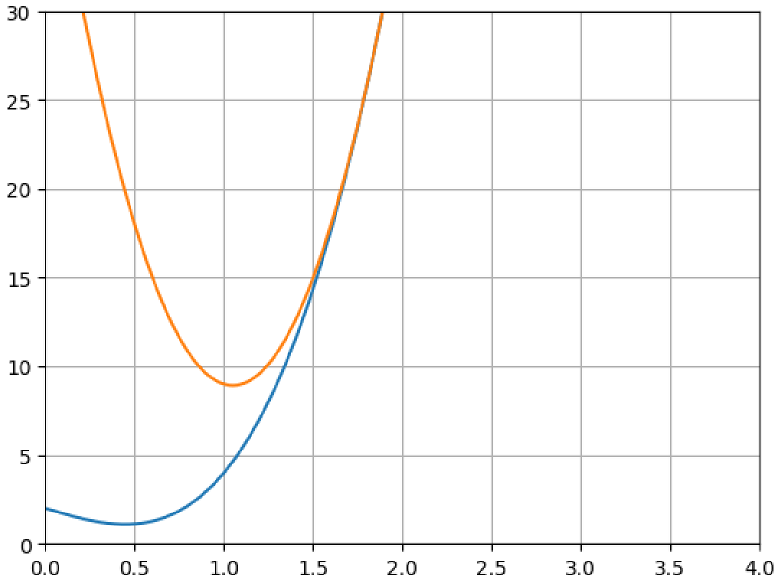
Example 1.1.1 Given $f(x) = 5x^3 - 3x + 2$ with $x_0 = 2$, answer the following:

(a) Determine $P_2(x)$

(b) Determine $R_2(x)$

(c) Use the second Taylor polynomial to approximate $f(1.5)$ and $f(1)$.

(d) Compute the absolute error and relative error $f(1.5)$ and $f(1)$.



Example 1.1.2 Let $g(x) = \cos x$ with $x_0 = 0$. Answer the following:

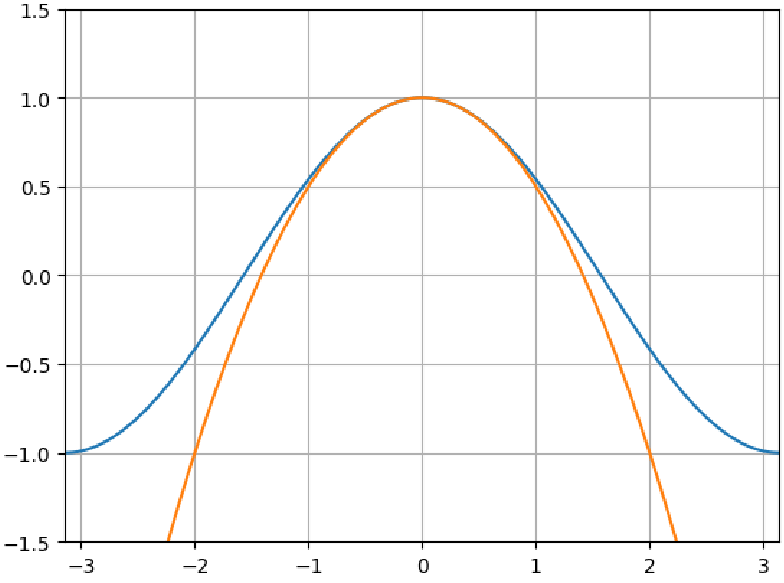
(a) Determine $P_2(x)$

(b) Compute $R_2(x)$

(c) Use the second Taylor polynomial to approximate $\cos(0.01)$.

(d) What is the absolute error?

(e) What is the relative error?



SECTION 1.3: ALGORITHM CONVERGENCE

An algorithm is a procedure that solves a problem in a finite number of steps. It is considered stable if small changes to an intial input produces small changes in the output. Some algorithms are only stable for certain inputs. These are called conditionally stable.

Definition of Convergence: Suppose $\{\beta_n\}_{n=1}^{\infty}$ is a sequence known to converge to zero and $\{\alpha_n\}_{n=1}^{\infty}$ converges to a number α. If a positive constant K exists with

$$|\alpha_n - \alpha| \leq K|\beta_n| \text{ for large } n,$$

then $\{\alpha_n\}_{n=1}^{\infty}$ converges to α with a rate of convergence $O\left(\beta_n\right)$.

$$\alpha_n = \alpha + O\left(\beta_n\right)$$

Example 1.3.1: Find the rates of convergence for the following sequences:

(a) $\alpha_n = \dfrac{n+1}{n^2}$

(b) $\alpha_n = \dfrac{n+3}{n^3}$

Example 1.3.2: Use Taylor Polynomials to determine the rates of convergence for the following as $h \to 0$:

(a) $\lim_{h\to 0} \left(\cos h + \dfrac{h^2}{2}\right) = 1$

(b) $\lim_{h\to 0} \left(\dfrac{\sin h - h\cos h}{h}\right) = 0$

(c) $\lim_{h\to 0} \left(\dfrac{1 - e^{-h}}{h}\right) = 1$

SECTION 2.1: BISECTION METHOD

Recall that a root or zero of an equation is any x that makes the statment $f(x) = 0$ true.

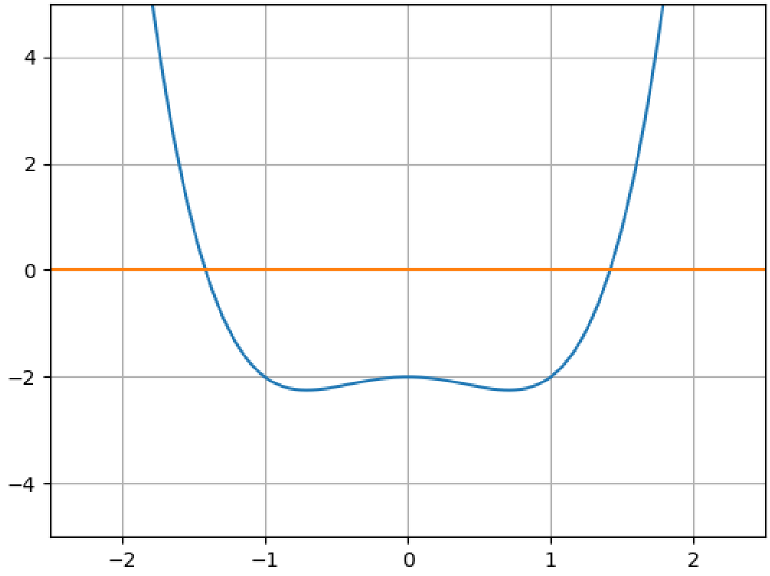
Intermediate Value Theorem: If $f \in C]a, b]$ and K is any number between $f(a)$ and $f(b)$, then there exists a number c in (a, b) for which $f(c) = K$.

To approximate a root, find x_0 and x_1 such that $x_0 < x_1$ and $f(x_0) \cdot f(x_1) < 0$.

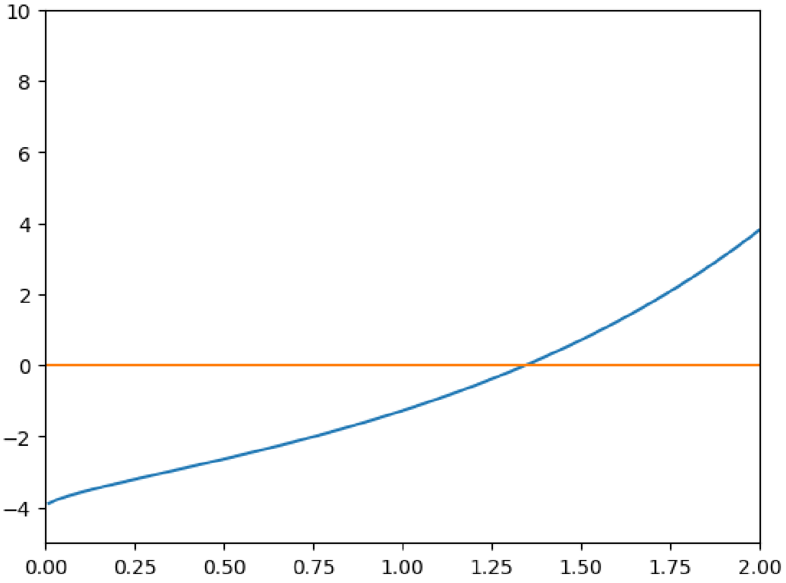
Bisection Method Process:

- $p_1 = \dfrac{x_1 + x_0}{2}$
- If $f(p_1) = 0$ then $p = p_1$ and we have found the root.
- If $f(p_1) \neq 0$, then compare the sign of $f(p_1)$ to $f(x_0)$ and $f(x_1)$
    - If $f(p_1)$ is the same sign as $f(x_0)$, replace x_0 with p_1
    - Else, replace x_1 with p_1
- repeat the process until the desired tolerance is reached

Example 2.1.1: Perform four iterations of the Bisection Method to solve $f(x) = x^4 - x^2 - 2 = 0$ using $x_0 = 1$ and $x_1 = 2$.



Example 2.1.2: Perform three iterations of the Bisection Method to solve $g(x) = e^x + \sqrt{x} - 5 = 0$ using $x_0 = 1.25$ and $x_1 = 1.5$.



The sequence $\{p_n\}_{n=1}^{\infty}$ converges to p with a rate of convergence of $O\left(\frac{1}{2^n}\right)$

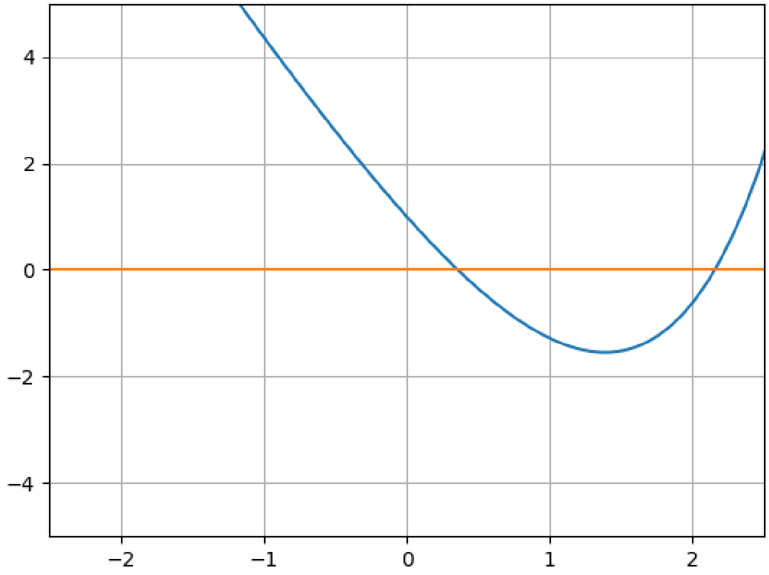
Example 2.1.3: Determine the number of iterations necessary to solve $f(x) = x^4 - x^2 - 2 = 0$ with accuracy 10^{-3} given $x_0 = 1$ and $x_1 = 2$

Example 2.1.4: Determine the number of iterations necessary to solve $g(x) = e^x + \sqrt{x} - 5 = 0$ with accuracy 10^{-6} given $x_0 = 1.25$ and $x_1 = 1.5$

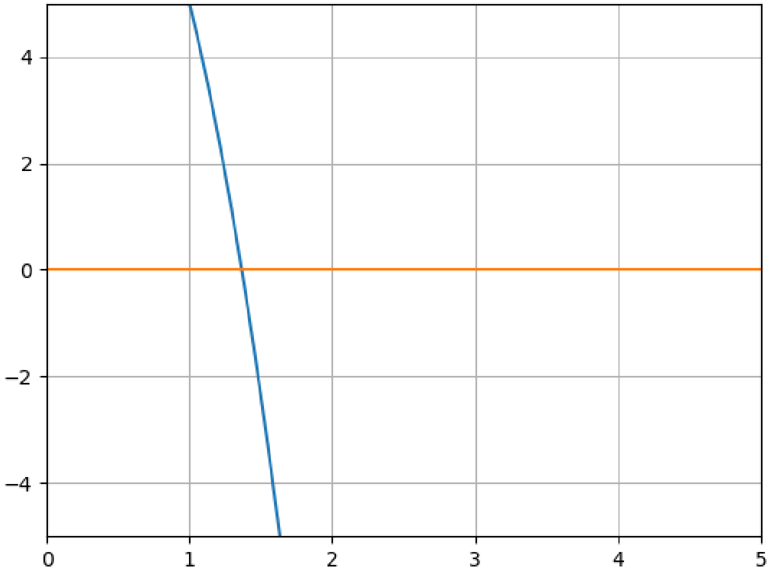
The number p is a fixed point for a given function g if $p = g(p)$. If a function has a fixed point p, then $f(x) = x - g(x)$ has a zero at p.

Example 2.2.1: Determine all fixed points for the function $g(x) = x^2 - 2$.

Example 2.2.2: Find a root of $g(x) = e^x - 4x$ using fixed point iteration given $p_0 = -1$.



Example 2.2.3: Find a root of $g(x) = -x^3 - 4x^2 + 10$ using fixed point iteration given $p_0 = 1.5$.



Fixed Point Existence Theorem: If $g \in C[a, b]$ and $g(x) \in [a, b]$ for all $x \in [a, b]$, then g has at least one fixed point in $[a, b]$. Also, if $g'(x)$ exists on (a, b) and a positive constant $k < 1$ exists such that

$$|g'(x)| \leq k \text{ for all } x \in (a, b),$$

then there is exactly one fixed point in $[a, b]$.

For the last example, consider the interval $[1, 2]$ for $p_0 = 1.5$.

$$g(x) = -x^3 - 4x^2 + x + 10 \text{ maps to the interval } [-12, 6]$$

$$g'(x) = -3x^2 - 8x + 1 \text{ maps to the interval } [-27, -10]$$

For Example 2.2.2, consider the interval $[-2, 1]$ for $p_0 = -1$.

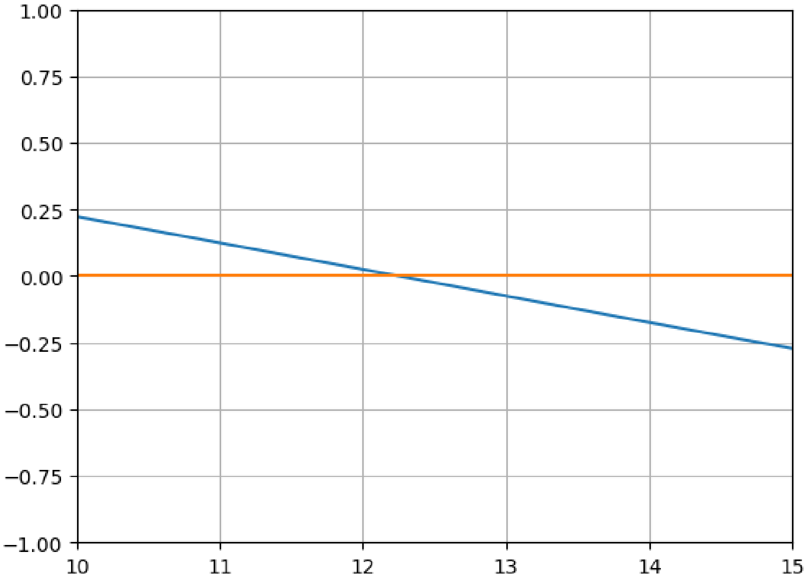
$$g(x) = \frac{e^x}{4} \text{ maps to the interval } [0.0338338208, 0.6795904571]$$

$$g'(x) = \frac{e^x}{4} \text{ maps to the interval } [0.0338338208, 0.6795904571]$$

Example 2.2.4: Confirm if a fixed point exists for

$$g(x) = \sin(0.1x + 8.2)$$

on the interval $[12, 13]$. If it does exist and the fixed point iteration converges, use $p_0 = 12.5$ to approximate the zero of g.



SECTION 2.3: NEWTON AND SECANT METHOD

Consider the first Taylor polynomial for a function $f(x)$ expanded about $p_0 \in [a, b]$ where $f'(p_0) \neq 0$ and $|p - p_0|$ is small.

$$f(p) = f(p_0) + (p - p_0)f'(p_0) + \frac{(p - p_0)^2}{2} f''\left(\mathcal{E}_x\right)$$

Newton's Method is derived by setting the Taylor Polynomial equal to 0 and assuming the remainder term approaches 0.

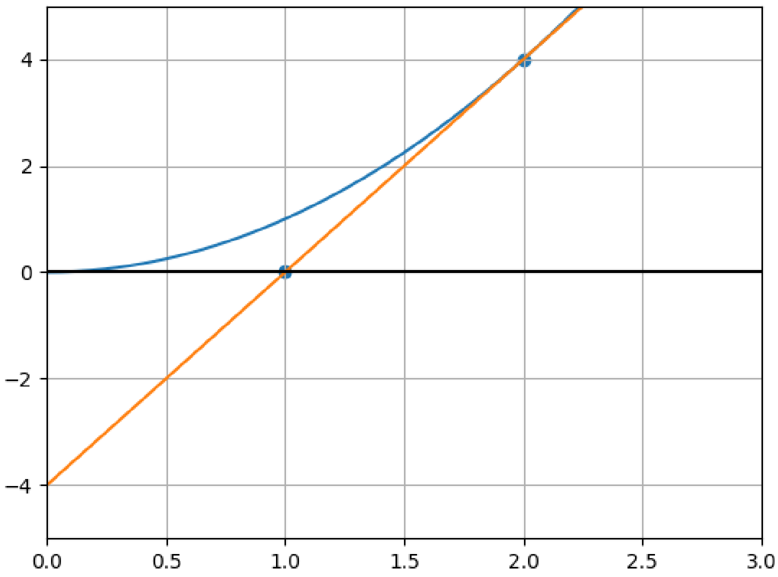
$$0 \approx f(p_0) + (p - p_0)f'(p_0)$$

which implies

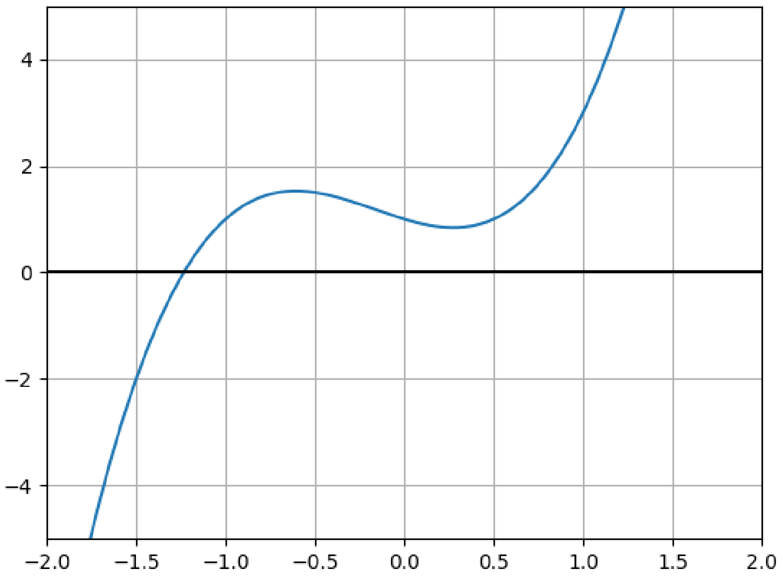
$$p \approx p_0 - \frac{f(p_0)}{f'(p_0)}$$

Newton's Method generates the sequence $\{p_n\}_{n=0}^{\infty}$ where

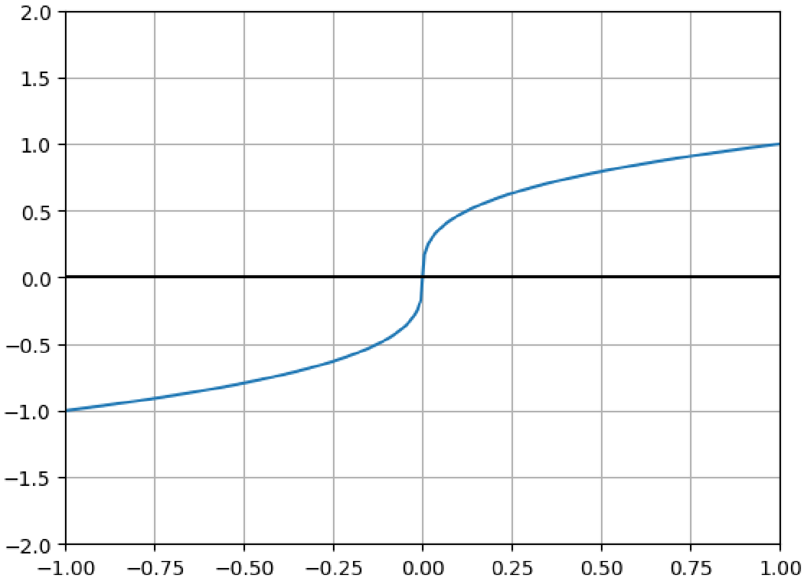
$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \text{ for } n \geq 1$$



Example 2.3.1: Use Newton's Method to approximate the zero of $f(x) = 2x^3 + x^2 - x + 1$ given $p_0 = -1.2$.



Example 2.3.2: Use Newton's Method to approximate the zero of $g(x) = \sqrt[3]{x}$ given $p_0 = 0.1$.



Sometimes it is difficult to compute the derivative of the function or is computationally expensive. Instead of using the derivate which is the slope of the tangent line, we can use the slope of the secant line instead.

$$f'(p_{n-1}) \approx \frac{f(p_{n-2}) - f(p_{n-1})}{p_{n-2} - p_{n-1}}$$

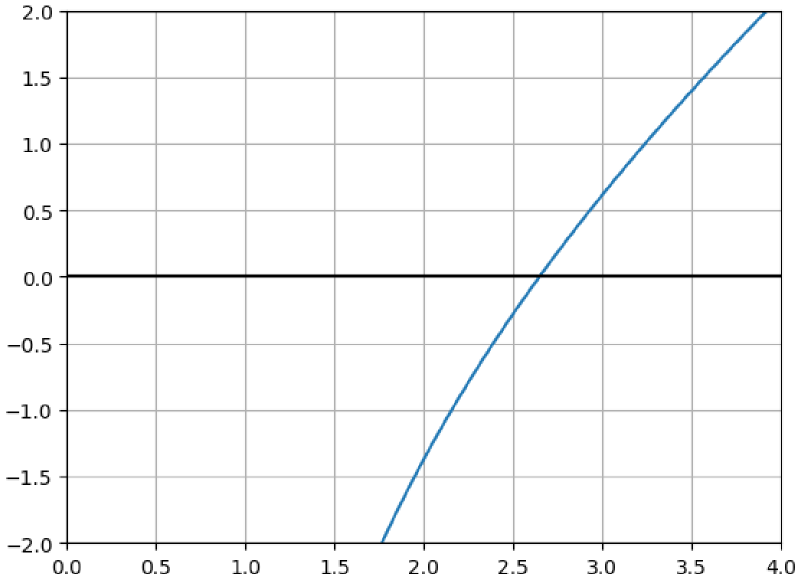
Thus, plugging in the secant formula into Newton's Method given close initial guesses of p_0 and p_1 results in

$$p_n = p_{n-1} - \frac{f(p_{n-1})(p_{n-1} - p_{n-2})}{f(p_{n-1}) - f(p_{n-2})}$$

This is called the Secant Method.

Example 2.3.3: Use the Secant Method to approximate the zero of $f(x) = 2x^3 + x^2 - x + 1$ given $p_0 = -1.2$ and $p_1 = -1.3$.

Example 2.3.4: Use the Secant Method to approximate the zero of $g(x) = \dfrac{x^2 - 0.4097x - 5.9161}{x}$ given $p_0 = 2.5$ and $p_1 = 3$.



SECTION 2.5: ACCELERATING CONVERGENCE

Order of Convergence: Suppose $\{p_n\}_{n=0}^{\infty}$ is a sequence that converges to p, with $p_n \neq p$ for all n. If positive constants λ and α exsit with

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^{\alpha}} = \lambda$$

then $\{p_n\}_{n=0}^{\infty}$ converges to p with order α, with asymptotic error constant λ.

- If $\alpha = 1$ (and $\lambda < 1$), the sequence is linearly convergent.
- If $\alpha = 2$, the sequence is quadratically convergent.

Method	**Advantages**	**Disadvantages**
Bisection	Always converges	slow
Fixed Point	Easy to implement	different formats for $p = g(p)$ does not always converge
Newton's Method	very fast	must use derivative function must use intial guesses close to root can error out if the root has a multiplicity
Secant Method	fast doesn't need derivative function	requires two inital values inital guesses must be close to root

Aitken's Δ^2 Method can accelerate the convergence of a linearly convergent sequence. The idea is that

$$\frac{p_{n+1} - p}{p_n - p} \approx \frac{p_{n+2} - p}{p_{n+1} - p}$$

Solving this approximation for p yields the formula

$$p = p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}$$

The **forward difference** (denoted as Δp_n) is defined by

$$\Delta p_n = p_{n+1} - p_n \text{ for } n \geq 0$$

Thus, the formula for the Aitken's Δ^2 Method can be written as

$$\hat{p_n} = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n}$$

Example 2.5.1: Generate the first three terms of the sequence $\{\hat{p_n}\}$ using Aitken's Δ^2 method given

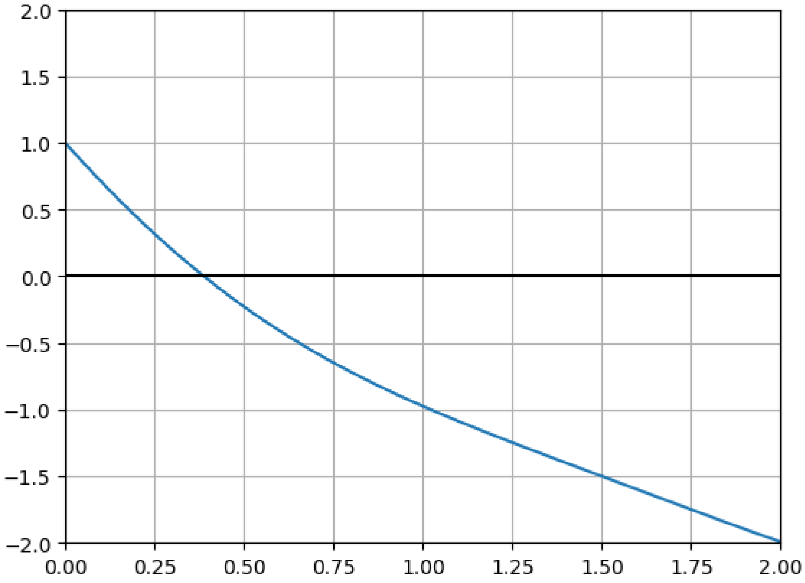
$$p_0 = 0.5, p_n = 3^{-p_{n-1}}$$

Steffensen's Method applies Aitken's Δ^2 method to fixed point iteration accelerating it to quadratic convergence. For this method, every third term is generated by Aitken's Method and the others use fixed-point iteration.

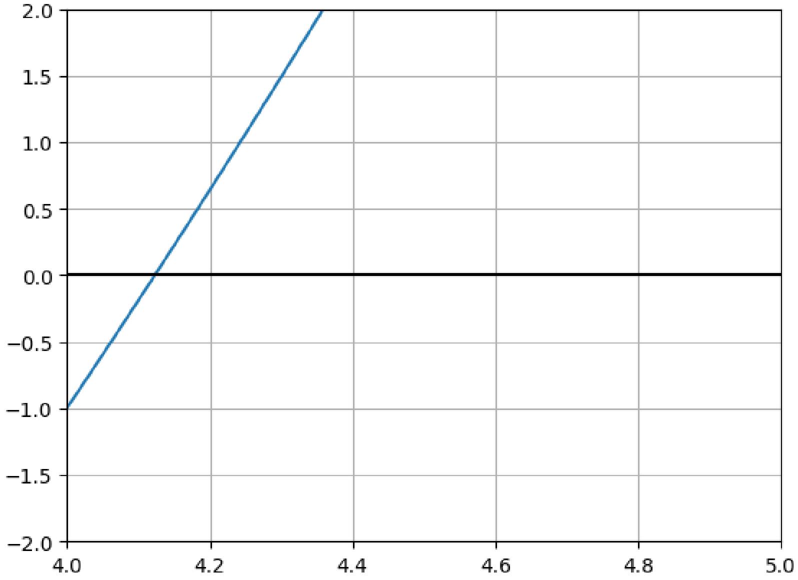
$$p_0^{(0)}, p_1^{(0)} = g\left(p_0^{(0)}\right), p_2^{(0)} = g\left(p_1^{(0)}\right),$$

$$p_0^{(1)} = \{\Delta^2\}\left(p_0^{(0)}\right), p_1^{(1)} = g\left(p_0^{(1)}\right), p_2^{(1)} = g\left(p_1^{(1)}\right), \ldots$$

Example 2.5.2: Let $g(x) = (\sin x - 1)^2 - x$. Use Steffensen's method with $p_0^{(0)} = 1$ to determine $p_0^{(1)}$.



Example 2.5.3: Approximate the solutions to $f(x) = x^2 - 17$ using the fixed point $x = \dfrac{17}{x}$. Use Steffensen's method with $p_0^{(0)} = 4$ to determine $p_0^{(2)}$.



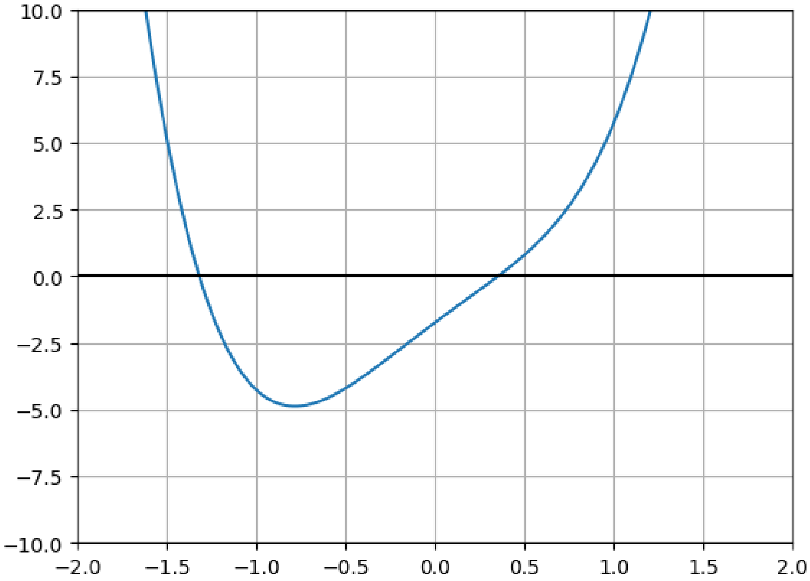
Fundamental Theorem of Algebra: If $P(x)$ is a polynomial of degree $n \geq 1$ with real or complex coefficients, then $P(x) = 0$ has at least one root.

$$P(x) = a_n\,(x - x_1)^{m_1}\,(x - x_2)^{m_2} \ldots (x - x_k)^{m_k}$$

where x_i are roots (possibly complex) of the polynomial with respective multiplicity m_i. Also, $\Sigma_{i=1}^{k} m_i = n$.

Horner's Method of finding roots for a polynomial uses Newton's method with the nesting technique applied. Let $P(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$.
Define $b_n = a_n$ and $b_k = a_k + b_{k+1}x_0$ for $k = n-1, n-2, \ldots, 1, 0$.
Then $b_0 = P(x_0)$ and $p(x) = (x - x_0)Q(x) + b_0$ where $Q(x) = b_n x^{n-1} + b_{n-1}x^{n-2} + \ldots + b_2 x + b_1$

Example 2.6.1: Use three iterations of Newton and Horner's Method (synthetic division) with four-digit rounding to approximate the root closest to $x_0 = 1$ for $P(x) = \pi x^4 - \frac{2}{3}x^2 + 5x - \sqrt{3}$.



Muller's Method helps finds roots that are complex. Complex roots come in conjugate pairs. In other words, if $z = a + bi$ is a root of a polynomial $P(x)$ with real coefficients, then $\bar{z} = a - bi$ is also a root of the polynomial.

Consider the polynomial $P(x) = a(x - p_2)^2 + b(x - p_2) + c$ that passes through the points $(p_0, f(p_0))$, $(p_1, f(p_1))$, and $(p_2, f(p_2))$. The constants a, b, and c can be found using the following conditions:

$$\begin{aligned} f(p_0) &= a(p_0 - p_2)^2 + b(p_0 - p_2) + c, \\ f(p_1) &= a(p_1 - p_2)^2 + b(p_1 - p_2) + c, \\ f(p_2) &= a(p_2 - p_2)^2 + b(p_2 - p_2) + c = c, \end{aligned}$$

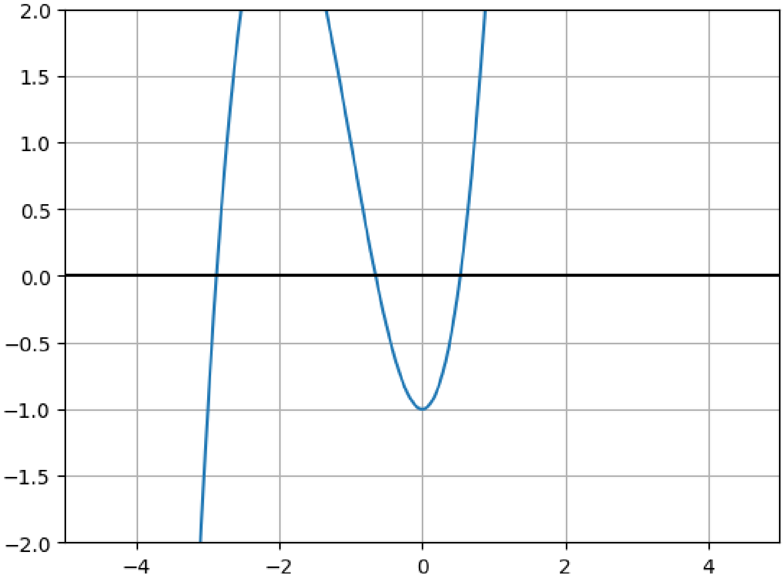
This leads to the formulas:

$$
\begin{aligned}
c &= f(p_2) \\
b &= \frac{(p_0 - p_2)^2 \left[f(p_1) - f(p_2)\right] - (p_1 - p_2)^2 \left[f(p_0) - f(p_2)\right]}{(p_0 - p_2)(p_1 - p_2)(p_0 - p_1)} \\
a &= \frac{(p_1 - p_2) \left[f(p_0) - f(p_2)\right] - (p_0 - p_2) \left[f(p_1) - f(p_2)\right]}{(p_0 - p_2)(p_1 - p_2)(p_0 - p_1)}
\end{aligned}
$$

Use the adjusted quadratic formula to compute p_3.

$$p_3 = p_2 - \frac{2c}{b + sgn(b)\sqrt{b^2 - 4ac}}$$

Example 2.6.2: Perform two iterations of Muller's Method on $f(x) = x^3 + 3x^2 - 1$ given $p_0 = 2$, $p_1 = 1$, and $p_2 = 0.5$.


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−4
−2
0
2
4