Homework 5 (root finding)

name

October 17, 2024

1. Based on a steady-state calculation, you know that the production feed P from a reactor depends on an input feed F as follows:

$$P(F) = 5F^2 + 2F$$

On the other hand, this input feed comes from a different stead-state calculation, and you know that it is a positive root of the function

$$Q(F) = -2F^2 + F + 8.$$

a) Show that Q has a positive root by applying the IVT to an interval of the form $[0, M]$

b) Using error propagation, estimate how the error for $P(F)$ depends on the error in F. Your answer will be a function of both F and σ_F.

c) Combine your M from part (a) with your answer to (b) to bound the error in terms of M and σ_F only.

d) Determine how many steps of the bisection method, applied to your interval $[0, M]$, are necessary to obtain an error less than 10^{-8} in $P(F)$.

e) Explain why the calculation in (c) was necessary for (d). Propose a way to sharpen the bound (c) at each step of the bisection method. Hint: the bound depends on M; can you improve it along the way? Even better, could you improve it by changing the left endpoint of the starting interval?

f) Determine the better error bound for the estimated $P(F)$ based on your proposal in (e). It should depend only on M, the step number k, and the approximate root F_k.

g) Explain why using the error estimate (f) is almost certainly not worth the extra complexity.

Solution.

(a)

Proof. Suppose $M = 3$, applying IVT on the interval $[0, M]$, we need to compute Q at endpoint:

$$Q(0) = -2(0)^2 + 0 + 8 > 0,\ Q(3) = -2(3)^2 + 3 + 8 = -7 < 0$$

Since Q is continuous and changes sign over $[0, 3]$, by IVT, there exists a positive root F in $[0, 3]$. □

(b)

Proof. The error $P(F)$ (i.e. δ_p) due to an error δ_f in F can be estimated using error propagation in form of:

$$\delta_P = |P'(F)| \cdot \delta_F$$

Compute the derivative:

$$P(F) = 5F^2 + 2F \Longrightarrow P'(F) = 10F + 2$$

Thus, the error in $P(F)$ is:

$$\delta_P = |10F + 2| \cdot \delta_F$$

□

(c)

Proof. From (a), $F \in [0, 3]$, so $F \leq M = 3$. The maximum value of $|P'(F)|$ over this interval is:

$$|P'(F)| \leq 10M + 2 = 10 \times 3 + 2 = 32$$

Therefore the error in $P(F)$ is bounded by:

$$\delta_P \leq (10M + 2) \cdot \delta_F = (32) \cdot \delta_F$$

□

(d)

Proof. In bisection method, after k steps, the error in F is:

$$\delta_F = \frac{M}{2^k}$$

Using the error bound from (c):

$$\delta_P \leq (10M + 2) \cdot \frac{M}{2^k}$$

We need $\delta_P \leq 10^{-8}$, so:

$$(10M + 2) \cdot \frac{M}{2^k} \leq 10^{-8}$$

Subbing $M = 3$ in:

$$32 \cdot \frac{3}{2^k} \leq 10^{-8} \Longrightarrow \frac{96}{2^k} \leq 10^{-8}$$

Solving for k, approximately:

$$k \geq \frac{13\log(2) + \log(3) + 8\log(5)}{\log(2)} \approx 33.1604$$

Therefore, at least 34 steps are needed □

(e)

The calculation in (c) provides a worst-case error bound based on the maximum possible value of $|P'(F)|$ over the entire interval $[0, M]$. In the bisection method, as k (i.e. iteration) increase, interval $[a_k, b_k]$ shrinks, and so does the maximum value of F within interval. Hence, we can sharpen the error bound by updating $M_k = b_k$ at each step.
Additionally we can improve the init. interval with better pick of left end point. Noticing that $Q(0) > 0$ and $Q(1.5) > 0$, wth $Q(2) < 0$, starting with the sharpened interval $[1.5, 2]$ shall reduce the error bound.

(f)

Proof. From (e), we pick interval $[1.5, 2]$, (NOTE: M here refer to the initial interval's width):

$$\delta_P = |P'(F_k)| \cdot \frac{M}{2^k} = |10F_k + 2| \cdot \frac{M}{2^k}$$

$$\leq 22 \cdot \frac{2 - 1.5}{2^k} = \frac{1}{2^k}$$

This provide a tighter estimate of the error in $P(F)$ comparing to (d) □

(g)

While the improved error estimate from part (f) offers a more precise bound on the error in $P(F)$, the added complexity—requiring extra computations like evaluating $|P'(F_k)|$ at each step—yields minimal benefit. Since the bisection method already halves the error at each iteration, the original error estimate is typically sufficient, and the marginal gains from tighter bounds do not justify the increased computational effort or implementation complexity.

2. Write a small computer program implementing the bisection method. Use it to estimate a root of

$$x^3 - 4x^2 - 8x + 30$$

on the interval $[2, 4]$ to an accuracy of 10^{-6}. Your program should produce a fraction of the form $\frac{N}{2^k}$. You do not need to include the program with your homework. Just report:

a) The number of steps you used, with justification.

b) The fraction $\frac{N}{2^k}$.

Solution.

Running Result:

Number of steps: 20
Root as a fraction: $2694241/2^{20}$
Approximate root: 2.5694284439086914

(a)

We want an accuracy of 10^{-6}, so we find the number of steps such that the interval length:

$$\frac{b-a}{2^n} \leq 10^{-6}$$

Starting with the interval length $b - a = 2$:

$$\frac{b-a}{2^n} \leq 10^{-6} \Rightarrow 2^{-n} \leq 5 \times 10^{-7} \Rightarrow n \geq \log_2\left(\frac{2}{10^{-6}}\right) \approx 20.93$$

So that we need around $n = 21$ steps to achieve the desired accuracy.

(b)

The root found of fraction form is in terms of:

$$\frac{N}{2^k} = \frac{2694241}{2^{20}}$$

3. Determine a polynomial $f(x)$ of degree 3 such that Newton's method applied to it with an initial guess of $x = 2$ fails to converge because the iterates alternates between 2 and another value.
Hint: draw a picture.
Hint: there are a lot of free parameters – you can pick the other value x in the NM sequence, as well as $f(2)$ and $f(x)$, after which Newton's method forces two constraints on f' – what is it, and why? In total, you will have a system of four linear equations in four variables to solve.

Solution.

Proof.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

We are looking for a scenario where:

- Starting from $x_0 = 2$, we get $x_1 = a$.
- Starting from $x_1 = a$, we get $x_2 = 2$.

This means the sequence alternates between $x = 2$ and $x = a$.
Referencing Newton's method formula, starting at $x = 2$:

$$x_1 = 2 - \frac{f(2)}{f'(2)} = a \Rightarrow \frac{f(2)}{f'(2)} = 2 - a$$

Starting at $x = a$:

$$x_2 = a - \frac{f(a)}{f'(a)} = 2 \Rightarrow \frac{f(a)}{f'(a)} = a - 2$$

Note that:

$$\frac{f(2)}{f'(2)} = -\left(\frac{f(a)}{f'(a)}\right)$$

For the sake of simplicity, let's choose $a = -2$, we get:

$$x_1 = 2 - \frac{f(2)}{f'(2)} = -2 \Rightarrow \frac{f(2)}{f'(2)} = 2 - (-2) = 4$$

Starting at $x = a$:

$$x_2 = -2 - \frac{f(-2)}{f'(-2)} = 2 \Rightarrow \frac{f(-2)}{f'(-2)} = -2 - 2 = -4$$

These conditions imply:

$$f(2) = 4f'(2);\ f(-2) = -4f'(-2)$$

Now let $f(x) = px^3 + qx^2 + rx + s$, it derivative is: $f'(x) = 3px^2 + 2qx + r$
Then, compute needed function values, $f(2), f'(2), f(-2)$, and $f'(-2)$:

1. $f(2) = 8p + 4q + 2r + s$
2. $f'(2) = 12p + 4q + r$
3. $f(-2) = -8p + 4q - 2r + s$
4. $f'(2) = 12p - 4q + r$

From our previously concluded conditions:

1. $f(2) = 4f'(2)$:

$$\begin{aligned} 8p + 4q + 2r + s &= 4(12p + 4q + r) \\ \Rightarrow 8p + 4q + 2r + s &= 48p + 16q + 4r \\ \Rightarrow -40p - 12q - 2r + s &= 0 \ (1) \end{aligned}$$

2. $f(-2) = 4f'(-2)$:

$$-8p + 4q - 2r + s = -4(12p - 4q + r)$$
$$\Rightarrow -8p + 4q - 2r + s = -48p + 16q - 4r$$
$$\Rightarrow 40p - 12q + 2r + s = 0 \ (2)$$

Since we have 4 vars need to solve, we want to get four euqations ideally.
So, we add euqaiton 1 and 2 first:

$$(-40p + 40p) + (-12q - 12q) + (-2r + 2r) + (s + s) = 0$$
$$\Rightarrow -24q + 2s = 0 \Longrightarrow s = 12q \ (3)$$

So, we substract euqaiton 1 and 2:

$$(40p + 40p) + (-12q + 12q) + (2r + 2r) + (s - s) = 0$$
$$\Rightarrow 80p + 4r = 0 \Longrightarrow r = -20p \ (4)$$

For double verification, subsitute r and s into Equaion 1:

$$-40p - 12q - 2(-20p) + (12q) = 0$$

Simplify

$$-40p - 12q + 40p + (12q) = 0 \Longrightarrow 0 = 0$$

This confirms that our subsitute are consistent.
Lastly, forming up the polynomial for the 4 equations we have, we can get:

$$f(x) = px^3 + qx^2 + (-20p)x + (12q)$$

Factor out p and q:

$$f(x) = p(x^3 - 20x) + q(x^2 + 12)$$

We can pick arbitrary p and q values here WLOG, so we let $p = 1$, and $q = 0$:

$$f(x) = x^3 - 20x$$

□

4. Generalize (3) as follows. Let $f(x) = ax^3 + bx^2 + cx + d$ be a cubic polynomial and $x \neq y$ two points.

1. Given two points $(x, f(x))$ and $(y, f(y))$, determine the NM constraints on $f'(x)$ and $f'(y)$ that ensure the NM sequence is $x, y, x, y, ...$
2. Observe that you have 4 linear constraints on f. Summarize them in matrix form.
3. Using a computer algebra system, calculate the determinant (as a polynomial in x and y. Then, show that it is nonzero.

Solution.

5. Using Newton's method, determine an efficient and accurate method for estimating the unique positive 6th root of a positive real number.

Solution.

6. [hard problem] Assume f is continuously twice differentiable, that $f'' > C > 0$, and that our sequence converges to some x which is a double root of f, so both $f(x)$ and $f'(x)$ are zero. Establish an error bound of the form

$$|e_n| \leq (constant)|e_{n-1}|.$$

Hint: we saw an example of this with $f(x) = x^2$. Hint: at Eqn 2 pg 84, develop the denominator by using the mean value theorem on f' on $[x_n, x]$, i.e. write the difference quotion as

$$\frac{f'(x_n) - f'(x)}{x_n - x} = f''(*)$$

(what is $*$?). Simplify and rearrange, then combine with the application of the MVT to the numerator from the error analysis in class, further simplify, and conclude.

Solution.

7. Suppose you're in a Question 6 situation:

1. If the Q6 constant is 0.8, would you rather use Newton's method or bisection? Justify your answer.
2. If the Q6 constant is 0.5, would you rather use Newton's method or bisection? Justify your answer.
3. If the Q6 constant is 0.0001, would you rather use Newton's method or bisection? Justify your answer.
4. Explain why the choice of Newton vs bisection might not be clear-cut if the Q6 constant is something like 0.45. What factors might you consider to make your decision?

Solution.