Homework 5 (root finding)

Hanzhang Yin

October 25, 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:

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

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)

In (c), we calculated how errors in F affect $P(F)$, establishing a bound necessary in (d) to determine the number of bisection steps required to achieve the desired precision in $P(F)$; without this calculation, we couldn't link the error in F to the error in $P(F)$. To sharpen the bound from (c) at each bisection step, we can update M to be the current upper bound of F, reducing the maximum possible error in $P(F)$ as the interval narrows. Furthermore, we can improve the bound by starting with a higher left endpoint—changing it from 0 to a larger value where $Q(F) > 0$ (like $L = 2$)—which decreases the interval length and the maximum value of F, resulting in a smaller M and a tighter error bound, ultimately reducing the number of steps needed in the bisection method.

(f)

Proof. From (e), at step k:

- Current interval: $[a_k, b_k]$
- Error in F:

$$\delta_F = \frac{b_k - a_k}{2} = \frac{M}{2}$$

- Current approx: $F_k = \frac{b_k + a_k}{2}$

Hence, the improved Error Bound will be:

$$\delta_P = |10F_k + 2|\delta_F = |10F_k + 2| \left(\frac{M}{2^{k+1}}\right)$$

□

(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: 21
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.

(1)

Proof. From (3) , for Newton's method to alternate between x and y, the iterations must satisfy:

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

Rewriting the equations, we get:

$$f'(x)(x - y) = f(x)$$
$$f'(y)(y - x) = f(y)$$

These are the NM constraints on $f'(x)$ and $f'(y)$. □

(2)

Proof. Let $f(x)$ be a cubic polynomial:

$$f(x) = ax^3 + bx^2 + cx + d$$

Assume the given cubic polynomial in general form is continuous and differentiable, compute $f(x), f(y), f'(x)$, and $f'(y)$:

$$\begin{aligned} f(x) &= ax^3 + bx^2 + cx + d \\ f(y) &= ay^3 + by^2 + cy + d \\ f'(x) &= 3ax^2 + 2bx + c \\ f'(y) &= 3ay^2 + 2by + c \end{aligned}$$

Applying the Newton's Method Constraints, we can set up the linear equations: Subsituting the expressions for $f(x), f(y), f'(x)$, and $f'(y)$:

1. $[3ax^2 + 2bx + c](x - y) = ax^3 + bx^2 + cx + d$
2. $[3ay^2 + 2by + c](y - x) = ay^3 + by^2 + cy + d$

Assuming $f(x)$ and $f(y)$ take specific values (which we can choose), but for generality, we keep them as itself.
Now, we can write the system of Equations:

1. $f(x) = ax^3 + bx^2 + cx + d$
2. $f(y) = ay^3 + by^2 + cy + d$
3. $[3ax^2 + 2bx + c](x - y) = ax^3 + bx^2 + cx + d$
4. $[3ay^2 + 2by + c](y - x) = ay^3 + by^2 + cy + d$

Now we can simplify equation 3 and 4:
For equation 3:

$$\begin{aligned} &[3ax^2 + 2bx + c](x - y) - (ax^3 + bx^2 + cx + d) = 0 \\ \Rightarrow\ &[3ax^2(x - y) + 2bx(x - y) + c(x - y)] - (ax^3 + bx^2 + cx + d) = 0 \\ \Rightarrow\ &[3ax^3 - 3ax^2y + 2bx^2 - 2bxy + cx - cy] - (ax^3 + bx^2 + cx + d) = 0 \\ \Rightarrow\ &2ax^3 - 3ax^2y + bx^2 - 2bxy - cy - d = 0 \end{aligned}$$

For equation 4:

$$\begin{aligned} &[3ay^2 + 2by + c](y - x) - (ay^3 + by^2 + cy + d) = 0 \\ \Rightarrow\ &[3ay^3 - 3ay^2x + 2by^2 - 2by^2 + cy - cx] - (ay^3 + by^2 + cy + d) = 0 \\ \Rightarrow\ &2ay^3 - 3ay^2x + by^2 - 2byx - cx - d = 0 \end{aligned}$$

Now we can write the 4 linear constraint equations we have into matrix representation:

$$\mathbf{A} = \begin{pmatrix} x^3 & x^2 & x & 1 \\ y^3 & y^2 & y & 1 \\ 2x^3 - 3x^2y & x^2 - 2xy & -y & -1 \\ 2y^3 - 3y^2x & y^2 - 2yx & -x & -1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} f(x) \\ f(y) \\ 0 \\ 0 \end{pmatrix}$$

□

(3)

Proof. To ensure a unique solution, $\det(A)$ must be non-zero.

$$\det(\mathbf{A}) = \begin{vmatrix} x^3 & x^2 & x & 1 \\ y^3 & y^2 & y & 1 \\ 2x^3 - 3x^2y & x^2 - 2xy & -y & -1 \\ 2y^3 - 3y^2x & y^2 - 2yx & -x & -1 \end{vmatrix} = (x-y)^6$$

Since $x \neq y$. $\det(\mathbf{A}) = (x-y)^6 \neq 0$. This nonzero determinant confirms a unique solution exists for $f(x)$. □

Therefore, by solving this system, we obtain a cubic polynomial $f(x)$ where Newton's method alternates between x and y, failing to converge as desired

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

Solution.

Proof. First, lets define the function and find its derivative:

$$f(x) = x^6 - a;\ f'(x) = 6x^5$$

Applying Newton's iteration formula on this function:

$$x_{n+1} = \frac{5}{6}x_n + \frac{a}{6x_n^5}$$

For our initial guess, we can use $x_0 = e^{\frac{\ln a}{6}}$ or $x_0 = a^{1/6}$, depending on a.
Then we iterate through the Newton's formula until it converges.
I have tried for $a = 64$, start with $x_0 \approx 2$. The iteration converges quickly, as $x_1 = 2$. Thus, the method efficiently estimates the 6th root. □

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.

Proof. To start with, Newton Method's updates are given by:

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

Define error at iteration n as:

$$e_n = x_n - x$$

Then,

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

Our goal is to express e_n in terms of e_{n-1}.
Since f is twice continuously differentiable, we can expand f and f' around x.
By Taylor's theorem:

$$f(x_{n-1}) = f(x + e_{n-1}) = f(x) + f'(x)e_{n-1} + \frac{1}{2}f''(x)e_{n-1}^2 + R_1$$

where the remainder R_1 satisties:

$$R_1 = \frac{1}{6}f'''(\xi)e_{n-1}^3$$

for some ξ between x and x_{n-1}. Since $f(x) = f'(x) = 0$:

$$f(x_{n-1}) = \frac{1}{2}f''(x)e_{n-1}^2 + \frac{1}{6}f'''(\xi)e_{n-1}^3$$

Similarly,

$$f'(x_{n-1}) = f'(x + e_{n-1}) = f'(x) + f''(x)e_{n-1} + R_2$$

Where the remainder R_2 satisfies:

$$R_2 = \frac{1}{2}f'''(\eta)e_{n-1}^2$$

for some η between x and x_{n-1}. Since $f'(x) = 0$:

$$f'(x_{n-1}) = f''(x)e_{n-1} + \frac{1}{2}f'''(\eta)e_{n-1}^2$$

Using MVT on f' over $[x_{n-1}, x]$:

$$\frac{f'(x_{n-1}) - f'(x)}{x_{n-1} - x} = f''(\zeta)e_{n-1}$$

for some ζ between x_{n-1} and x.
Since $f'(x) = 0$ and $e_{n-1} = x_{n-1} - x$

$$f'(x_{n-1}) = f''(\zeta)e_{n-1}$$

Because $f''(t) \geq C > 0$ for all t in the interval, we have:

$$f''(\zeta) \geq C > 0$$

Then, subsituting the previous result into Newton's Method Error Expression, we have:

$$e_n = e_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$$

Subsitute the expressions for $f(x_{n-1})$ and $f'(x_{n-1})$:

$$e_n = e_{n-1} - \frac{f''(x)e_{n-1}^2 + \frac{1}{2}f'''(\xi)e_{n-1}^3}{f''(\zeta)e_{n-1}}$$

$$\Rightarrow e_n = e_{n-1} - \frac{f''(x)e_{n-1} + \frac{1}{2}f'''(\xi)e_{n-1}^2}{f''(\zeta)}$$

For simplication, let's denote:

$$\alpha = \frac{f''(x)}{2f''(\zeta)},\ \beta = \frac{f'''(\xi)}{2f''(\zeta)}$$

Then,

$$e_n = e_{n-1} - \alpha e_{n-1} - \beta e_{n-1} = (1 - \alpha)e_{n-1} - \beta e_{n-1}$$

Since $\beta e_{n-1} = O(e_{n-1}^2)$, it becomes negligible compared to e_{n-1} as n increases.
Then we can bound this expression on α only, since $f''(x) \geq C > 0$ and $f''(\zeta) \leq f''_{max}$ due to continuity, we have:

$$\alpha = \frac{f''(x)}{2f''(\zeta)} \leq \frac{f''(x)}{2C} = k_0 < \frac{1}{2}$$

Similarly, since $f''(\zeta) \to f''(x)$ as $e_{n-1} \to 0$, we can make α arbitrary close to $\frac{1}{2}$ for large n.
Lastly, for large n, we can write:

$$|e_n| \leq |1 - \alpha||e_{n-1}| + |\beta|e_{n-1}$$

Since $\alpha \approx \frac{1}{2}$, we have $|1 - \alpha| \approx \frac{1}{2}$ Let $k = |1 - \alpha| + |\beta|$, For large n, $|\beta|$ is small, so k can be made less than 1. Therefore,

$$|e_n| \leq k|e_{n-1}$$

The convergence is linear in this case, and errors decrease by at least a constant factor less than 1 for sufficiently large n. □

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.

1. **Reasoning:** If the constant (denote k) $k = 0.8$, this represents that the error at each iteration reduces by a factor of 0.8 for Newton's method (i.e., $|e_n| \leq 0.8|e_{n-1}$). The *bisection method* reduces the error by a factor of 0.5 at each iteration, as it halves the interval containing the root.
   In comparison, Newton's Method reduce the error by 20% per iteration, while Bisection Method reduce the error by 50% per iteration. Therefore, I will prefer the bisection method in this case for faster convergence.

2. **Reasoning:** If the constant (denote k) $k = 0.5$, this represents that the error at each iteration reduces by a factor of 0.5 for Newton's method (i.e., $|e_n| \leq 0.5|e_{n-1}$). The *bisection method* reduces the error by a factor of 0.5 at each iteration, as it halves the interval containing the root.
   In comparison, Newton's Method reduce the error by 50% per iteration, which is the same comparing to Bisection Metho. Theoretically, we can pick both methods in the case, but due to the simplicity and lower computation cost (No derivative calculations needed) that Bisection Metho possess per iteration, I prefer picking Bisection Method.

3. **Reasoning:** If the constant (denote k) $k = 0.0001$, this represents that the error at each iteration reduces by a factor of 0.0001 for Newton's method (i.e., $|e_n| \leq 0.0001|e_{n-1}$). The *bisection method* reduces the error by a factor of 0.5 at each iteration, as it halves the interval containing the root.
   In comparison, Newton's Method reduce the error by 99.99% per iteration, while Bisection Method reduce the error by 50% per iteration. Therefore, I will prefer the Newton's Method in this case for faster convergence.

4. **Reasoning:** When the Q6 constant is around 0.45, indicating that Newton's method reduces the error by approximately 55% per iteration compared to bisection's 50%, the choice between the two methods isn't clear-cut. The marginal gain in convergence speed with Newton's method may not justify its higher computational cost per iteration due to derivative evaluations and potential instability near multiple roots. Consider factors like the ease of computing derivatives, the function's behavior near the root, robustness and guaranteed convergence (offered by bisection), implementation complexity, and available resources to decide if Newton's slight efficiency outweighs bisection's simplicity and reliability