CS/MATH 3414 — Numerical Methods

Lecture 3
Numerically Solving Equations

Virginia Tech Department of Computer Science, Fall 2026
Instructor: Sumeet Khatri
Date: September 8, 2026

Learning Objectives

- Understand and numerically implement the bisection, fixed-point, Newton, and secant methods for finding roots of equations.

Primary References

[1] T. A. Driscoll and R. J. Braun, *Fundamentals of Numerical Computation*, Includes Python, Julia, and MATLAB examples (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2022), https://fncbook.com/.

[2] T. Sauer, *Numerical Analysis*, Third (Pearson, 2018), https://www.pearson.com/en-ca/subject-catalog/p/numerical-analysis/P200000006340/9780134696454.

[3] C. T. Kelley, *Solving Nonlinear Equations with Newton's Method*, Vol. 1, Fundamentals of Algorithms (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2003), https://doi.org/10.1137/1.9780898718898.

[4] R. P. Brent, *Algorithms for Minimization without Derivatives*, Reprinted by Dover Publications, 2002 (Prentice-Hall, Englewood Cliffs, NJ, 1973), https://maths-people.anu.edu.au/brent/pub/pub011.html.

[5] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Second (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2002), https://doi.org/10.1137/1.9780898718027.

[6] G. Strang and E. Herman, *Calculus Volume 1* (OpenStax, Houston, TX, 2016), https://openstax.org/details/books/calculus-volume-1.

[7] R. L. Burden, J. D. Faires, and A. M. Burden, *Numerical analysis*, 10th ed. (Cengage Learning, 2016), https://www.cengage.com/c/numerical-analysis-10e-burden/9781305253667/?filterBy=Student.

[8] E. W. Cheney and D. R. Kincaid, *Numerical mathematics and computing*, 7th ed. (Brooks/Cole, Cengage Learning, 2013), https://web.ma.utexas.edu/CNA/NMC7/.

Note: much of the text here comes directly from, or is a summarized form of, the text in the above two references.

1 Introduction

Many problems can be written as

$$f(x) = 0, \qquad x \in D \subseteq \mathbb{R}. \tag{1.1}$$

A solution α is a *root* or *zero* of f. An equation $u(x) = v(x)$ takes this form by setting $f(x) = u(x) - v(x)$, with the domain chosen so both sides are defined. For example, the function $f(x) = 6x^2 - 7x + 2$ has $\frac{1}{2}$ and $\frac{2}{3}$ as zeros, as can be verified by direct substitution or by writing f in its factored form as $f(x) = (2x-1)(3x-2)$. For another example, the function $g(x) = \cos(3x) - \cos(7x)$ as not only the obvious zero $x = 0$, but every integer multiple of $\frac{\pi}{5}$ and $\frac{\pi}{2}$ as well, which can be seen by applying the trigonometric identity $\cos(a) - \cos(b) = 2\sin\left(\frac{1}{2}(a+b)\right)\sin\left(\frac{1}{2}(b-a)\right)$. Consequently, we find that $g(x) = 2\sin(5x)\sin(2x)$.

1.1 Why is locating roots important?

Frequently, the solution to a scientific problem is a number about which we have little information other than that it satisfies some equation. Since every equation caneb written so that a function stands on one side and zero on the other, the desired number must be a zero of the function. Thus, if we possess an arsenal of methods for locating zeros of functions, then we will be tabl to solve such problems.

Consider the following specific engineering problem, whose solution is the root of an equation. In a certain electrical circuit, the voltage V and current I are related by two equations of the form

$$I = a(\mathrm{e}^{bV} - 1), \quad c = dI + V, \tag{1.2}$$

where a, b, c, d are constants. For our purpose, these four numbers are assumed to be known. When these equations are combined by eliminating I between them, the result is a single equation:

$$c = ad(\mathrm{e}^{bV} - 1) + V. \tag{1.3}$$

In a concrete case, this might be

$$12 = 14.3(\mathrm{e}^{2V} - 1) + V, \tag{1.4}$$

and its solution V is required. It turns out that $V \approx 0.299$ in this case.

In some problems in which a root of an equation is sought, we can perform the required calculation with a hand calculator. But how can we locate zeros of complicated functions such as these?

$$f(x) = 3.24x^2 - 2.42x^7 + 10.34x^6 + 11.01x^2 + 47.98, \tag{1.5}$$

$$g(x) = 2^{x^2} - 10x + 1, \tag{1.6}$$

$$h(x) = \cosh\left(\sqrt{x^2+1} - \mathrm{e}^x\right) + \log|\sin(x)|. \tag{1.7}$$

What is needed is a general *numerical method* that does not depend on special properties of our functions. Of course, continuity and differentiability are special properties, but they are common attributes of functions that are usually encountered. The sort of special property that we probably *cannot* easily exploit in general-purpose codes is typified by the trigonometric identity mentioned previously.

Example 1 (A recurring test problem)**.** Throughout the note we return to

$$f(x) = x^3 - x - 2. \tag{1.8}$$

Since $f(1) = -2$ and $f(2) = 4$, continuity guarantees a root in $(1, 2)$. Moreover, $f'(x) = 3x^2 - 1 \geq 2$ on $[1, 2]$, so there is exactly one root in that interval. A high-precision value for the root is

$$\alpha = 1.5213797068045675696040808322544385\ldots. \tag{1.9}$$

Example 2 (A physical constraint)**.** A sphere of radius R contains liquid to depth h, measured from its lowest point. The liquid volume is $V(h) = \pi h^2(R - h/3)$, with $0 \leq h \leq 2R$. Given a target volume V_*, the depth solves $f(h) = \pi h^2(R - h/3) - V_* = 0$.

2 The bisection method

The bisection method exploits a property of continuous functions called the *Intermediate-Value Theorem*; see Appendix A for a detailed background on calculus and a formal statement of the theorem. Informally, let f be a function that has values of opposite sign at the two ends of an interval. Suppose also that f is continuous on that interval. To fix the notation, let $a < b$ and $f(a)f(b) < 0$. It then follows that f has a root in the interval

(a, b). In other words, there must exist a number r that satisfies the two conditions $a < r < b$ and $f(r) = 0$. This is essentially the content of the Intermediate-Value Theorem; the function f must take on the value of zero somewhere in the interval, because f is continuous and because $f(a)$ and $f(b)$ have opposite signs (on account of the fact that $f(a)f(b) < 0$).

2.1 Description of the method

At each step in the algorithm, we have an interval $[a, b]$ and the values $u = f(a)$ and $v = f(b)$. The numbers u and v satisfy $uv < 0$ by design (i.e., a and b are chosen such that this is true). Next, we construct the midpoint of the interval, $c = \frac{1}{2}(a + b)$, and we compute $w = f(c)$. It can happen fortuitously that $f(c) = 0$. If so, the objective of the algorithm has been fulfilled. In the usual case, $w \neq 0$ and either $wu < 0$ or $wv < 0$ (why?). If $wu < 0$, we can be sure that a root of f exists in the interval $[a, c]$. Consequently, we store the value of c in b and w in v. If $wu > 0$, then we cannot be sure that f has a root in $[a, c]$, but since $wv < 0$, f must have a root in $[c, b]$. In this case, we store the value of c in a and w in u. In either case, the situation at the end of this step is just like that at the beginning, except that the final interval is *half* as large as the initial interval. This step can not be repeated until the interval is satisfactorily small, say $|b - a| < \frac{1}{2} \times 10^{-6}$. At the end, the best estimate of the root would be $\frac{1}{2}(a + b)$, where $[a, b]$ is the last interval in the procedure. We write this whole procedure as pseudocode below.

Algorithm: The Bisection Method

First evaluate $f(a_0)$ and $f(b_0)$ and return an endpoint if it is a root. Reject an interval without opposite endpoint signs. For $n = 0, 1, 2, \ldots$:

1. Set $c_n = (a_n + b_n)/2$ and evaluate $f(c_n)$.
2. If $f(c_n) = 0$, return c_n. If the half-width is within the requested error tolerance, return c_n and its bracket.
3. If $f(a_n)$ and $f(c_n)$ have opposite signs, set $[a_{n+1}, b_{n+1}] = [a_n, c_n]$; otherwise set $[a_{n+1}, b_{n+1}] = [c_n, b_n]$.

Store endpoint values so that only one new function value is needed per midpoint tested. In a program, also impose a finite iteration limit.

Example 3. For $f(x) = x^3 - x - 2$, the first rows are shown below. Every sign is taken from the displayed midpoint's function value; no knowledge of α is used to select the next interval.

n	a_n	b_n	c_n	$f(c_n)$	$(b_n - a_n)/2$
0	1.000000	2.000000	1.500000	-0.12500000	0.5000000
1	1.500000	2.000000	1.750000	+1.60937500	0.2500000
2	1.500000	1.750000	1.625000	+0.66601562	0.1250000
3	1.500000	1.625000	1.562500	+0.25219727	0.0625000
4	1.500000	1.562500	1.531250	+0.05911255	0.0312500
5	1.500000	1.531250	1.515625	-0.03405380	0.0156250
6	1.515625	1.531250	1.523438	+0.01225042	0.0078125

TABLE 1: Bisection on $[1, 2]$. The final column is a rigorous bound in exact arithmetic, not the actual error.

The midpoint 1.5 is already fairly close to the root, whereas the next midpoint 1.75 is farther away. This does not contradict convergence. See Figure 1 for a depiction of the intervals for the few iterations.

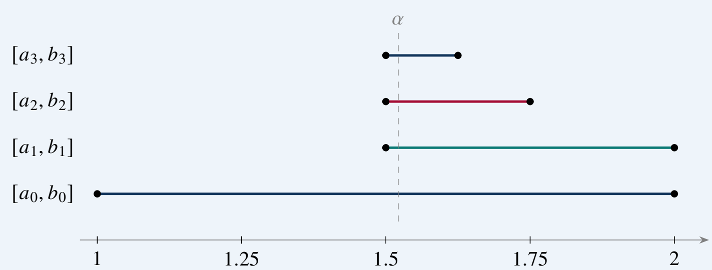


FIGURE 1: The first few iterations of the bisection method for $x^3 - x - 2 = 0$ on $[1, 2]$, as in Example 3. Each retained interval contains the root. The interval width, rather than the distance of the latest midpoint to the root, decreases by exactly a factor of two.

The bisection method is the simplest way to solve a non-linear equation $f(x) = 0$. It arrives at the root by constraining the interval in which a root lies, and it eventually makes the interval quite small. Because the bisection method halves the width of the interval at each step, one can predict exactly how long it takes to find the root within any desired degree of accuracy. In the bisection method, not every guess is closer to the root than the previous guess, because the bisection method does not use the nature of the function itself. Often the bisection method is used to get close to the root before switching to a faster method. Observe that root finding by the bisection method uses the same idea as in the **binary search** method taught in data structures.

2.2 Convergence

The bisection method is conceptually clear, but it exhibits rather slow convergence, being only linear — we will say more about that later. But, importantly, the method always converges to a solution, and for that reason it is often used as a starter for the more efficient methods we will see later.

Theorem 1 (Bisection error bound)**.** Let f be a continuous function, and assume that in the interval $[a_0, b_0]$ the function satisfies $f(a_0)f(b_0) < 0$. If bisection does not encounter an exact zero, the nested intervals converge to a common point α with $f(\alpha) = 0$. After n interval halvings,

$$b_n - a_n = \frac{b_0 - a_0}{2^n}, \qquad |c_n - \alpha| \le \frac{b_0 - a_0}{2^{n+1}}. \tag{2.1}$$

Proof. Assume that f is continuous on $[a_0, b_0]$, where $a_0 < b_0$, and that $f(a_0)f(b_0) < 0$. We consider the case in which bisection never encounters an exact zero. Thus, for every $n \ge 0$, the midpoint

$$c_n = \frac{a_n + b_n}{2} \tag{2.2}$$

satisfies $f(c_n) \ne 0$.

1. The intervals remain nested sign brackets, and their lengths tend to zero.

   Suppose that $f(a_n)f(b_n) < 0$. Since the endpoint values have opposite signs and $f(c_n) \ne 0$, exactly one of the products $f(a_n)f(c_n)$ and $f(c_n)f(b_n)$ is negative. The bisection update is therefore well defined:

$$(a_{n+1}, b_{n+1}) = \begin{cases} (a_n, c_n), & \text{if } f(a_n)f(c_n) < 0, \\ (c_n, b_n), & \text{if } f(c_n)f(b_n) < 0. \end{cases} \tag{2.3}$$

   In either case,

$$a_n \le a_{n+1} < b_{n+1} \le b_n, \qquad f(a_{n+1})f(b_{n+1}) < 0, \tag{2.4}$$

   and

$$b_{n+1} - a_{n+1} = \frac{b_n - a_n}{2}. \tag{2.5}$$

   Starting from the initial bracket and applying induction, these properties hold at every iteration. In particular,

$$[a_{n+1}, b_{n+1}] \subseteq [a_n, b_n], \qquad b_n - a_n = \frac{b_0 - a_0}{2^n} \quad \text{for every } n \ge 0. \tag{2.6}$$

   Continuity and the intermediate value theorem also guarantee that each interval $[a_n, b_n]$ contains at least one zero of f.

2. Both endpoint sequences converge to the same point.

   The nesting property shows that (a_n) is nondecreasing and bounded above by b_0, while (b_n) is nonincreasing and bounded below by a_0. By the monotone convergence theorem for real sequences, there exist $A, B \in [a_0, b_0]$ such that

$$\lim_{n\to\infty} a_n = A, \qquad \lim_{n\to\infty} b_n = B. \tag{2.7}$$

   Using the interval-length formula from Step 1 gives

$$B - A = \lim_{n\to\infty} (b_n - a_n) = \lim_{n\to\infty} \frac{b_0 - a_0}{2^n} = 0. \tag{2.8}$$

   Hence $A = B$. Denote their common value by α.

   To see that α belongs to every interval, fix $n \ge 0$. For every $m \ge n$, nesting gives

$$a_n \le a_m \le b_m \le b_n. \tag{2.9}$$

Letting $m \to \infty$ yields

$$a_n \le \alpha \le b_n. \tag{2.10}$$

Moreover, if another point β belonged to every interval, then

$$0 \le |\beta - \alpha| \le b_n - a_n = \frac{b_0 - a_0}{2^n} \quad \text{for every } n \ge 0. \tag{2.11}$$

Letting $n \to \infty$ would give $\beta = \alpha$. Thus the nested intervals have exactly one common point.

3. The common limit is a zero of f.

   Since $a_n \to \alpha$, $b_n \to \alpha$, and f is continuous at α,

$$f(a_n) \to f(\alpha), \qquad f(b_n) \to f(\alpha). \tag{2.12}$$

   The sign-bracket property gives $f(a_n)f(b_n) < 0$ for every n. Taking limits therefore yields

$$f(\alpha)^2 = \lim_{n\to\infty} f(a_n)f(b_n) \le 0. \tag{2.13}$$

   A real square is nonnegative, so $f(\alpha)^2 = 0$ and consequently $f(\alpha) = 0$.

4. The midpoint error satisfies the claimed bound.

   Because $a_n \le \alpha \le b_n$ and $c_n = (a_n + b_n)/2$, subtracting c_n gives

$$-\frac{b_n - a_n}{2} \le \alpha - c_n \le \frac{b_n - a_n}{2}. \tag{2.14}$$

   Consequently,

$$|c_n - \alpha| \le \frac{b_n - a_n}{2} = \frac{b_0 - a_0}{2^{n+1}}. \tag{2.15}$$

   The right-hand side tends to zero, so $c_n \to \alpha$, with the stated error bound after n interval halvings. ■

Remark 2. *Even if the initial interval contains several roots, the method converges to a root selected by its successive sign decisions. It does not identify all roots in the interval.*

To guarantee $|c_n - \alpha| \le \varepsilon$, it suffices to take

$$n \ge \max\left\{0, \left\lceil \log_2 \frac{b_0 - a_0}{2\varepsilon} \right\rceil\right\}. \tag{2.16}$$

Example 4 (How much work buys six decimal places?)**.** For $b_0 - a_0 = 1$ and absolute error tolerance 10^{-6}, $n \ge \lceil \log_2(500000) \rceil = 19$. The midpoint of the bracket after 19 halvings has error at most $2^{-20} \approx 9.537 \times 10^{-7}$. If every midpoint from c_0 through c_{19} is evaluated, this uses 20 midpoint evaluations in addition to the two initial endpoint evaluations. To guarantee absolute error 10^{-10}, 33 halvings suffice.

Linear convergence of the bisection method. Assume that bisection does not terminate at an exact zero, and let α be the limiting root. Define the actual midpoint error E_n and the guaranteed error bound B_n by

$$E_n = |c_n - \alpha|, \qquad B_n = \frac{b_n - a_n}{2}. \tag{2.17}$$

Theorem 1 gives

$$E_n \le B_n = \frac{b_0 - a_0}{2^{n+1}} = B_0 \left(\frac{1}{2}\right)^n. \tag{2.18}$$

Moreover,

$$\frac{B_{n+1}}{B_n} = \frac{1}{2} \qquad \text{for every } n \ge 0. \tag{2.19}$$

Thus the sequence of error bounds (B_n) converges to zero Q-linearly with asymptotic constant $1/2$, according to the ratio-based definition in Definition 4. Each additional interval halving reduces the guaranteed error bound by a factor of two.

The midpoint sequence (c_n) is therefore said to converge *R-linearly* to α: its errors are bounded by a geometrically decreasing sequence. More generally, an estimate

$$|x_n - \alpha| \le Cq^n, \qquad C > 0, \quad 0 < q < 1, \tag{2.20}$$

valid for all sufficiently large n, establishes R-linear convergence. For bisection, we may take $C = (b_0 - a_0)/2$ and $q = 1/2$.

The distinction between the error and its bound matters. The estimate above does not imply that $E_{n+1} \le E_n/2$, or that E_{n+1}/E_n tends to $1/2$. For example, apply bisection to $f(x) = x - 2/5$ on $[0, 1]$. The first two midpoints are $c_0 = 1/2$ and $c_1 = 1/4$, so

$$E_0 = \left|\frac{1}{2} - \frac{2}{5}\right| = \frac{1}{10}, \qquad E_1 = \left|\frac{1}{4} - \frac{2}{5}\right| = \frac{3}{20} > E_0. \tag{2.21}$$

Indeed, in this example the successive error ratios alternate between $3/2$ and $1/6$, and therefore have no limit. Nevertheless, the guaranteed error bound halves at every step, and the midpoint sequence converges R-linearly.

2.3 Strengths and limitations

Bisection requires no derivative and has a predictable worst-case cost. It is especially valuable when function evaluation is reliable but the graph is otherwise poorly understood. Its information use is modest: each new sign selects one half of the current interval, giving about one binary digit of location information per halving.

Its limitations are equally clear. A suitable bracket must be found; a sign scan can miss even-multiplicity roots or two sign crossings inside one scan cell; and fast local methods can gain many more digits per evaluation. A bracket found by expanding around an initial point must remain inside the function's domain and must not cross singularities.

3 The fixed-point method

For a non-linear equation $f(x) = 0$, we seek a point where the curve of f intersects the x-axis (i.e., $y = 0$). An alternative to the bisection approach to solving this problem is to recast the problem as a fixed-point problem of finding the point x at which $x = g(x)$ for a related non-linear function g. In other words, for the fixed-point problem, we seek a point where the curve g intersects the diagonal line $y = x$. A value of x such that $x = g(x)$ is called a **fixed point** of g, because x is unchanged when g is applied to it. Root-finding problems and fixed-point problems are equivalent in the following sense.

- Given the root-finding problem $f(x) = 0$, we can define functions g with a fixed point at x in a number of ways; for example, as
$$g(x) = x - f(x), \quad g(x) = x + 3f(x). \tag{3.1}$$
More generally, observe that for every $\lambda \in \mathbb{R}$, the function
$$g(x) = x - \lambda f(x) \tag{3.2}$$
has any root of f as its fixed point. As we will see, the Newton and secant methods will arise from particular choices of λ.
- Conversely, if the function g has a fixed point at α, then the function defined by $f(x) = x - g(x)$ has a zero at α.

Although the problems we sih to solve are in the root-finding form, th fixed-point form is easier to analyze, and certain fixed-point choices lead to very powerful root-finding techniques.

3.1 Fixed points of a function

Let us first become comfortable with the concept of fixed points of a function. We will also understand when a function even has a fixed point. Then, we will understand how the fixed points (if they exist) can be approximated to a specified accuracy in the next subsection.

Example 5. Determine any fixed points of the function $g(x) = x^2 - 2$.

Solution: A fixed point α should have the property that $\alpha = g(\alpha) = \alpha^2 - 2$, which implies that $0 = \alpha^2 - \alpha - 2 = (\alpha + 1)(\alpha - 2)$. A fixed point for g occurs precisely when the graph of $y = g(x)$ intersects the graph of $y = x$, so g has two fixed points, one at $\alpha = -1$ and the other at $\alpha = 2$.

The following theorem gives a sufficient condition for the existence and uniqueness of a fixed point.

Theorem 3 (Fixed point)**.**

1. If $g : \mathbb{R} \to \mathbb{R}$ is a continuous function in the interval $[a, b]$ and $g(x) \in [a, b]$ for all $x \in [a, b]$, then g has at least one fixed point in $[a, b]$.
2. If, in addition, $g'(x)$ exists on (a, b) and a positive constant $k < 1$ exists with $|g'(x)| \le k$ for all $x \in (a, b)$, then there is exactly one fixed point in $[a, b]$.

Proof.

1. If $g(a) = a$ or $g(b) = b$, then g has a fixed point at an endpoint. If not, then $g(a) > a$ and $g(b) < b$ by assumption. The function $h(x) = g(x) - x$ is continuous on $[a, b]$, with $h(a) = g(a) - a > 0$ and $h(b) = g(b) - b < 0$. The Intermediate Value Theorem implies that there exists $\alpha \in (a, b)$ for which $h(\alpha) = 0$. The number α is a fixed point for g, because $0 = h(\alpha) = g(\alpha) - \alpha$ implies $g(\alpha) = \alpha$.

2. Suppose, in addition, that $|g'(x)| \leq k < 1$ and that α and β are both fixed points in $[a, b]$. If $\alpha \neq \beta$, then the Mean Value Theorem implies that a number ξ exists between α and β, and hence in $[a, b]$, with

$$\frac{g(\alpha) - g(\beta)}{\alpha - \beta} = g'(\xi). \tag{3.3}$$

Thus,

$$|\alpha - \beta| = |g(\alpha) - g(\beta)| = |g'(\xi)|\,|\alpha - \beta| \leq k\,|\alpha - \beta| < |\alpha - \beta|\,, \tag{3.4}$$

which is a contradiction. This contradiction must come from the only supposition, $\alpha \neq \beta$. Hence, $\alpha = \beta$, and the fixed point in $[a, b]$ is unique. ■

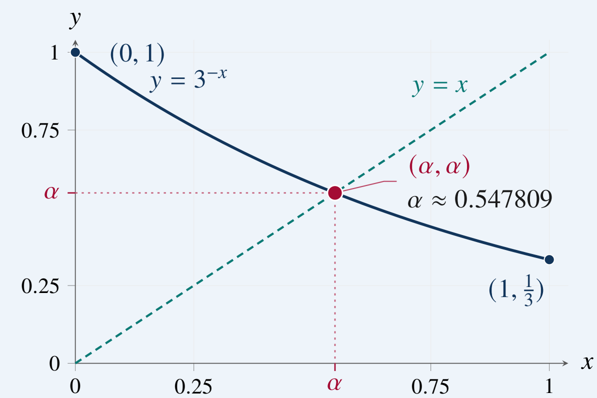
Example 6. Show that $g(x) = \frac{1}{3}(x^2 - 1)$ has a unique fixed point in the interval $[-1, 1]$.

Solution: The maximum and minimum values of $g(x)$, for $x \in [-1, 1]$, must occur either when x is an endpoint of the interval or when the derivative is zero. Since $g'(x) = \frac{2x}{3}$, the function g is continuous, and $g'(x)$ exists on $[-1, 1]$. The maximum and minimum values of $g(x)$ occur at $x = -1$, $x = 0$, or $x = 1$. But $g(-1) = 0$, $g(1) = 0$, and $g(0) = -\frac{1}{3}$, so an absolute maximum for $g(x)$ on $[-1, 1]$ occurs at $x = -1$ and $x = 1$ and an absolute minimum at $x = 0$. Moreover, $|g'(x)| = \left|\frac{2x}{3}\right| \leq \frac{2}{3}$ for all $x \in (-1, 1)$. So g satisfies all of the hypotheses of Theorem 3 and has a unique fixed point in $[-1, 1]$.

Remark: For this function, the unique fixed point α in the interval $[-1, 1]$ can be determined algebraically. If $\alpha = g(\alpha) = \frac{1}{3}(\alpha^2 - 1)$, then $\alpha^2 - 3\alpha - 1 = 0$, which can be readily solved by the quadratic formula to yield $\alpha = \frac{1}{2}(3 - \sqrt{13})$. We note also that g has a unique fixed point $\alpha = \frac{1}{2}(3 + \sqrt{13})$ for the interval $[3, 4]$. However, $g(4) = 5$ and $g'(4) = \frac{8}{3} > 1$, so g does not satisfy the hypotheses of Theorem 3 on $[3, 4]$. This demonstrates that the hypotheses of Theorem 3 are sufficient to guarantee a unique fixed point, but they are not necessary.

Example 7. Show that Theorem 3 does not ensure a unique fixed point of $f(x) = 3^{-x}$ on the interval $[0, 1]$, even though a unique fixed point does exist on this interval.

Solution: We have that $f'(x) = -3^{-x}\ln(3) < 0$ on $[0, 1]$, so the function f is strictly decreasing on $[0, 1]$. Therefore, $f(1) = \frac{1}{3} \leq f(x) \leq 1 = f(0)$ for all $x \in [0, 1]$. This means that for $x \in [0, 1]$, we have $f(x) \in [0, 1]$. The first part of Theorem 3 ensures that there is at least one fixed point in $[0, 1]$. However, $f'(0) = -\ln(3) = -1.098612289\ldots$, so $|f'(x)| \nleq 1$ on $(0, 1)$, and Theorem 3 cannot be used to determine uniqueness. But f is always decreasing, and it is clear from the diagram below that the fixed point must be unique.



3.2 Description of the method

In Example 6, we encounter a function $f(x) = 3^{-x}$ whose fixed point cannot be determined explicitly algebraically, because there is no way to algebraically solve for α in the equation $\alpha = g(\alpha) = 3^{-\alpha}$. We can, however, determine approximations to this fixed point to any specified degree of accuracy. We will now consider how this can be done.

To approximate the fixed point of a function g, we choose an initial approximation x_0 and generate the seqeuence $(x_n)_{n=0}^{\infty}$ in the following way: $f(x) = 0$ as $x = g(x)$, we can try

$$x_{n+1} = g(x_n), \qquad n = 0, 1, 2, \ldots. \tag{3.5}$$

If the sequence converges to α and g is continuous, then

$$\alpha = \lim_{n\to\infty} x_n = \lim_{n\to\infty} g(x_{n-1}) = g\big(\lim_{n\to\infty} x_{n-1}\big) = g(\alpha), \tag{3.6}$$

and thus we have a solution to the fixed-point problem. This technique is called **fixed-point iteration** or **functional iteration**.

Algorithm: Fixed-point iteration

Input: Initial approximation x_0; tolerance ε; maximum number of iterations, $n_{\max}$.
Output: Approximate solution α or message of failure.
For $n = 1, \ldots, n_{\max}$:

1. Set $x_n = g(x_{n-1})$.
2. If $|x_n - x_{n-1}| < \varepsilon$, then output x_{n+1}.

If $\left|x_{n_{\max}} - x_{n_{\max}-1}\right| \geq \varepsilon$, then output failure.

Example 8. Let us return to the example $f(x) = x^3 - x - 2$. Three possible functions whose fixed points give the required root of f are

$$g_1(x) = (x+2)^{1/3}, \qquad g_2(x) = x^3 - 2, \qquad g_3(x) = \sqrt{1 + 2/x} \quad (x > 0). \tag{3.7}$$

They have the same fixed point in $[1,2]$, but their iterations need not have the same behavior. Let us additional consider two other functions:

$$g_+(x) = x - \frac{1}{16} f(x), \quad g_-(x) = x - \frac{2}{7} f(x). \tag{3.8}$$

We plot these below to demonstrate how the fixed-point iterations converge to the root of f.

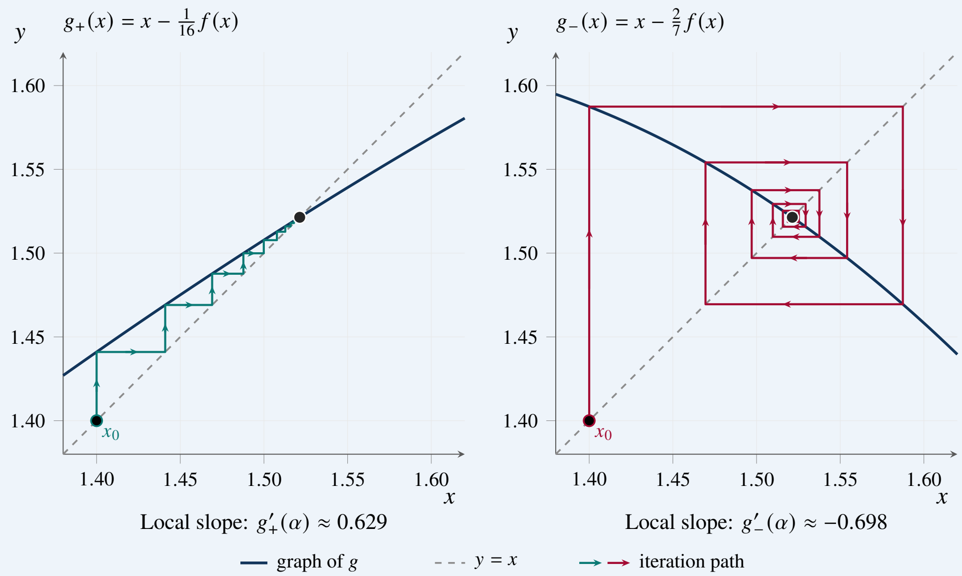


FIGURE 2: Two fixed-point iterations for $f(x) = x^3 - x - 2$, both starting from $x_0 = 1.4$. Each path starts at (x_0, x_0); a vertical segment reaches the graph of g, and a horizontal segment reaches $y = x$. The black dot marks the common fixed point (α, α), where $\alpha \approx 1.521379707$. For $g_+(x) = x - f(x)/16$, the iterates increase monotonically toward α. For $g_-(x) = x - 2f(x)/7$, the iterates alternate about α, forming an inward spiral. Both maps are contractions on $[1.4, 1.6]$. The first 12 and 14 updates, respectively, are shown on identical axis scales.

3.3 Convergence

Definition 1 (Contraction)**.** A map $g : I \longrightarrow I$ is a contraction on I if there is a constant $q < 1$, $q \geq 0$, such that

$$|g(x) - g(y)| \leq q|x - y| \quad \text{for all } x, y \in I. \tag{3.9}$$

The notation $g : I \longrightarrow I$ includes the *self-mapping condition*: $g(x)$ must remain in I whenever $x \in I$. If g is differentiable, the mean value theorem shows that $|g'| \leq q < 1$ is sufficient for the distance inequality, but it does not by itself prove self-mapping.

Theorem 4 (Contraction theorem on an interval)**.** Let $I = [a, b]$ be a nonempty closed interval, let $g(I) \subseteq I$, and suppose g is a contraction with constant $0 \leq q < 1$. Then:

1. g has exactly one fixed point $\alpha \in I$;

2. for every $x_0 \in I$, iteration (3.5) stays in I and converges to α;

3. for $n \geq 1$, the following bounds hold:

$$|x_n - \alpha| \leq q^n |x_0 - \alpha|, \tag{3.10}$$

$$|x_n - \alpha| \leq \frac{q^n}{1-q}|x_1 - x_0|, \tag{3.11}$$

$$|x_n - \alpha| \leq \frac{q}{1-q}|x_n - x_{n-1}|. \tag{3.12}$$

Proof. If $q = 0$, the contraction inequality implies that $g(x) = g(y)$ for all $x, y \in I$, so g is constant. Denote its constant value by α. The self-mapping condition gives $\alpha \in I$, and $g(\alpha) = \alpha$. This is the unique fixed point, and every iteration satisfies $x_n = \alpha$ for $n \geq 1$. All three error bounds then hold, with both sides equal to zero. For the remainder of the proof, assume that $0 < q < 1$.

1. The iteration stays in I, and successive differences decrease geometrically.

   Fix an arbitrary starting point $x_0 \in I$ and define

   $$x_{n+1} = g(x_n), \qquad n \geq 0. \tag{3.13}$$

   If $x_n \in I$, then $x_{n+1} = g(x_n) \in I$ because $g(I) \subseteq I$. Starting from $x_0 \in I$, induction therefore shows that every iterate is well defined and belongs to I.

   For $n \geq 0$, set

   $$d_n = |x_{n+1} - x_n|. \tag{3.14}$$

   For every $n \geq 1$, the contraction inequality gives

   $$d_n = |g(x_n) - g(x_{n-1})| \leq q|x_n - x_{n-1}| = q d_{n-1}. \tag{3.15}$$

   Applying this inequality repeatedly, or equivalently using induction, yields

   $$d_n \leq q^n d_0 = q^n |x_1 - x_0| \qquad \text{for every } n \geq 0. \tag{3.16}$$

2. The sequence converges to a point of I.

   For integers $m > n \geq 0$, write $x_m - x_n$ as a telescoping sum. The triangle inequality and the bound on d_j give

   $$\begin{aligned} |x_m - x_n| &= \left|\sum_{j=n}^{m-1}(x_{j+1} - x_j)\right| \\ &\leq \sum_{j=n}^{m-1}|x_{j+1} - x_j| \\ &\leq d_0 \sum_{j=n}^{m-1} q^j \\ &= d_0 \frac{q^n(1-q^{m-n})}{1-q} \\ &\leq \frac{q^n}{1-q} d_0. \end{aligned} \tag{3.17}$$

The final bound depends on n but not on m.

To verify the Cauchy property explicitly, let $\varepsilon > 0$. Since $q^N \to 0$, there is an integer $N \geq 0$ such that

$$\frac{q^N}{1-q} d_0 < \varepsilon. \tag{3.18}$$

For any $m, n \geq N$ with $m > n$, the preceding estimate gives

$$|x_m - x_n| \leq \frac{q^n}{1-q} d_0 \leq \frac{q^N}{1-q} d_0 < \varepsilon. \tag{3.19}$$

The same conclusion holds when $n > m$ by interchanging the indices, and it is immediate when $m = n$. Thus (x_n) is a Cauchy sequence.

By completeness of $\mathbb{R}$, there exists $\alpha \in \mathbb{R}$ such that $x_n \to \alpha$. Since $a \leq x_n \leq b$ for every n, passing to the limit gives $a \leq \alpha \leq b$. Consequently, $\alpha \in I$.

3. The limit is the unique fixed point.

   Because α and every x_n belong to I, we may apply the contraction inequality to obtain

$$\begin{aligned} |g(\alpha) - \alpha| &\leq |g(\alpha) - g(x_n)| + |g(x_n) - \alpha| \\ &\leq q|\alpha - x_n| + |x_{n+1} - \alpha|. \end{aligned} \tag{3.20}$$

   Both terms on the right tend to zero as $n \to \infty$. The nonnegative quantity on the left is independent of n, so it must be zero. Hence $g(\alpha) = \alpha$.

   Now suppose that $\beta \in I$ is another fixed point. Then

$$|\alpha - \beta| = |g(\alpha) - g(\beta)| \leq q|\alpha - \beta|, \tag{3.21}$$

   and therefore

$$(1-q)|\alpha - \beta| \leq 0. \tag{3.22}$$

   Since $1 - q > 0$, this forces $\alpha = \beta$. Thus the fixed point is unique.

   The starting point $x_0 \in I$ was arbitrary. The preceding argument therefore proves that every starting point in I produces a sequence converging to this same fixed point.

4. The three error bounds hold.

   First, since $g(\alpha) = \alpha$, for every $n \geq 1$,

$$|x_n - \alpha| = |g(x_{n-1}) - g(\alpha)| \leq q|x_{n-1} - \alpha|. \tag{3.23}$$

   Induction on n gives the first bound:

$$|x_n - \alpha| \leq q^n |x_0 - \alpha|. \tag{3.24}$$

   Next, fix $n \geq 1$ and let $m \to \infty$ in the telescoping-sum estimate established above. Since $x_m \to \alpha$ and the absolute value is continuous, we obtain the a priori bound

$$|x_n - \alpha| = \lim_{m\to\infty} |x_m - x_n| \leq \frac{q^n}{1-q}|x_1 - x_0|. \tag{3.25}$$

Finally, the contraction inequality and the triangle inequality give, for every $n \geq 1$,

$$\begin{aligned} |x_n - \alpha| &\leq q|x_{n-1} - \alpha| \\ &\leq q|x_{n-1} - x_n| + q|x_n - \alpha|. \end{aligned} \tag{3.26}$$

Subtracting the last term from both sides yields

$$(1-q)|x_n - \alpha| \leq q|x_n - x_{n-1}|. \tag{3.27}$$

Dividing by $1 - q > 0$ proves the a posteriori bound:

$$|x_n - \alpha| \leq \frac{q}{1-q}|x_n - x_{n-1}|. \tag{3.28}$$

All conclusions of the theorem follow. ■

An *a priori* bound uses information known before iterating. An *a posteriori* bound uses information obtained during the computation. There is also a defect bound valid for any $x \in I$:

$$|x - \alpha| \leq \frac{|g(x) - x|}{1-q}. \tag{3.29}$$

Indeed, $|x - \alpha| \leq |x - g(x)| + q|x - \alpha|$. If $q = 0$, g is constant and reaches its fixed point in one step.

Example 9. Let us return again to our running example $f(x) = x^3 - x - 2$. For $g_1(x) = (x+2)^{1/3}$ on $I = [1,2]$,

$$g_1(I) = [\sqrt[3]{3}, \sqrt[3]{4}] \subset [1,2], \qquad 0 < g_1'(x) = \frac{1}{3(x+2)^{2/3}} \leq q := \frac{1}{3 \cdot 3^{2/3}} \approx 0.16025. \tag{3.30}$$

Both hypotheses have been checked. Starting anywhere in $[1,2]$ converges to the unique fixed point and therefore to the root of the cubic in this interval. With $x_0 = 1$, the iterates are given in Table 2.

TABLE 2: Fixed-point iteration $x_{n+1} = (x_n + 2)^{1/3}$ from $x_0 = 1$. The last column uses the interval contraction constant, not the unknown error.

n	x_n	$\lvert x_n - \alpha\rvert$	$\frac{q}{1-q}\lvert x_n - x_{n-1}\rvert$
0	1.00000000000000	5.21×10^{-1}	–
1	1.44224957030741	7.91×10^{-2}	8.44×10^{-2}
2	1.50989744933236	1.15×10^{-2}	1.29×10^{-2}
3	1.51972430499164	1.66×10^{-3}	1.88×10^{-3}
4	1.52114126906273	2.38×10^{-4}	2.70×10^{-4}
5	1.52134536775190	3.43×10^{-5}	3.89×10^{-5}
6	1.52137476149789	4.95×10^{-6}	5.61×10^{-6}
7	1.52137899461282	7.12×10^{-7}	8.08×10^{-7}
8	1.52137960423926	1.03×10^{-7}	1.16×10^{-7}
12	1.52137970676045	4.41×10^{-11}	5.00×10^{-11}

For example, to guarantee an error at most ε using consecutive iterates, it suffices to check

$$|x_n - x_{n-1}| \le \frac{1-q}{q}\varepsilon \qquad (q > 0). \tag{3.31}$$

This is a proof-based use of the step size. The same test with the factor omitted has no such general guarantee.

3.4 Designing an iteration by relaxation

For a nonzero constant λ, the map

$$g_\lambda(x) = x - \lambda f(x) \tag{3.32}$$

has exactly the same fixed points as the roots of f. Its derivative is $g'_\lambda = 1 - \lambda f'$. If $0 < m \le f'(x) \le M$ on an interval, then

$$|g'_\lambda(x)| \le q(\lambda) := \max\{|1 - \lambda m|, |1 - \lambda M|\}. \tag{3.33}$$

For $0 < \lambda < 2/M$, this upper bound is less than one. To apply the interval contraction theorem, self-mapping must still be checked. Locally about a root, the derivative condition supplies a small invariant interval.

The choice that minimizes this derivative bound is obtained by balancing the endpoint errors, $1 - \lambda m = -(1 - \lambda M)$:

$$\lambda_* = \frac{2}{m + M}, \qquad q_* = \frac{M - m}{M + m}. \tag{3.34}$$

If $m \ll M$, even this best bound is close to one and convergence can be slow. Newton's method adapts the factor to the local derivative instead of using a single constant.

4 Newton's method

The procedure known as Newton's method is also called the **Newton–Raphson iteration**. It has a more general form than the one seen here, and the more general form can be used to find roots of systems of (non-linear) equations. It is one of the most important procedures in numerical analysis, and its applicability extends to differential equations and integral equations. Here, we will apply to a single equation of the form $f(x) = 0$. As before, we seek one or more points at which the value of the function f is zero.

4.1 Description of the method

In Newton's method, it is assumed at once that the function f is differentiable. This implies that the graph of f has a definite slope at each point and hence a unique tangent line.

In fact, Newton's method is nothing but a particular case of the fixed-point iteration method: for a given function f whose roots we wish to find, Newton's method takes the function $g(x) = x + \lambda f(x)$, with $\lambda = -\frac{1}{f'(x)}$, which is well defined as long as $f'(x) \neq 0$. So the iterates take the form

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}}, \quad n \in \{0, 1, \dots\}. \tag{4.1}$$

Algorithm: Newton's method

Input: A function f and its derivative f'; an initial approximation x_0; absolute tolerance $\varepsilon_{\text{abs}} > 0$; relative tolerance $\varepsilon_{\text{rel}} \geq 0$; residual tolerance $\varepsilon_f > 0$; maximum number of iterations $n_{\max} \geq 1$.
Output: An approximate root, or a message of failure.
Throughout the algorithm, terminate with failure if an arithmetic operation or function evaluation is undefined or produces a nonfinite value.

1. Evaluate and store $y_0 = f(x_0)$. If $y_0 = 0$, **return** x_0.
2. **For** $n = 0, \ldots, n_{\max} - 1$:
   (a) Evaluate $d_n = f'(x_n)$.
   (b) If $d_n = 0$, **return failure**: "The derivative is zero at the current approximation."
   (c) Compute the next approximation:
   $$x_{n+1} = x_n - \frac{y_n}{d_n}. \tag{4.2}$$
   (d) Evaluate and store $y_{n+1} = f(x_{n+1})$.
   (e) If $y_{n+1} = 0$, **return** x_{n+1}.
   (f) Set the step tolerance:
   $$\tau_{n+1} = \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \max\{|x_n|, |x_{n+1}|\}. \tag{4.3}$$
   If both $|x_{n+1} - x_n| \leq \tau_{n+1}$ and $|y_{n+1}| \leq \varepsilon_f$, **return** x_{n+1}.
   (g) If $x_{n+1} = x_n$, **return failure**: "The iteration has stagnated without satisfying the residual tolerance."
3. **Return failure**: "The maximum number of iterations has been reached."

Remark 5 (Absolute, relative, and residual tolerances)**.** *The Newton's method algorithm uses stopping criteria to decide when an approximation is sufficiently satisfactory. One practical choice requires both a small step and a small residual:*
$$|x_{n+1} - x_n| \leq \tau_{n+1} \qquad \textit{and} \qquad |f(x_{n+1})| \leq \varepsilon_f, \tag{4.4}$$
where
$$\tau_{n+1} = \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \max\{|x_n|, |x_{n+1}|\}. \tag{4.5}$$
The quantity τ_{n+1} is calculated during the iteration; it is not an additional parameter that the user must choose.

The three input tolerances have different roles:

1. ***Absolute tolerance.*** *The parameter $\varepsilon_{\text{abs}} > 0$ provides a fixed threshold for the change in x. This is particularly useful near zero, where a purely relative tolerance can demand unnecessarily small steps.*
2. ***Relative tolerance.*** *The parameter $\varepsilon_{\text{rel}} \geq 0$ allows the step tolerance to grow with the magnitude of the iterates. For example, a change of 10^{-3} is relatively small when x is approximately 1000, but substantial when x is approximately 10^{-3}.*

3. ***Residual tolerance.*** *The parameter $\varepsilon_f > 0$ specifies how closely the approximation must satisfy the equation $f(x) = 0$. The quantity $|f(x_{n+1})|$ is called the* absolute residual.

For example, suppose that $\varepsilon_{\text{abs}} = 10^{-8}$ and $\varepsilon_{\text{rel}} = 10^{-6}$. When the iterates are near 1000, the step tolerance is approximately 10^{-3}. When both iterates are sufficiently close to zero, the step tolerance is approximately 10^{-8}. Thus, the absolute term supplies a fixed baseline, and the relative term adjusts the tolerance to the scale of the iterates.

The step and residual tests provide different information. A small step indicates that successive approximations have stopped changing appreciably. This can occur because of numerical stagnation, so a small step alone does not establish that the equation is nearly satisfied. The residual test provides that additional check.

Separate step and residual tolerances are useful because x and $f(x)$ can have different numerical scales, and may even have different physical units. For instance, replacing f by $10^{12} f$ leaves both the Newton and secant iterates unchanged in exact arithmetic, but multiplies every absolute residual by 10^{12}. The residual tolerance should therefore reflect the scaling of the function.

Neither the step size nor the residual directly measures the unknown root error $|x_{n+1} - \alpha|$. Additional information can relate these quantities. For example, suppose that f is continuously differentiable on the interval joining an approximation x to a root α, and that $|f'(t)| \geq m > 0$ throughout that interval. The mean value theorem gives

$$|x - \alpha| \leq \frac{|f(x)|}{m}. \tag{4.6}$$

Under these assumptions, satisfying the residual tolerance guarantees a root error of at most ε_f / m.

Three tolerance parameters are therefore a practical refinement, rather than a requirement of either iteration formula. For a simpler presentation, set $\varepsilon_{\text{rel}} = 0$ and write $\varepsilon_x = \varepsilon_{\text{abs}}$. The stopping criteria then become

$$|x_{n+1} - x_n| \leq \varepsilon_x \qquad \text{and} \qquad |f(x_{n+1})| \leq \varepsilon_f. \tag{4.7}$$

For dimensionless, suitably scaled examples, one may further choose $\varepsilon_x = \varepsilon_f = \varepsilon$, giving a single input tolerance while retaining both tests.

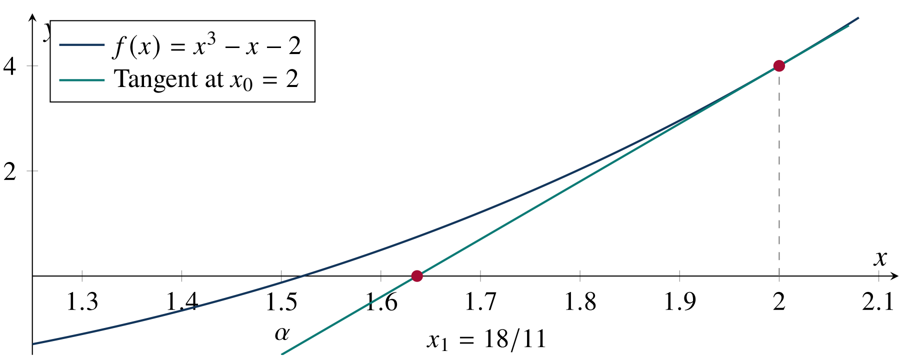


Figure 3: Newton's first step from $x_0 = 2$. The tangent intercept is closer to the root here, but a tangent intercept is not guaranteed to stay inside an initial interval in general.

Example 10. For our recurring example of $f(x) = x^3 - x - 2$, the Newton step is

$$x_{n+1} = x_n - \frac{x_n^3 - x_n - 2}{3x_n^2 - 1}, n \in \{0, 1, \dots\}. \tag{4.8}$$

Starting at $x_0 = 1.5$ gives $f(x_0) = -0.125$ and $f'(x_0) = 5.75$, so

$$x_1 = 1.5 + \frac{0.125}{5.75} = 1.5217391304347826\dots. \tag{4.9}$$

The next rows show the rapid gain in accuracy.

n	x_n	$\lvert f(x_n)\rvert$	$\lvert x_n - \alpha\rvert$
0	1.5000000000000000	1.25×10^{-1}	2.14×10^{-2}
1	1.5217391304347826	2.14×10^{-3}	3.59×10^{-4}
2	1.5213798059647862	5.89×10^{-7}	9.92×10^{-8}
3	1.5213797068045751	4.49×10^{-14}	7.55×10^{-15}
4	1.5213797068045676	2.60×10^{-28}	4.38×10^{-29}

TABLE 3: Newton's method for the cubic, starting at 1.5. Iterates were computed with high-precision arithmetic before rounding the displayed values.

Remember to not feed rounded table entries back into the recurrence. Doing so changes the computation, especially when very small errors are being illustrated.

4.2 Convergence

Theorem 6 (Convergence of Newton's method)**.** Let $f \in C^2$ on an interval $I = [\alpha - r, \alpha + r]$, where $f(\alpha) = 0$. Suppose

$$|f'(x)| \geq m > 0, \qquad |f''(x)| \leq M \quad \text{for every } x \in I. \tag{4.10}$$

Set $K = M/(2m)$. Choose $0 < \delta \leq r$ so that $K\delta \leq 1/2$. If $|x_0 - \alpha| \leq \delta$, Newton's method is well defined, stays within distance δ of α, and converges with

$$|e_{n+1}| \leq K|e_n|^2. \tag{4.11}$$

If $M = 0$, Newton reaches the root in at most one step.

Proof. Let

$$J = [\alpha - \delta, \alpha + \delta] \subseteq I, \tag{4.12}$$

and write $e_n = x_n - \alpha$ whenever the iterate x_n is defined. We prove the conclusions in the following steps.

1. A single Newton step satisfies a quadratic error estimate.

Suppose that $x_n \in I$ has been constructed. The hypothesis $|f'(x_n)| \geq m > 0$ guarantees that

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{4.13}$$

is well defined.

If $x_n \neq \alpha$, Taylor's theorem applied to $f(\alpha)$, with expansion point x_n, gives a point ξ_n between x_n and α such that

$$0 = f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + \frac{1}{2} f''(\xi_n)(\alpha - x_n)^2. \tag{4.14}$$

The segment joining x_n to α lies in I, so $\xi_n \in I$. Since $\alpha - x_n = -e_n$, rearranging gives

$$f(x_n) = f'(x_n) e_n - \frac{1}{2} f''(\xi_n) e_n^2. \tag{4.15}$$

Substituting this expression into the Newton update yields the exact error identity

$$\begin{aligned} e_{n+1} &= e_n - \frac{f(x_n)}{f'(x_n)} \\ &= e_n - \frac{f'(x_n)e_n - \frac{1}{2} f''(\xi_n) e_n^2}{f'(x_n)} \\ &= \frac{f''(\xi_n)}{2f'(x_n)} e_n^2. \end{aligned} \tag{4.16}$$

If $x_n = \alpha$, then $f(x_n) = 0$ and the Newton update gives $x_{n+1} = \alpha$. In this case the same identity holds if we choose $\xi_n = \alpha$.

Consequently, in either case, the derivative bounds imply

$$|e_{n+1}| = \frac{|f''(\xi_n)|}{2|f'(x_n)|} |e_n|^2 \leq \frac{M}{2m} |e_n|^2 = K |e_n|^2. \tag{4.17}$$

At this stage, this estimate has been established whenever $x_n \in I$; we next prove that this condition continues to hold.

2. Every iterate is well defined and remains in J.

   We argue by induction. The assumption $|x_0 - \alpha| \leq \delta$ gives $x_0 \in J$.

   Suppose that x_n has been constructed and belongs to J. Then $x_n \in I$, so the preceding step shows that x_{n+1} is well defined. Moreover, using $|e_n| \leq \delta$ and $K\delta \leq 1/2$, we obtain

$$\begin{aligned} |e_{n+1}| &\leq K|e_n|^2 \\ &\leq K\delta |e_n| \\ &\leq \frac{1}{2} |e_n| \\ &\leq \frac{\delta}{2} \leq \delta. \end{aligned} \tag{4.18}$$

   Thus $x_{n+1} \in J$. Induction proves that every iterate is well defined and remains within distance δ of α. In particular, the denominator in every Newton step has absolute value at least m.

3. The iterates converge to α with the claimed quadratic error bound.

   The preceding argument establishes

$$|e_{n+1}| \le \frac{1}{2}|e_n| \qquad \text{for every } n \ge 0. \tag{4.19}$$

   Applying this inequality repeatedly gives

$$0 \le |x_n - \alpha| = |e_n| \le 2^{-n}|e_0| \le 2^{-n}\delta. \tag{4.20}$$

   Since the right-hand side tends to zero, $x_n \to \alpha$.

   Because every x_n belongs to I, the stronger estimate from the first step also holds at every iteration:

$$|e_{n+1}| \le K|e_n|^2. \tag{4.21}$$

   This proves the asserted quadratic error bound.

   More explicitly, if $K > 0$, set $t_n = K|e_n|$. Then

$$t_{n+1} \le t_n^2, \qquad 0 \le t_0 \le K\delta \le \frac{1}{2}. \tag{4.22}$$

   Induction therefore gives $t_n \le t_0^{2^n}$, and hence

$$|e_n| \le \frac{1}{K}\big(K|e_0|\big)^{2^n} \le \frac{1}{K}\,2^{-2^n}. \tag{4.23}$$

   This stronger estimate exhibits the repeated squaring of the error bound.

4. If $M = 0$, the root is reached in at most one step.

   In this case $K = 0$, so the one-step error estimate gives

$$0 \le |e_1| \le K|e_0|^2 = 0. \tag{4.24}$$

   Thus $x_1 = \alpha$. If $x_0 = \alpha$, the root has already been reached; otherwise, one Newton step reaches it exactly. ■

Remark 7. *If $f'(\alpha) \neq 0$ and $f \in C^2$ near α, continuity supplies some interval with the required bounds. This is the usual statement that Newton's method converges locally at least quadratically at a simple root. When $f''(\alpha) \neq 0$ and the iteration does not terminate exactly,*

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^2} = \left|\frac{f''(\alpha)}{2f'(\alpha)}\right|. \tag{4.25}$$

For the cubic this constant is $3\alpha/(3\alpha^2 - 1) \approx 0.768$. If $f''(\alpha) = 0$, the coefficient vanishes and the order can be higher. For an affine function $f(x) = ax + b$, $a \neq 0$, one Newton step is exact.

4.3 Pitfalls

In the use of Newton's method, consideration must be given to the proper choice of starting point. Usually, one must have some insight into the shape of the graph of the function. Sometimes, a coarse graph is adequate, but in other cases a step-by-step evaluation of the function at various points may be necessary to find a point near the root. Often, several steps of the bisection method are used to obtain a suitable starting point, so that Newton's method converges more rapidly.

Below we illustrate some of the things that can go wrong when applying Newton's method.

Example 11 (A zero derivative away from a root)**.** For $f(x) = x^3 - 1$, starting at $x_0 = 0$ makes the denominator zero even though the function is nonzero. There is no Newton step to compute. A near-zero denominator can instead produce an enormous step.

Example 12 (An exact two-cycle)**.** For $f(x) = x^3 - 2x + 2$, Newton maps 0 to 1 and 1 to 0:

$$N(0) = 0 - \frac{2}{-2} = 1, \qquad N(1) = 1 - \frac{1}{1} = 0. \tag{4.26}$$

The iteration never reaches the negative root. Smoothness and a nonzero derivative at the iterates do not imply convergence from arbitrary starts.

Example 13 (Leaving the domain)**.** For $f(x) = \log x$, $x > 0$, Newton gives $N(x) = x(1 - \log x)$. Starting at $x_0 = e^2$ produces $x_1 = -e^2$, outside the domain of the real logarithm.

Other common problems include convergence to an unintended root, inaccurate derivatives, noisy function values, and stagnation caused by rounding. These are reasons to monitor the calculation and, when possible, safeguard it with a bracket. Kelley [3] develops Newton methods and their practical convergence control in much greater depth, including systems of equations.

5 The secant method

Newton's method is an extrenely powerful technique, but is has a major weakness: the need to know the value of the deriative of f at each approximation. Frequently, $f'(x)$ is far more difficult to compute than f itself, needing more arithmetic operations to calculate than f.

5.1 Description of the method

To circumvent this problem of the derivative evaluation in Newton's method, we introduce a slight variation. By definition,

$$f'(x_{n-1}) = \lim_{y \to x_{n-1}} \frac{f(y) - f(x_{n-1})}{y - x_{n-1}}. \tag{5.1}$$

If x_{n-2} is close to x_{n-1}, then

$$f'(x_{n-1}) \approx \frac{f(x_{n-2}) - f(x_{n-1})}{x_{n-2} - x_{n-1}} = \frac{f(x_{n-1}) - f(x_{n-2})}{x_{n-1} - x_{n-2}}. \tag{5.2}$$

Using this approximation for $f'(x_{n-1})$ in the Newton's method iteration gives the **secant method**:

$$\boxed{x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}}, \qquad n \in \{1, 2, \ldots\}. \tag{5.3}$$

Two distinct initial points x_0, x_1 are required. The function values must also differ unless a root has already been found. No sign change is required.

The line through $(x_{n-1}, f(x_{n-1}))$ and $(x_n, f(x_n))$ is

$$L_n(x) = f(x_n) + d_n(x - x_n), \qquad d_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}. \tag{5.4}$$

Equation (5.3) is simply its intercept $L_n(x_{n+1}) = 0$. After the first two function evaluations, one new function value is needed for each new point.

Algorithm: Secant method

Input: A function f; initial approximations x_0, x_1; absolute tolerance $\varepsilon_{\text{abs}} > 0$; relative tolerance $\varepsilon_{\text{rel}} \geq 0$; residual tolerance $\varepsilon_f > 0$; maximum number of iterations $n_{\max} \geq 1$.
Output: An approximate root, or a message of failure.
Throughout the algorithm, terminate with failure if an arithmetic operation or function evaluation is undefined or produces a nonfinite value.

1. If $x_0 = x_1$, **return failure**: "Distinct initial approximations are required."
2. Evaluate and store $y_0 = f(x_0)$ and $y_1 = f(x_1)$.
   - If $y_0 = 0$, **return** x_0.
   - If $y_1 = 0$, **return** x_1.
3. **For** $n = 1, \ldots, n_{\max}$:
   (a) Set $d_n = y_n - y_{n-1}$.
   (b) If $d_n = 0$, **return failure**: "The secant denominator is zero."
   (c) Compute the next approximation:
   $$x_{n+1} = x_n - y_n \frac{x_n - x_{n-1}}{d_n}. \tag{5.5}$$
   (d) Evaluate and store $y_{n+1} = f(x_{n+1})$.
   (e) If $y_{n+1} = 0$, **return** x_{n+1}.
   (f) Set the step tolerance:
   $$\tau_{n+1} = \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \max\{|x_n|, |x_{n+1}|\}. \tag{5.6}$$
   If both $|x_{n+1} - x_n| \leq \tau_{n+1}$ and $|y_{n+1}| \leq \varepsilon_f$, **return** x_{n+1}.
   (g) If $x_{n+1} = x_n$, **return failure**: "The iteration has stagnated without satisfying the residual tolerance."
4. **Return failure**: "The maximum number of iterations has been reached."

Remark 8. *See Remark 5 for an explanation of the different error tolerances used in the secant method algorithm.*

Example 14. Consider our running example of $f(x) = x^3 - x - 2$. Using $x_0 = 1$ and $x_1 = 2$, we have

$$x_2 = 2 - 4\frac{2-1}{4-(-2)} = \frac{4}{3}. \tag{5.7}$$

In Figure 4, we provide a plot of these first three points of the secant method for this problem.

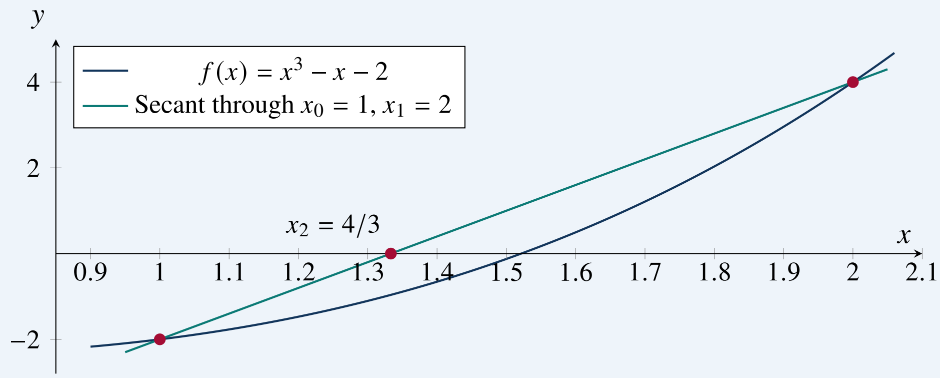


FIGURE 4: The first secant intercept for the cubic. Future secants use the two most recent points, even when their function values have the same sign.

Then, using $x_1 = 2$ and $x_2 = 4/3$, we compute x_3, and continue to obtain the following.

n	x_n	$\lvert f(x_n)\rvert$	$\lvert x_n - \alpha\rvert$
0	1.00000000000000	2.00×10^{0}	5.21×10^{-1}
1	2.00000000000000	4.00×10^{0}	4.79×10^{-1}
2	1.33333333333333	9.63×10^{-1}	1.88×10^{-1}
3	1.46268656716418	3.33×10^{-1}	5.87×10^{-2}
4	1.53116943214120	5.86×10^{-2}	9.79×10^{-3}
5	1.52092642051528	2.69×10^{-3}	4.53×10^{-4}
6	1.52137631666974	2.02×10^{-5}	3.39×10^{-6}
7	1.52137970798487	7.02×10^{-9}	1.18×10^{-9}
8	1.52137970680456	1.83×10^{-14}	3.07×10^{-15}
9	1.52137970680457	1.66×10^{-23}	2.78×10^{-24}

TABLE 4: Secant iteration for the cubic. Values are computed at high precision, then rounded for display.

This convergence is faster than a typical linear fixed-point iteration but slower per step than a generic quadratic Newton iteration. Per-step speed is only part of the cost comparison.

5.2 Convergence

We first establish an exact error identity, then use it to prove local convergence and determine the order of convergence. Throughout this subsection, the iteration is understood in exact arithmetic. If either initial point or a subsequent iterate is a root, the algorithm stops. Write

$$e_n = x_n - \alpha, \qquad E_n = |e_n|, \tag{5.8}$$

whenever x_n has been constructed.

For distinct points u, v, the first divided difference is

$$f[u,v] = \frac{f(v) - f(u)}{v - u}. \tag{5.9}$$

For three pairwise distinct points u, v, w, the second divided difference is

$$f[u,v,w] = \frac{f[v,w] - f[u,v]}{w - u}. \tag{5.10}$$

Only distinct nodes will be needed below, because we stop if a root is reached.

Proposition 2 (Exact error identity for the secant method)**.** Let $f \in C^2(I)$, where I is an interval containing a root α. Suppose $x_{n-1}, x_n \in I$ are distinct, neither is a root, and $f(x_{n-1}) \neq f(x_n)$. Then the secant step is well defined, and

$$e_{n+1} = \frac{f[x_{n-1}, x_n, \alpha]}{f[x_{n-1}, x_n]} e_n e_{n-1} = \frac{f''(\xi_n)}{2f'(\eta_n)} e_n e_{n-1}, \tag{5.11}$$

where η_n lies between x_{n-1} and x_n, and ξ_n lies in the smallest interval containing x_{n-1}, x_n, α.

Proof.

1. We establish the required interpolation remainder.

   Let $u, v, w \in I$ be pairwise distinct, and let

   $$L(t) = f(u) + f[u,v](t - u) \tag{5.12}$$

   be the line interpolating f at u and v. Direct algebra gives

   $$\begin{aligned} f(w) - L(w) &= f(w) - f(v) - f[u,v](w - v) \\ &= (w - v)\big(f[v,w] - f[u,v]\big) \\ &= f[u,v,w](w - u)(w - v). \end{aligned} \tag{5.13}$$

   Consequently, the function

   $$H(t) = f(t) - L(t) - f[u,v,w](t - u)(t - v) \tag{5.14}$$

   vanishes at u, v, w.

   Order these three nodes as $t_0 < t_1 < t_2$. Rolle's theorem provides points $s_0 \in (t_0, t_1)$ and $s_1 \in (t_1, t_2)$ with $H'(s_0) = H'(s_1) = 0$. Applying Rolle's theorem to H' on $[s_0, s_1]$ gives a point $\xi \in (s_0, s_1)$ such that

   $$0 = H''(\xi) = f''(\xi) - 2f[u,v,w]. \tag{5.15}$$

   Thus $f[u,v,w] = f''(\xi)/2$, and

   $$f(w) - L(w) = \frac{1}{2} f''(\xi)(w - u)(w - v). \tag{5.16}$$

   This argument does not require w to lie between u and v; all three nodes need only belong to I.

2. We apply the remainder formula at the root.

   Let L_n be the line interpolating f at x_{n-1}, x_n, and write

$$d_n = f[x_{n-1}, x_n] \neq 0. \tag{5.17}$$

   The secant update is $x_{n+1} = x_n - f(x_n)/d_n$, so $L_n(x_{n+1}) = 0$. Since L_n has slope d_n,

$$L_n(\alpha) = d_n(\alpha - x_{n+1}) = -d_n e_{n+1}. \tag{5.18}$$

   Using $f(\alpha) = 0$ and the interpolation identity from the first step, we obtain

$$\begin{aligned} 0 &= L_n(\alpha) + f[x_{n-1}, x_n, \alpha](\alpha - x_{n-1})(\alpha - x_n) \\ &= -d_n e_{n+1} + f[x_{n-1}, x_n, \alpha] e_{n-1} e_n. \end{aligned} \tag{5.19}$$

   Division by $d_n \neq 0$ proves the first equality in (5.11).

3. We express the divided differences using derivatives.

   The first step gives $f[x_{n-1}, x_n, \alpha] = f''(\xi_n)/2$, with ξ_n in the stated interval. The mean value theorem gives

$$d_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} = f'(\eta_n) \tag{5.20}$$

   for some η_n between x_{n-1} and x_n. Substituting these expressions proves the second equality in (5.11). ■

The product of the two most recent errors in (5.11) replaces the square of the latest error in Newton's method. We now show that sufficiently close starting points produce a convergent sequence.

Theorem 9 (Local convergence of the secant method)**.** Let f be twice continuously differentiable on an open interval containing $I = [\alpha - r, \alpha + r]$, where $r > 0$ and $f(\alpha) = 0$. Suppose

$$|f'(x)| \geq m > 0, \qquad |f''(x)| \leq M \qquad \text{for every } x \in I. \tag{5.21}$$

Set $K = M/(2m)$, and choose $0 < \delta \leq r$ such that $K\delta \leq 1/2$. Let x_0, x_1 be distinct points in

$$J = [\alpha - \delta, \alpha + \delta]. \tag{5.22}$$

Then the secant method is well defined until a root is reached, and every iterate belongs to J. The method either terminates at α or generates an infinite sequence converging to α. Every secant update satisfies

$$E_{n+1} \leq K E_n E_{n-1}, \qquad n \geq 1. \tag{5.23}$$

If the sequence is infinite, then

$$E_n \leq 2^{-(n-1)} E_1 \quad (n \geq 1), \qquad \lim_{n\to\infty} \frac{E_{n+1}}{E_n} = 0. \tag{5.24}$$

In particular, convergence is superlinear. If $M = 0$, at most one secant update is needed to reach the root.

Proof.

1. A secant step from two distinct nonroots in I is well defined and satisfies the product error bound.

   For any distinct $u, v \in I$, the mean value theorem gives

   $$|f(v) - f(u)| = |f'(\eta)|\,|v - u| \geq m|v - u| > 0 \tag{5.25}$$

   for some η between u and v. Thus distinct points in I have distinct function values. In particular, α is the unique root in I.

   Suppose $x_{n-1}, x_n \in I$ are distinct and neither is a root. Their function values differ, so the next secant point is well defined. By Proposition 2,

   $$\begin{aligned} E_{n+1} &= \frac{|f''(\xi_n)|}{2|f'(\eta_n)|} E_n E_{n-1} \\ &\leq \frac{M}{2m} E_n E_{n-1} = K E_n E_{n-1}. \end{aligned} \tag{5.26}$$

   Both intermediate points belong to I, since $x_{n-1}, x_n, \alpha \in I$. This estimate does not assume in advance that $x_{n+1} \in I$.

   Moreover,

   $$x_{n+1} - x_n = -\frac{f(x_n)}{f[x_{n-1}, x_n]} \neq 0. \tag{5.27}$$

   Hence the next pair of defining points is distinct.

2. Every constructed iterate remains in J.

   The starting points belong to J. If either is a root, it must be α, and the algorithm stops.

   Otherwise, suppose inductively that $x_{n-1}, x_n \in J$ are distinct nonroots. The first step constructs x_{n+1} and gives

   $$\begin{aligned} E_{n+1} &\leq K E_n E_{n-1} \\ &\leq K\delta E_n \\ &\leq \frac{1}{2} E_n \leq \frac{\delta}{2} \leq \delta. \end{aligned} \tag{5.28}$$

   Therefore $x_{n+1} \in J$. If it is a root, it equals α and the algorithm stops. Otherwise, x_n, x_{n+1} are distinct nonroots in J, so the induction continues. This proves both invariance of J and the absence of a zero denominator before termination.

3. An infinite sequence converges to α superlinearly.

   If the algorithm never reaches a root, the preceding step gives $E_{n+1} \leq E_n/2$ for every $n \geq 1$. Repeated application yields

   $$0 < E_n \leq 2^{-(n-1)} E_1 \leq 2^{-(n-1)}\delta, \qquad n \geq 1. \tag{5.29}$$

   The right-hand side tends to zero, so $x_n \to \alpha$.

   The stronger product bound now gives

   $$0 \leq \frac{E_{n+1}}{E_n} \leq K E_{n-1}. \tag{5.30}$$

   Since $E_{n-1} \to 0$, the squeeze theorem implies $E_{n+1}/E_n \to 0$. This is superlinear convergence.

4. If $M = 0$, one update reaches the root unless it has already been reached.

   Here $K = 0$. If neither starting point is a root, the first secant update satisfies

$$0 \le E_2 \le K E_1 E_0 = 0. \tag{5.31}$$

   Thus $x_2 = \alpha$, as claimed. ∎

Remark 10. *If $f \in C^2$ near a simple root α, then $f'(\alpha) \neq 0$. Continuity of f' supplies a sufficiently small interval I on which $|f'(x)| \ge |f'(\alpha)|/2 > 0$, and continuity of f'' supplies a finite bound M on I. The theorem therefore proves local convergence whenever both starting points are sufficiently close to a simple root. The starting points need not lie on opposite sides of the root.*

Theorem 11 (Order of convergence of the secant method)**.** Under the hypotheses of Theorem 9, suppose in addition that $f''(\alpha) \neq 0$ and that no iterate equals α. Set

$$A = \left| \frac{f''(\alpha)}{2f'(\alpha)} \right| > 0, \tag{5.32}$$

and let

$$\varphi = \frac{1 + \sqrt{5}}{2} \approx 1.6180339887. \tag{5.33}$$

Then

$$\lim_{n \to \infty} \frac{E_{n+1}}{E_n^{\varphi}} = A^{1/\varphi}. \tag{5.34}$$

Thus the secant method has order of convergence φ, with asymptotic error constant $A^{1/\varphi}$.

Proof.

1. We determine the limiting coefficient in the error identity.

   Theorem 9 gives $x_n \to \alpha$. The intermediate points in (5.11) satisfy

$$|\xi_n - \alpha| \le \max\{E_{n-1}, E_n\}, \qquad |\eta_n - \alpha| \le \max\{E_{n-1}, E_n\}. \tag{5.35}$$

   Consequently, both ξ_n and η_n tend to α. Continuity of f' and f'', together with $f'(\alpha) \neq 0$, yields

$$e_{n+1} = \left(\frac{f''(\alpha)}{2f'(\alpha)} + o(1) \right) e_n e_{n-1}. \tag{5.36}$$

   Since no error is zero, we may define

$$A_n = \frac{E_{n+1}}{E_n E_{n-1}} = \left| \frac{f''(\xi_n)}{2f'(\eta_n)} \right| > 0, \qquad n \ge 1. \tag{5.37}$$

   It follows that $A_n \to A > 0$.

2. We obtain a recurrence for the logarithms of the errors.

   The number φ satisfies $\varphi^2 = \varphi + 1$. Set $q = 1/\varphi$, so that $0 < q < 1$, $1 - \varphi = -q$, and $1 + q = \varphi$. Define

$$D_n = \log E_n - \varphi \log E_{n-1}, \qquad n \ge 1. \tag{5.38}$$

All logarithms are well defined because $E_n > 0$. Taking logarithms of the exact identity $E_{n+1} = A_n E_n E_{n-1}$ gives

$$\begin{aligned} D_{n+1} &= \log A_n + (1-\varphi)\log E_n + \log E_{n-1} \\ &= \log A_n - qD_n. \end{aligned} \tag{5.39}$$

The candidate limit is

$$D = \frac{\log A}{1+q} = \frac{\log A}{\varphi}, \tag{5.40}$$

which satisfies $D = \log A - qD$.

3. We prove that $D_n \to D$.

   Put $z_n = D_n - D$ and $r_n = \log A_n - \log A$. Since $A_n \to A > 0$, continuity of the logarithm gives $r_n \to 0$. Subtracting the equation for D from the recurrence for D_n yields

$$z_{n+1} = -qz_n + r_n. \tag{5.41}$$

   Let $\varepsilon > 0$. Choose $N \geq 1$ such that

$$|r_n| \leq \frac{(1-q)\varepsilon}{2} \qquad \text{for every } n \geq N. \tag{5.42}$$

   Iterating the recurrence and applying the triangle inequality gives, for every integer $k \geq 1$,

$$\begin{aligned} |z_{N+k}| &\leq q^k |z_N| + \sum_{j=0}^{k-1} q^{k-1-j}|r_{N+j}| \\ &\leq q^k |z_N| + \frac{(1-q)\varepsilon}{2}\sum_{j=0}^{k-1} q^j \\ &\leq q^k |z_N| + \frac{\varepsilon}{2}. \end{aligned} \tag{5.43}$$

   Because $0 < q < 1$, the first term is less than $\varepsilon/2$ for all sufficiently large k. Thus $|z_{N+k}| < \varepsilon$ eventually. This proves $z_n \to 0$, and hence $D_n \to D$.

4. We recover the asymptotic error constant.

   By the definition of D_{n+1},

$$\frac{E_{n+1}}{E_n^{\varphi}} = \exp(D_{n+1}) \longrightarrow \exp\left(\frac{\log A}{\varphi}\right) = A^{1/\varphi}. \tag{5.44}$$

   This limit is finite and strictly positive, so it establishes the claimed order and asymptotic error constant. ■

Remark 12. *The order theorem requires both $f'(\alpha) \neq 0$ and $f''(\alpha) \neq 0$, and it excludes exact termination. If $f''(\alpha) = 0$, the limiting coefficient in (5.36) vanishes. The local convergence theorem still guarantees superlinear convergence for an infinite sequence, but the specific order φ is no longer asserted. Further analysis is needed to determine an exact order in that case. Multiple roots, for which $f'(\alpha) = 0$, are outside the hypotheses of these theorems. For further discussion of the secant method, see Driscoll and Braun [1].*

5.3 Pitfalls

For $f(x) = x^2 - 1$, the starts $x_0 = -2$, $x_1 = 2$ give equal function values and an undefined secant step. Nearly equal computed values can produce an unreliable slope or a very large step. As with Newton's method, the secant method can leave the domain, diverge, or converge to an unintended root. Also, a small difference $x_n - x_{n-1}$ is not itself an error certificate; it may instead signal that rounding has made the two points numerically indistinguishable.

A Calculus background

A.1 Continuity, signs, and the existence of a root

Continuity at c means $\lim_{x\to c} f(x) = f(c)$: sufficiently small changes in the input cause sufficiently small changes in the output near c. Polynomials are continuous on $\mathbb{R}$. Rational functions are continuous wherever their denominators are nonzero. Logarithms require positive arguments.

Theorem 13 (Intermediate value theorem)**.** If f is continuous on $[a, b]$ and y lies between $f(a)$ and $f(b)$, then there is a $c \in [a, b]$ with $f(c) = y$. In particular, if $f(a)$ and $f(b)$ have opposite signs, there is a root in (a, b).

This theorem supplies the basic certificate used by bisection. It guarantees existence, not uniqueness. A *bracket* in this note is an interval on which continuity is known and whose endpoints have opposite signs, unless an endpoint is already a root.

A.2 Derivatives as local linear models

The derivative is defined by

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}, \tag{A.1}$$

when the limit exists. It is a rate of change and the slope of the tangent line. Differentiability at x is equivalently the expansion

$$f(x+h) = f(x) + f'(x)h + o(h). \tag{A.2}$$

Here $o(h)$ denotes a remainder whose ratio to h tends to zero as $h \to 0$. For a small increment h, the tangent model $f(x) + f'(x)h$ approximates $f(x + h)$. Solving this model for zero is the idea behind Newton's method.

$f(x)$	$f'(x)$	$f(x)$	$f'(x)$
c	0	e^x	e^x
x^p	px^{p-1}	$\log x$	$1/x$
$\sin x$	$\cos x$	$\cos x$	$-\sin x$

The power rule is used on domains where the real-valued power is defined and differentiable. Trigonometric derivatives assume angles in radians. The main differentiation rules are

$$(au + bv)' = au' + bv', \qquad (uv)' = u'v + uv',$$

$$\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2} \quad (v \neq 0), \qquad\qquad (u \circ v)'(x) = u'(v(x))v'(x).$$

The final rule is the *chain rule*. For example,

$$\frac{\mathrm{d}}{\mathrm{d}x}(x+2)^{1/3} = \frac{1}{3(x+2)^{2/3}}, \qquad \frac{\mathrm{d}}{\mathrm{d}x}\log(1+x^2) = \frac{2x}{1+x^2}. \tag{A.3}$$

Differentiating again gives the second derivative f''. It measures how the slope changes and controls the leading error in a tangent approximation. For $f(x) = x^3 - x - 2$, we have $f' = 3x^2 - 1$ and $f'' = 6x$.

Example 15 (Derive a derivative from its definition)**.** For $f(x) = x^2$,

$$\frac{(x+h)^2 - x^2}{h} = 2x + h \longrightarrow 2x. \tag{A.4}$$

The exact identity $(x+h)^2 = x^2 + 2xh + h^2$ also exhibits the tangent error: it is h^2, which is small compared with h near zero.

A.3 The mean value theorem and useful consequences

Theorem 14 (Mean value theorem)**.** If f is continuous on $[a, b]$ and differentiable on (a, b), then there is a $\xi \in (a, b)$ such that

$$f(b) - f(a) = f'(\xi)(b - a). \tag{A.5}$$

Geometrically, at some intermediate point the tangent has the same slope as the secant joining the endpoints. When $f(a) = f(b)$, this is Rolle's theorem: $f'(\xi) = 0$ somewhere between them.

Three consequences are used repeatedly.

1. If $f'(x) > 0$ throughout an interval, f is strictly increasing there, so it has at most one root. The corresponding statement holds for $f' < 0$.
2. If $|f'(x)| \leq L$ throughout an interval, then $|f(x) - f(y)| \leq L|x - y|$ for any two points in it. This is a *Lipschitz bound*; it limits amplification of input differences.
3. If $f(\alpha) = 0$ and $|f'(t)| \geq m > 0$ throughout the segment joining x and α, then

   $$|x - \alpha| \leq \frac{|f(x)|}{m}. \tag{A.6}$$

   This turns an observable residual into an error bound.

For the third statement, use $f(x) = f'(\xi)(x - \alpha)$ and take absolute values. The bound requires a derivative lower bound on the whole segment; substituting $|f'(x)|$ for m without justification is not a proof.

A.4 Taylor's theorem and the size of an approximation error

The notation $f \in C^k(I)$ means that f has continuous derivatives through order k on I (with the usual one-sided interpretation at endpoints).

Theorem 15 (Taylor's theorem with remainder)**.** Suppose $f \in C^{p+1}$ on an interval containing x and $x+h$. There is a point ξ between them such that

$$f(x+h) = \sum_{j=0}^{p} \frac{f^{(j)}(x)}{j!}h^j + \frac{f^{(p+1)}(\xi)}{(p+1)!}h^{p+1}. \tag{A.7}$$

The two-term version is especially important:

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(\xi)h^2. \tag{A.8}$$

If $|f''| \le M$, the tangent approximation has absolute error at most $Mh^2/2$. This gives a quantitative meaning to "the model is accurate near the current point." It does not justify a tangent model arbitrarily far away.

We write $r(h) = O(h^p)$ as $h \to 0$ if $|r(h)| \le C|h|^p$ for some fixed C and all sufficiently small nonzero h. We write $r(h) = o(h^p)$ if $r(h)/h^p \to 0$. Thus, for $f \in C^2$,

$$f(\alpha+e) = f(\alpha) + f'(\alpha)e + \tfrac{1}{2}f''(\alpha)e^2 + o(e^2). \tag{A.9}$$

Taylor's theorem is a finite approximation statement; it does not require an infinite Taylor series to converge to f.

A.5 Simple roots, multiple roots, and convexity

Definition 3 (Root multiplicity)**.** A differentiable function has a *simple root* at α if $f(\alpha) = 0$ and $f'(\alpha) \neq 0$. A sufficiently smooth function has a root of multiplicity $m \ge 2$ if locally

$$f(x) = (x-\alpha)^m h(x), \qquad h(\alpha) \neq 0, \tag{A.10}$$

with h sufficiently smooth. Equivalently, the first nonzero derivative at the root is $f^{(m)}(\alpha)$, under the corresponding smoothness assumptions.

Simple roots cross the axis locally. Finite odd-multiplicity roots change sign; finite even-multiplicity roots do not. A root need not have finite multiplicity: smooth functions can have every derivative vanish at a root. Our multiple-root analysis concerns the stated finite factorization.

If $f'' \ge 0$ on an interval, f is *convex*: its graph lies above its tangents. In symbols, $f(y) \ge f(x) + f'(x)(y-x)$. This property will give a useful one-sided convergence result for Newton's method.

B Convergence order and the design of iterative methods

We now investigate how rapidly an iterative method approaches its solution. We will relate the convergence rate of a fixed-point iteration to the derivatives of its iteration function and use this relationship to derive Newton's method.

B.1 Order of convergence

Let $\{x_n\}_{n=0}^{\infty}$ converge to α, and write $e_n = x_n - \alpha$ for the error in the nth approximation.

Definition 4 (Order of convergence)**.** Suppose that $x_n \neq \alpha$ for all sufficiently large n. If there exist constants $p \geq 1$ and $\lambda > 0$ such that

$$\lim_{n\to\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^p} = \lambda, \tag{B.1}$$

then the sequence converges to α with *order* p and *asymptotic error constant* λ. Equivalently,

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = \lambda. \tag{B.2}$$

The definition means that, for sufficiently large n, the errors approximately satisfy

$$|e_{n+1}| \approx \lambda |e_n|^p. \tag{B.3}$$

Thus, the order specifies the power of the current error that determines the next error. The asymptotic error constant supplies the corresponding multiplicative factor.

Two cases are particularly important:

1. **Linear convergence:** $p = 1$ and $0 < \lambda < 1$. In this case,

$$|e_{n+1}| \approx \lambda |e_n|. \tag{B.4}$$

   Each iteration eventually reduces the error by approximately the same factor.

2. **Quadratic convergence:** $p = 2$ and $\lambda > 0$. In this case,

$$|e_{n+1}| \approx \lambda |e_n|^2. \tag{B.5}$$

   Once the error is small, squaring it can produce a much smaller next error. The constant λ need not be less than one.

An iteration $x_{n+1} = g(x_n)$ is said to have a particular order of convergence when the sequence it generates has that order. Statements about the order of a method therefore include conditions on the iteration function and the starting point.

Remark 16 (Exact order, error bounds, and finite termination)**.** *The limit in the definition uses a strictly positive asymptotic error constant. It therefore describes an* exact *order.*

The phrase at least quadratic convergence *will mean that the sequence converges and that there exist* $C > 0$ *and an index* n_0 *such that*

$$|e_{n+1}| \leq C |e_n|^2, \qquad n \geq n_0. \tag{B.6}$$

This bound also allows convergence faster than exact order two.

If a fixed-point iteration reaches α *after finitely many steps, then all subsequent iterates equal* α*. The iteration has terminated exactly, and the error ratios above are no longer defined. We treat finite termination separately.*

B.2 Comparing linear and quadratic convergence

A higher order generally produces a faster asymptotic decrease in the error once the iterates are sufficiently close to the solution. The following model comparison illustrates the size of this difference.

Example 16 (Linear and quadratic error reduction)**.** Consider two nonnegative error sequences satisfying

$$a_{n+1} = \frac{1}{2}a_n, \qquad b_{n+1} = \frac{1}{2}b_n^2. \tag{B.7}$$

These exact recurrences model linear and quadratic convergence, respectively, with the same asymptotic error constant $1/2$.
For the linear recurrence, repeated substitution gives

$$a_n = \left(\frac{1}{2}\right)^n a_0. \tag{B.8}$$

For the quadratic recurrence, the first few substitutions give

$$\begin{aligned} b_1 &= \left(\frac{1}{2}\right) b_0^2, \\ b_2 &= \left(\frac{1}{2}\right)^3 b_0^4, \\ b_3 &= \left(\frac{1}{2}\right)^7 b_0^8. \end{aligned} \tag{B.9}$$

Induction gives the general formula

$$b_n = \left(\frac{1}{2}\right)^{2^n-1} b_0^{2^n}. \tag{B.10}$$

Indeed, if the formula holds at index n, then

$$\begin{aligned} b_{n+1} &= \frac{1}{2}\left[\left(\frac{1}{2}\right)^{2^n-1} b_0^{2^n}\right]^2 \\ &= \left(\frac{1}{2}\right)^{2^{n+1}-1} b_0^{2^{n+1}}. \end{aligned} \tag{B.11}$$

Taking $a_0 = b_0 = 1$ gives the following comparison:

n	Linear error a_n	Quadratic error b_n
1	5.0000×10^{-1}	5.0000×10^{-1}
2	2.5000×10^{-1}	1.2500×10^{-1}
3	1.2500×10^{-1}	7.8125×10^{-3}
4	6.2500×10^{-2}	3.0518×10^{-5}
5	3.1250×10^{-2}	4.6566×10^{-10}
6	1.5625×10^{-2}	1.0842×10^{-19}
7	7.8125×10^{-3}	5.8775×10^{-39}

The quadratic error is below 10^{-38} after seven updates. For the linear recurrence, the corresponding requirement is

$$2^{-n} \le 10^{-38}, \tag{B.12}$$

which is equivalent to

$$n \ge \lceil 38 \log_2(10) \rceil = 127. \tag{B.13}$$

In fact, $b_7 = a_{127} = 2^{-127}$: seven quadratic updates achieve the same error as 127 linear updates in this model.
These recurrences are idealized models. An asymptotic error formula describes sufficiently late iterations and need not accurately describe the first few steps of an actual method.

B.3 Linear convergence of fixed-point iteration

The derivative of the iteration function at its fixed point determines the leading error behavior.

Theorem 17 (Linear convergence of a fixed-point iteration)**.** Let g be continuously differentiable on an open interval containing $[a, b]$. Suppose that

$$g([a, b]) \subseteq [a, b] \qquad \text{and} \qquad |g'(x)| \le q < 1 \quad \text{for every } x \in [a, b], \tag{B.14}$$

where $0 < q < 1$.
Let α be the unique fixed point of g in $[a, b]$, and suppose that $g'(\alpha) \ne 0$. For every $x_0 \in [a, b]$, the iteration

$$x_{n+1} = g(x_n), \qquad n = 0, 1, 2, \ldots, \tag{B.15}$$

either reaches α after finitely many steps or converges linearly to α, with asymptotic error constant

$$\lambda = |g'(\alpha)|. \tag{B.16}$$

Proof.

1. Establish convergence.

   By the mean value theorem, for any $u, v \in [a, b]$,

   $$|g(u) - g(v)| \le q|u - v|. \tag{B.17}$$

   Thus, g is a contraction of $[a, b]$ into itself. The contraction theorem gives a unique fixed point $\alpha \in [a, b]$ and convergence $x_n \to \alpha$ for every $x_0 \in [a, b]$.

2. Relate successive errors. If some $x_n = \alpha$, then $x_{n+1} = g(\alpha) = \alpha$, so the iteration has terminated exactly.

   Otherwise, $x_n \ne \alpha$ for every n. Applying the mean value theorem between x_n and α gives a point ξ_n between them such that

   $$\begin{aligned} x_{n+1} - \alpha &= g(x_n) - g(\alpha) \\ &= g'(\xi_n)(x_n - \alpha). \end{aligned} \tag{B.18}$$

   Hence,

   $$\frac{x_{n+1} - \alpha}{x_n - \alpha} = g'(\xi_n). \tag{B.19}$$

3. Evaluate the asymptotic error constant. Since ξ_n lies between x_n and α,

$$|\xi_n - \alpha| \le |x_n - \alpha| \longrightarrow 0. \tag{B.20}$$

Continuity of g' therefore gives

$$\lim_{n\to\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|} = \lim_{n\to\infty} |g'(\xi_n)| = |g'(\alpha)|. \tag{B.21}$$

Because $0 < |g'(\alpha)| \le q < 1$, the convergence is linear.

■

Consequently, to obtain an order greater than one from a differentiable fixed-point iteration, we must arrange for $g'(\alpha) = 0$. Indeed, if a nonterminating iteration has order $p > 1$, then

$$\frac{|e_{n+1}|}{|e_n|} = \frac{|e_{n+1}|}{|e_n|^p} |e_n|^{p-1} \longrightarrow 0. \tag{B.22}$$

On the other hand, differentiability at α gives

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n} = \lim_{n\to\infty} \frac{g(x_n) - g(\alpha)}{x_n - \alpha} = g'(\alpha). \tag{B.23}$$

These two limits imply $g'(\alpha) = 0$.

B.4 Quadratic convergence of fixed-point iteration

The next result provides sufficient conditions for the desired quadratic error reduction.

Theorem 18 (Local quadratic convergence of a fixed-point iteration)**.** Let I be an open interval containing α, and suppose that $g \in C^2(I)$ satisfies

$$g(\alpha) = \alpha, \qquad g'(\alpha) = 0. \tag{B.24}$$

Suppose also that, for some $M \ge 0$,

$$|g''(x)| \le M, \qquad x \in I. \tag{B.25}$$

There exists $\delta > 0$ such that, for every $x_0 \in [\alpha - \delta, \alpha + \delta]$, the iteration $x_{n+1} = g(x_n)$ remains in this interval and converges to α. Moreover,

$$|x_{n+1} - \alpha| \le \frac{M}{2} |x_n - \alpha|^2, \qquad n \ge 0. \tag{B.26}$$

If the iteration does not terminate exactly, then

$$\lim_{n\to\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^2} = \frac{|g''(\alpha)|}{2}. \tag{B.27}$$

In particular, if $g''(\alpha) \ne 0$, the convergence has exact order two.

Proof.

1. Choose an interval on which the derivative is small. Fix q with $0 < q < 1$. Since g' is continuous and $g'(\alpha) = 0$, there exists $\delta > 0$ such that
$$J = [\alpha - \delta, \alpha + \delta] \subset I \qquad \text{and} \qquad |g'(x)| \le q \quad \text{for every } x \in J. \tag{B.28}$$

2. Prove that the iteration stays in J and converges. For $x \in J$, the mean value theorem gives
$$|g(x) - \alpha| = |g(x) - g(\alpha)| \le q|x - \alpha| \le q\delta \le \delta. \tag{B.29}$$
Thus, $g(J) \subseteq J$. The same derivative bound gives
$$|g(u) - g(v)| \le q|u - v|, \qquad u, v \in J. \tag{B.30}$$
Therefore, g is a contraction on J. The contraction theorem shows that every iteration starting in J remains there and converges to α.

3. Derive the quadratic error bound. Taylor's theorem about α gives, for each n,
$$g(x_n) = g(\alpha) + g'(\alpha)(x_n - \alpha) + \frac{g''(\xi_n)}{2}(x_n - \alpha)^2, \tag{B.31}$$
where ξ_n lies between x_n and α. If $x_n = \alpha$, the same identity holds by taking $\xi_n = \alpha$.
Using $g(\alpha) = \alpha$, $g'(\alpha) = 0$, and $x_{n+1} = g(x_n)$, we obtain
$$x_{n+1} - \alpha = \frac{g''(\xi_n)}{2}(x_n - \alpha)^2. \tag{B.32}$$
Since $\xi_n \in J \subset I$, it follows that
$$|x_{n+1} - \alpha| \le \frac{M}{2}|x_n - \alpha|^2. \tag{B.33}$$
This bound holds for every $n \ge 0$.

4. Determine the exact order when possible. Suppose that the iteration does not terminate exactly. Dividing the error identity by $|x_n - \alpha|^2$ gives
$$\frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^2} = \frac{|g''(\xi_n)|}{2}. \tag{B.34}$$
Since $x_n \longrightarrow \alpha$ and ξ_n lies between x_n and α, we have $\xi_n \longrightarrow \alpha$. Continuity of g'' now gives
$$\lim_{n\to\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^2} = \frac{|g''(\alpha)|}{2}. \tag{B.35}$$
If $g''(\alpha) \ne 0$, this limit is positive, establishing exact quadratic convergence.

■

Remark 19 (When the quadratic error constant vanishes)**.** *If $g''(\alpha) = 0$ and the iteration does not terminate exactly, the preceding proof gives*
$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^2} = 0. \tag{B.36}$$

The convergence is therefore faster than a quadratic error law with a positive asymptotic constant. Under the stated C^2 assumption, however, this does not by itself establish an exact order $p > 2$.

Additional derivative information can identify a higher order. For example, if $g \in C^3$ near α and

$$g'(\alpha) = g''(\alpha) = 0, \qquad g'''(\alpha) \neq 0, \tag{B.37}$$

then Taylor's theorem gives, for a nonterminating iteration,

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^3} = \frac{|g'''(\alpha)|}{6}. \tag{B.38}$$

The convergence then has exact order three.

B.5 Deriving Newton's method from fixed-point iteration

The preceding results suggest a way to design a rapidly convergent iteration for solving $f(x) = 0$. We seek an iteration function g satisfying

$$g(\alpha) = \alpha \qquad \text{and} \qquad g'(\alpha) = 0 \tag{B.39}$$

at the desired root α.

Suppose that $f \in C^2$ near α and that

$$f(\alpha) = 0, \qquad f'(\alpha) \neq 0. \tag{B.40}$$

Consider iteration functions of the form

$$g(x) = x - \phi(x) f(x), \tag{B.41}$$

where ϕ is differentiable and is to be chosen. Since $f(\alpha) = 0$, every such choice gives

$$g(\alpha) = \alpha. \tag{B.42}$$

If ϕ is nonzero throughout the neighborhood under consideration, the fixed points of g there are precisely the zeros of f.

Differentiation gives

$$g'(x) = 1 - \phi'(x) f(x) - \phi(x) f'(x). \tag{B.43}$$

Evaluating at α and using $f(\alpha) = 0$, we obtain

$$g'(\alpha) = 1 - \phi(\alpha) f'(\alpha). \tag{B.44}$$

Consequently,

$$g'(\alpha) = 0 \quad \Longleftrightarrow \quad \phi(\alpha) = \frac{1}{f'(\alpha)}. \tag{B.45}$$

Although α is unknown, we can ensure this identity by choosing the entire function

$$\phi(x) = \frac{1}{f'(x)}. \tag{B.46}$$

Continuity of f' and the condition $f'(\alpha) \neq 0$ ensure that this choice is well-defined on a sufficiently small interval containing α.

The resulting fixed-point function is

$$g(x) = x - \frac{f(x)}{f'(x)}, \tag{B.47}$$

and its iteration is

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}}. \tag{B.48}$$

This is Newton's method.

For this choice of g, direct differentiation confirms that

$$g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}, \tag{B.49}$$

so $g'(\alpha) = 0$ as required.