Chapter 2

Solving Nonlinear Equations

2.1 The bisection method

Algorithm 2.1. The *bisection method* finds a root of a continuous function $f : \mathbb{R} \to \mathbb{R}$ by repeatedly reducing the interval to the half interval where the root must lie.

Input: $f : [a, b] \to \mathbb{R},\ a \in \mathbb{R},\ b \in \mathbb{R},$
$M \in \mathbb{N}^+,\ \delta \in \mathbb{R}^+,\ \epsilon \in \mathbb{R}^+$
Preconditions : $f \in \mathcal{C}[a, b],$
$\operatorname{sgn}(f(a)) \neq \operatorname{sgn}(f(b))$
Output: c, h, k
Postconditions: $|f(c)| < \epsilon$ or $|h| < \delta$ or $k = M$

$u \leftarrow f(a)$
$v \leftarrow f(b)$
for $k = 1 : M$ **do**
  $h \leftarrow b - a$
  $c \leftarrow a + h/2$
  $w \leftarrow f(c)$
  if $|h| < \delta$ **or** $|w| < \epsilon$ **then**
    break
  else if $\operatorname{sgn}(w) \neq \operatorname{sgn}(u)$ **then**
    $b \leftarrow c$
    $v \leftarrow w$
  else
    $a \leftarrow c$
    $u \leftarrow w$
  end
end

2.2 The signature of an algorithm

Definition 2.2. An *algorithm* is a step-by-step procedure that takes some set of values as its *input* and produces some set of values as its *output*.

Definition 2.3. A *precondition* is a condition that holds for the input prior to the execution of an algorithm.

Definition 2.4. A *postcondition* is a condition that holds for the output after the execution of an algorithm.

Definition 2.5. The *signature of an algorithm* consists of its input, output, preconditions, postconditions, and how input parameters violating preconditions are handled.

2.3 Proof of correctness and simplification of algorithms

Definition 2.6. An *invariant* is a condition that holds during the execution of an algorithm.

Definition 2.7. A variable is *temporary or derived* for a loop if it is initialized inside the loop. A variable is *persistent or primary* for a loop if it is initialized before the loop and its value changes across different iterations.

Exercise 2.1. What are the invariants in Algorithm 2.1? Which quantities do a, b, c, h, u, v, w represent? Which of them are primary? Which of these variables are temporary? Draw pictures to illustrate the life spans of these variables.

Algorithm 2.8. A simplified bisection algorithm.

Input: $f : [a, b] \to \mathbb{R},\ a \in \mathbb{R},\ b \in \mathbb{R},$
$M \in \mathbb{N}^+,\ \delta \in \mathbb{R}^+,\ \epsilon \in \mathbb{R}^+$
Preconditions : $f \in \mathcal{C}[a, b],$
$\operatorname{sgn}(f(a)) \neq \operatorname{sgn}(f(b))$
Output: c, h, k
Postconditions: $|f(c)| < \epsilon$ or $|h| < \delta$ or $k = M$

$h \leftarrow b - a$
$u \leftarrow f(a)$
for $k = 1 : M$ **do**
  $h \leftarrow h/2$
  $c \leftarrow a + h$
  $w \leftarrow f(c)$
  if $|h| < \delta$ **or** $|w| < \epsilon$ **then**
    break
  else if $\operatorname{sgn}(w) = \operatorname{sgn}(u)$ **then**
    $a \leftarrow c$
  end
end

2.4 Q-order convergence

Definition 2.9 (Q-order convergence)**.** A convergent sequence $\{x_n\}$ is said to *converge* to L with *Q-order* p ($p \geq 1$) if

$$\lim_{n\to\infty} \frac{|x_{n+1} - L|}{|x_n - L|^p} = c > 0; \tag{2.1}$$

the constant c is called the *asymptotic factor*. In particular, $\{x_n\}$ has *Q-linear convergence* if $p = 1$ and *Q-quadratic convergence* if $p = 2$.

Definition 2.10. A sequence of iterates $\{x_n\}$ is said to *converge linearly* to L if

$$\exists c \in (0,1), \exists d > 0, \text{ s.t. } \forall n \in \mathbb{N}, \ |x_n - L| \le c^n d. \tag{2.2}$$

In general, the *order of convergence* of a sequence $\{x_n\}$ converging to L is the maximum $p \in \mathbb{R}^+$ satisfying

$$\exists c > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall n > N, \ |x_{n+1} - L| \le c|x_n - L|^p. \tag{2.3}$$

In particular, $\{x_n\}$ *converges quadratically* if $p = 2$.

Theorem 2.11 (Monotonic sequence theorem)**.** Every bounded monotonic sequence is convergent.

Theorem 2.12 (Convergence of the bisection method)**.** For a continuous function $f : [a_0, b_0] \to \mathbb{R}$ satisfying $\operatorname{sgn}(f(a_0)) \ne \operatorname{sgn}(f(b_0))$, the sequence of iterates in the bisection method converges linearly with asymptotic factor $\frac{1}{2}$,

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n = \alpha, \tag{2.4}$$

$$f(\alpha) = 0, \tag{2.5}$$

$$|c_n - \alpha| \le 2^{-(n+1)}(b_0 - a_0), \tag{2.6}$$

where $[a_n, b_n]$ is the interval in the nth iteration of the bisection method and $c_n = \frac{1}{2}(a_n + b_n)$.

Proof. It follows from the bisection method that

$$\begin{aligned} a_0 \le a_1 \le a_2 \le \cdots \le b_0, \\ b_0 \ge b_1 \ge b_2 \ge \cdots \ge a_0, \\ b_{n+1} - a_{n+1} = \frac{1}{2}(b_n - a_n). \end{aligned}$$

In the rest of this proof, "lim" is a shorthand for "$\lim_{n\to\infty}$." By Theorem 2.11, both $\{a_n\}$ and $\{b_n\}$ converge. Also, $\lim(b_n - a_n) = \lim \frac{1}{2^n}(b_0 - a_0) = 0$, hence $\lim b_n = \lim a_n = \alpha$. By the given condition and the algorithm, the invariant $f(a_n)f(b_n) \le 0$ always holds. Since f is continuous, $\lim f(a_n)f(b_n) = f(\lim a_n)f(\lim b_n)$, then $f^2(\alpha) \le 0$ implies $f(\alpha) = 0$. (2.6) is another important invariant that can be proven by induction. Comparing (2.6) to (2.2) yields convergence of the bisection method. Also, the convergence is linear with asymptotic factor as $c = \frac{1}{2}$. □

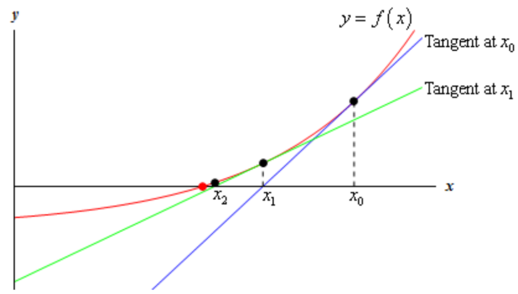
2.5 Newton's method

Algorithm 2.13. *Newton's method* finds the root of $f : \mathbb{R} \to \mathbb{R}$ near an initial guess x_0 by the iteration formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n \in \mathbb{N}. \tag{2.7}$$

Input: $f : \mathbb{R} \to \mathbb{R}$, f', $x_0 \in \mathbb{R}$, $M \in \mathbb{N}^+$, $\epsilon \in \mathbb{R}^+$
Preconditions : $f \in \mathcal{C}^2$ and x_0 is sufficiently close to a root of f
Output: x, k
Postconditions: $|f(x)| < \epsilon$ or $k = M$

$x \leftarrow x_0$
for $k = 0 : M$ **do**
  $u \leftarrow f(x)$
  if $|u| < \epsilon$ **then**
    break
  end
  $x \leftarrow x - u/f'(x)$
end



Theorem 2.14 (Convergence of Newton's method)**.** Consider a $\mathcal{C}^2$ function $f : \mathcal{B} \to \mathbb{R}$ on $\mathcal{B} = [\alpha - \delta, \alpha + \delta]$ satisfying $f(\alpha) = 0$ and $f'(\alpha) \ne 0$. If x_0 is chosen sufficiently close to α, then the sequence of iterates $\{x_n\}$ in the Newton's method converges quadratically to the root α, i.e.

$$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{(\alpha - x_n)^2} = -\frac{f''(\alpha)}{2f'(\alpha)}. \tag{2.8}$$

Proof. By Taylor's theorem (Theorem 0.48) and the assumption $f \in \mathcal{C}^2$,

$$f(\alpha) = f(x_n) + (\alpha - x_n)f'(x_n) + \frac{(\alpha - x_n)^2}{2} f''(\xi)$$

where ξ is between α and x_n. $f(\alpha) = 0$ yields

$$-\alpha = -x_n + \frac{f(x_n)}{f'(x_n)} + \frac{(\alpha - x_n)^2}{2} \frac{f''(\xi)}{f'(x_n)}.$$

By (2.7), we have

$$(*): \quad x_{n+1} - \alpha = x_n - \frac{f(x_n)}{f'(x_n)} - \alpha = (x_n - \alpha)^2 \frac{f''(\xi)}{2f'(x_n)}.$$

The continuity of f' and the assumption $f'(\alpha) \ne 0$ yield

$$\exists \delta_1 \in (0, \delta) \text{ s.t. } \forall x \in \mathcal{B}_1, \ f'(x) \ne 0$$

where $\mathcal{B}_1 = [\alpha - \delta_1, \alpha + \delta_1]$. Define

$$M = \frac{\max_{x\in\mathcal{B}_1} |f''(x)|}{2 \min_{x\in\mathcal{B}_1} |f'(x)|}$$

and pick x_0 sufficiently close to α such that

(i) $|x_0 - \alpha| = \delta_0 < \delta_1$;

(ii) $M\delta_0 < 1$.

The definition of M and (*) imply

$$|x_{n+1} - \alpha| \le M |x_n - \alpha|^2.$$

Comparing the above to (2.3) implies that if $\{x_n\}$ converges, then the order of convergence is 2. We must still show that (a) it converges and (b) it converges to α.

By (i) and (ii), we have $M|x_0 - \alpha| < 1$. Then it is easy to obtain the following via induction,

$$|x_n - \alpha| \le \frac{1}{M} \left(M|x_0 - \alpha|\right)^{2^n},$$

which shows both (a) and (b) and completes the proof. □

Theorem 2.15. A continuous function $f : [a, b] \to [c, d]$ is bijective if and only if it is strictly monotonic.

Theorem 2.16. If a $\mathcal{C}^2$ function $f : \mathbb{R} \to \mathbb{R}$ satisfies $f(\alpha) = 0$, $f' > 0$ and $f'' > 0$, then α is the only root of f and, $\forall x_0 \in \mathbb{R}$, the sequence of iterates $\{x_n\}$ in the Newton's method converges quadratically to α.

Proof. By Theorem 2.15, f is a bijection since f is continuous and strictly monotonic. With 0 in its range, f must have a unique root. When proving Theorem 2.14, we had

$$x_{n+1} - \alpha = (x_n - \alpha)^2 \frac{f''(\xi)}{2f'(x_n)}. \tag{2.9}$$

Then $f' > 0$ and $f'' > 0$ further imply that $x_{n+1} > \alpha$ for all $n > 0$. f being strictly increasing implies that $f(x_n) > f(\alpha) = 0$ for all $n > 0$. By the definition of Newton's method, $x_{n+1} - \alpha = x_n - \alpha - \frac{f(x_n)}{f'(x_n)}$, hence the sequence $\{x_n - \alpha : n > 0\}$ is strictly monotonically decreasing with 0 as a lower bound. By Theorem 2.11 it converges.

Suppose the sequence $\{x_n\}$ converges to $\alpha + c$ for some fixed $c > 0$. Define $\delta = \frac{f(\alpha+c)}{f'(\alpha+c)}$. The Taylor series of $f(\alpha + c)$ expanded at α and $f'(x) > 0$ imply $\delta > 0$. Because the Newton iteration $\{x_n\}$ converges, we have

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall n > N, |x_n - x_{n+1}| = \left|\frac{f(x_n)}{f'(x_n)}\right| < \epsilon,$$

which holds in particular for $\epsilon = \frac{1}{2}\delta$. On the other hand,

$$\begin{aligned} \left| x_n - x_{n+1} - \frac{f(\alpha+c)}{f'(\alpha+c)} \right| &\ge \left| |x_n - x_{n+1}| - \left|\frac{f(\alpha+c)}{f'(\alpha+c)}\right| \right| \\ &> \delta - \frac{1}{2}\delta = \epsilon. \end{aligned}$$

This contradicts the assumption that the Newton iteration $\{x_n\}$ converges to $\alpha + c$.

The quadratic convergence rate can be proved by an induction using (2.9), as in Theorem 2.14. □

Definition 2.17. Let $\mathcal{V}$ be a vector space. A subset $\mathcal{U} \subseteq \mathcal{V}$ is a *convex set* iff

$$\forall x, y \in \mathcal{U}, \forall t \in (0, 1), \qquad tx + (1-t)y \in \mathcal{U}. \tag{2.10}$$

A function $f : \mathcal{U} \to \mathbb{R}$ is *convex* iff

$$\begin{aligned} &\forall x, y \in \mathcal{U}, \forall t \in (0, 1), \\ &f(tx + (1-t)y) \le tf(x) + (1-t)f(y). \end{aligned} \tag{2.11}$$

In particular, f is *strictly convex* if we replace "$\le$" with "$<$" in the above equation.

2.6 The secant method

Algorithm 2.18. The *secant method* finds a root of $f : \mathbb{R} \to \mathbb{R}$ near initial guesses x_0, x_1 by the iteration

$$x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \qquad n \in \mathbb{N}^+. \tag{2.12}$$

Input: $f : \mathbb{R} \to \mathbb{R},\ x_0 \in \mathbb{R},\ x_1 \in \mathbb{R},$
$M \in \mathbb{N}^+,\ \delta \in \mathbb{R}^+,\ \epsilon \in \mathbb{R}^+$
Preconditions : $f \in \mathcal{C}^2$; x_0, x_1 are sufficiently close to a root of f
Output: x_n, x_{n-1}, k
Postconditions: $|f(x_n)| < \epsilon$ or $|x_n - x_{n-1}| < \delta$ or $k = M$

$x_n \leftarrow x_1$
$x_{n-1} \leftarrow x_0$
$u \leftarrow f(x_n)$
$v \leftarrow f(x_{n-1})$
for $k = 2 : M$ **do**
  if $|u| > |v|$ **then**
    $x_n \leftrightarrow x_{n-1}$
    $u \leftrightarrow v$
  end
  $s \leftarrow \frac{x_n - x_{n-1}}{u - v}$
  $x_{n-1} \leftarrow x_n$
  $v \leftarrow u$
  $x_n \leftarrow x_n - u \times s$
  $u \leftarrow f(x_n)$
  if $|x_n - x_{n-1}| < \delta$ **or** $|u| < \epsilon$ **then**
    break
  end
end

Definition 2.19. The sequence $\{F_n\}$ of *Fibonacci numbers* is defined as

$$F_0 = 0,\ F_1 = 1, \qquad F_{n+1} = F_n + F_{n-1}. \tag{2.13}$$

Theorem 2.20 (Binet's formula)**.** Denote the golden ratio by $r_0 = \frac{1+\sqrt{5}}{2} \approx 1.618$ and let $r_1 = 1 - r_0 = \frac{1-\sqrt{5}}{2}$, then

$$F_n = \frac{r_0^n - r_1^n}{\sqrt{5}}. \tag{2.14}$$

Corollary 2.21. The ratios r_0, r_1 in Theorem 2.20 satisfy

$$F_{n+1} = r_0 F_n + r_1^n. \tag{2.15}$$

Proof. This follows from (2.14) and values of r_0 and r_1. □

Lemma 2.22 (Error relation of the secant method)**.** For the secant method (2.12), there exist ξ_n between x_{n-1} and x_n and ζ_n between $\min(x_{n-1}, x_n, \alpha)$ and $\max(x_{n-1}, x_n, \alpha)$ such that

$$x_{n+1} - \alpha = (x_n - \alpha)(x_{n-1} - \alpha) \frac{f''(\zeta_n)}{2f'(\xi_n)}. \tag{2.16}$$

Proof. Define a divided difference as

$$f[a,b] = \frac{f(a)-f(b)}{a-b}. \tag{2.17}$$

Then it takes some algebra to show that the formula (2.12) is equivalent to

$$x_{n+1}-\alpha = (x_n-\alpha)(x_{n-1}-\alpha)\frac{\frac{f[x_{n-1},x_n]-f[x_n,\alpha]}{x_{n-1}-\alpha}}{f[x_{n-1},x_n]}. \tag{2.18}$$

By (2.17) and the mean value theorem (Theorem 0.35), there exists ξ_n between x_{n-1} and x_n such that

$$f[x_{n-1},x_n] = f'(\xi_n). \tag{2.19}$$

Define a function $g(x) := f[x, x_n]$, apply the mean value theorem to $g(x)$, and we have

$$\frac{f[x_{n-1},x_n]-f[x_n,\alpha]}{x_{n-1}-\alpha} = g'(\beta) \tag{2.20}$$

for some β between x_{n-1} and α. Compute the derivative of $g'(\beta)$ from (2.17), use the Lagrangian remainder Theorem 0.48, and we have

$$\frac{f[x_{n-1},x_n]-f[x_n,\alpha]}{x_{n-1}-\alpha} = \frac{f''(\zeta_n)}{2} \tag{2.21}$$

for some ζ_n between $\min(x_{n-1},x_n,\alpha)$ and $\max(x_{n-1},x_n,\alpha)$. The proof is completed by substituting (2.19) and (2.21) into (2.18). □

Theorem 2.23 (Convergence of the secant method)**.** Consider a $\mathcal{C}^2$ function $f:\mathcal{B}\to\mathbb{R}$ on $\mathcal{B}=[\alpha-\delta,\alpha+\delta]$ satisfying $f(\alpha)=0$ and $f'(\alpha)\neq 0$. If both x_0 and x_1 are chosen sufficiently close to α and $f''(\alpha)\neq 0$, then the iterates $\{x_n\}$ in the secant method converges to the root α with order $p=\frac{1}{2}(1+\sqrt{5})\approx 1.618$.

Proof. The continuity of f' and the assumption $f'(\alpha)\neq 0$ yield

$$\exists \delta_1\in(0,\delta) \text{ s.t. } \forall x\in\mathcal{B}_1,\ f'(x)\neq 0$$

where $\mathcal{B}_1=[\alpha-\delta_1,\alpha+\delta_1]$. Define $E_i=|x_i-\alpha|$,

$$M=\frac{\max_{x\in\mathcal{B}_1}|f''(x)|}{2\min_{x\in\mathcal{B}_1}|f'(x)|},$$

and we have from Lemma 2.22

$$ME_{n+1}\le ME_nME_{n-1}.$$

Pick x_0,x_1 such that

(i) $E_0<\delta$, $E_1<\delta$;

(ii) $\max(ME_1,ME_0)=\eta<1$,

then an induction by the above equation shows that $E_n<\delta$, $ME_n<\eta$. To prove convergence, we write $ME_0<\eta$, $ME_1<\eta$, $ME_2<ME_1ME_0<\eta^2$, $ME_3<ME_2ME_1<\eta^3$, $\cdots$, $ME_{n+1}<ME_nME_{n-1}<\eta^{q_n+q_{n-1}}=\eta^{q_{n+1}}$, i.e.

$$E_n<B_n:=\frac{1}{M}\eta^{q_n}.$$

$\{q_n\}$ is a Fibonacci sequence starting from $q_0=1, q_1=1$. By Theorem 2.20, as $n\to\infty$ we have $q_n\to\frac{1.618^{n+1}}{\sqrt{5}}$ since $|r_1|\approx 0.618<1$. Hence $\lim_{n\to\infty}E_n=0$.

To guestimate the convergence rate, we first examine the rate at which the upper bounds $\{B_n\}$ decrease:

$$\frac{B_{n+1}}{B_n^{r_0}}=\frac{\frac{1}{M}\eta^{q_{n+1}}}{\left(\frac{1}{M}\right)^{r_0}\eta^{r_0q_n}}=M^{r_0-1}\eta^{q_{n+1}-r_0q_n}\le M^{r_0-1}\eta^{-1}$$

where $q_{n+1}-r_0q_n=r_1^{n+1}>-1$.

To prove convergence rates, we define

$$m_n:=\left|\frac{f''(\zeta_n)}{2f'(\xi_n)}\right|,\qquad m_\alpha:=\left|\frac{f''(\alpha)}{2f'(\alpha)}\right|, \tag{2.22}$$

where ζ_n and ξ_n are the same as those in Lemma 2.22. By induction, we have

$$\begin{aligned}E_n&=E_1^{F_n}E_0^{F_{n-1}}m_1^{F_{n-1}}\cdots m_{n-1}^{F_1},\\ E_{n+1}&=E_1^{F_{n+1}}E_0^{F_n}m_1^{F_n}\cdots m_{n-1}^{F_2}m_n^{F_1},\end{aligned}$$

where F_n is a Fibonacci number as in Definition 2.19. Then

$$\begin{aligned}\frac{E_{n+1}}{E_n^{r_0}}=&E_1^{F_{n+1}-r_0F_n}E_0^{F_n-r_0F_{n-1}}m_1^{F_n-r_0F_{n-1}}m_2^{F_{n-1}-r_0F_{n-2}}\\ &\cdots m_{n-2}^{F_3-r_0F_2}m_{n-1}^{F_2-r_0F_1}m_n^{F_1}\\ =&E_1^{r_1^n}E_0^{r_1^{n-1}}m_1^{r_1^{n-1}}m_2^{r_1^{n-2}}\cdots m_{n-1}^{r_1^1}m_n^1,\end{aligned} \tag{2.23}$$

where the second step follows from Corollary 2.21. (2.22) and the convergence we just proved yield

$$\lim_{n\to+\infty}m_n=m_\alpha, \tag{2.24}$$

which means

$$\forall\epsilon>0,\ \exists N\in\mathbb{N} \text{ s.t. } \forall n>N, m_n\in(m_\alpha-\epsilon,m_\alpha+\epsilon). \tag{2.25}$$

We define

$$\begin{aligned}A&:=E_1^{r_1^n}\cdot E_0^{r_1^{n-1}}m_1^{r_1^{n-1}}\cdot m_2^{r_1^{n-2}}\cdots m_{N-1}^{r_1^{n-N+1}}\\ B&:=m_N^{r_1^{n-N}}\cdot m_{N+1}^{r_1^{n-N-1}}\cdots m_{n-1}^{r_1^1}\cdot m_n^1\end{aligned}$$

so that $\frac{E_{n+1}}{E_n^{r_0}}=AB$. Since $|r_1|<1$, we have $\lim_{n\to\infty}A=1$. As for B, we have from (2.25)

$$B\le(m_\alpha+\epsilon)^{1+r_1^1+r_1^2+\cdots+r_1^{n-N}},$$

and then

$$\begin{aligned}\lim_{n\to\infty}\frac{E_{n+1}}{E_n^{r_0}}&=\lim_{n\to\infty}A\lim_{n\to\infty}B\\ &=\lim_{n\to\infty}B\le(m_\alpha)^{\frac{1}{1-r_1}}=(m_\alpha)^{\frac{1}{r_0}}.\end{aligned}$$

The proof is then completed by Definition 2.9. □

Corollary 2.24. Consider solving $f(x) = 0$ near a root α. Let m and sm be the time to evaluate $f(x)$ and $f'(x)$ respectively. The minimum time to obtain the desired absolute accuracy ϵ with Newton's method and the secant method are respectively

$$T_N = (1+s)m\lceil \log_2 K \rceil, \tag{2.26}$$
$$T_S = m\lceil \log_{r_0} K \rceil, \tag{2.27}$$

where $r_0 = \frac{1+\sqrt{5}}{2}$, $c = \left|\frac{f''(\alpha)}{2f'(\alpha)}\right|$,

$$K = \frac{\log c\epsilon}{\log c|x_0 - \alpha|}, \tag{2.28}$$

and $\lceil \cdot \rceil$ denotes the rounding-up operator, i.e. it rounds towards $+\infty$.

Proof. We showed $|x_n - \alpha| \le \frac{1}{M}\left(M|x_0-\alpha|\right)^{2^n}$ in proving Theorem 2.14. Denote $E_n = |x_n - \alpha|$, we have

$$ME_n \le (ME_0)^{2^n}.$$

Let $i \in \mathbb{N}^+$ denote the smallest number of iterations such that the desired accuracy ϵ is satisfied, i.e. $(ME_0)^{2^i} \le M\epsilon$. When ϵ is sufficiently small, $M \to c$. Hence we have

$$i = \lceil \log_2 K \rceil.$$

For each iteration, Newton's method incurs one function evaluation and one derivative evaluation, which cost time m and sm, respectively. Therefore (2.26) holds.

For the secant method, assume $ME_0 \ge ME_1$. By the proof of Theorem 2.23, we have

$$ME_n \le (ME_0)^{r_0^{n+1}/\sqrt{5}}.$$

Let $j \in \mathbb{N}^+$ denote the smallest number of iterations such that the desired accuracy ϵ is satisfied, i.e. $r_0^j \le \frac{\sqrt{5}}{r_0}K$. Hence

$$j = \left\lceil \log_{r_0} K + \log_{r_0}\frac{\sqrt{5}}{r_0} \right\rceil \le \lceil \log_{r_0} K \rceil + 1.$$

Since the first two values x_0 and x_1 are given in the secant method, the least number of iterations is $\lceil \log_{r_0} K \rceil$ (compare to Newton's method!). Finally, only the function value $f(x_n)$ needs to be evaluated per iteration because $f(x_{n-1})$ has already been evaluated in the previous iteration. □

2.7 Fixed-point iterations

Definition 2.25. A *fixed point* of a function g is an independent parameter of g satisfying $g(\alpha) = \alpha$.

Example 2.2. A fixed point of $f(x) = x^2 - 3x + 4$ is $x = 2$.

Lemma 2.26. If $g : [a, b] \to [a, b]$ is continuous, then g has at least one fixed point in $[a, b]$.

Proof. The function $f(x) = g(x) - x$ satisfies $f(a) \ge 0$ and $f(b) \le 0$. The proof is then completed by the intermediate value theorem (Theorem 0.32). □

Exercise 2.3. Let $A = [-1, 0) \cup (0, 1]$. Give an example of a continuous function $g : A \to A$ that does not have a fixed point. Give an example of a continuous function $f : \mathbb{R} \to \mathbb{R}$ that does not have a fixed point.

Theorem 2.27 (Brouwer's fixed point)**.** Any function $f : \mathbb{D}^n \to \mathbb{D}^n$ with

$$\mathbb{D}^n := \{\mathbf{x} \in \mathbb{R}^n : \|x\| \le 1\}$$

has a fixed point.

Exercise 2.4. Take two pieces of the same-sized paper and lay one on top of the other. Every point on the top sheet of paper is associated with some point right below it on the bottom sheet. Crumple the top sheet into a ball without ripping it. Place the crumpled ball on top of (and simultaneously within the realm of) the bottom sheet of paper. Use Theorem 2.27 to prove that there always exists some point in the crumpled ball that sits above the same point it sat above prior to crumpling.

Example 2.5. Take a map of your country C and place it on the ground of your room. Let f be the function assigning to each point in your country the point on the map corresponding to it. Then f can be considered as a continuous function $C \to C$. If C is homeomorphic to $\mathbb{D}^2$, then there must exist a point on the map that corresponds exactly to the point on the ground directly beneath it.

Definition 2.28. A *fixed-point iteration* is a method for finding a fixed point of g with a formula of the form

$$x_{n+1} = g(x_n), \qquad n \in \mathbb{N}. \tag{2.29}$$

Example 2.6. Newton's method is a fixed-point iteration.

Exercise 2.7. To calculate the square root of some positive real number a, we can formulate the problem as finding the root of $f(x) = x^2 - a$. For $a = 1$, the initial guess of $x_0 = 2$, and the three choices of $g_1(x) := x^2 + x - a$, $g_2(x) := \frac{a}{x}$, and $g_3(x) := \frac{1}{2}(x + \frac{a}{x})$, verify that g_1 diverges, g_2 oscillates, g_3 converges. The theorems in this section will explain why.

Definition 2.29. A function $f : [a, b] \to [a, b]$ is a *contraction* or *contractive mapping* on $[a, b]$ if

$$\exists \lambda \in [0, 1) \text{ s.t. } \forall x, y \in [a, b],\ |f(x) - f(y)| \le \lambda |x - y|. \tag{2.30}$$

Example 2.8. Any linear function $f(x) = \lambda x + c$ with $0 \le \lambda < 1$ is a contraction.

Theorem 2.30 (Convergence of contractions)**.** If $g(x)$ is a continuous contraction on $[a, b]$, then it has a unique fixed point α in $[a, b]$. Furthermore, the fixed-point iteration (2.29) converges to α for any choice $x_0 \in [a, b]$ and

$$|x_n - \alpha| \le \frac{\lambda^n}{1-\lambda}|x_1 - x_0|. \tag{2.31}$$

Proof. By Lemma 2.26, g has at least one fixed point in $[a,b]$. Suppose there are two distinct fixed points α and β, then $|\alpha-\beta| = |g(\alpha)-g(\beta)| \le \lambda|\alpha-\beta|$, which implies $|\alpha-\beta| \le 0$, i.e. the two fixed points are identical.

By Definition 2.29, $x_{n+1} = g(x_n)$ implies that all x_n's stay in $[a,b]$. To prove convergence,

$$|x_{n+1}-\alpha| = |g(x_n)-g(\alpha)| \le \lambda|x_n-\alpha|.$$

By induction and the triangle inequality,

$$\begin{aligned}|x_n-\alpha| &\le \lambda^n|x_0-\alpha|\\ &\le \lambda^n(|x_1-x_0|+|x_1-\alpha|)\\ &\le \lambda^n(|x_1-x_0|+\lambda|x_0-\alpha|).\end{aligned}$$

From the first and last right-hand sides (RHSs), we have $|x_0-\alpha| \le \frac{1}{1-\lambda}|x_1-x_0|$, which yields (2.31). □

Theorem 2.31. Consider $g:[a,b]\to[a,b]$. If $g\in\mathcal{C}^1[a,b]$ and $\lambda = \max_{x\in[a,b]}|g'(x)| < 1$, then g has a unique fixed point α in $[a,b]$. Furthermore, the fixed-point iteration (2.29) converges to α for any choice $x_0\in[a,b]$, the error bound (2.31) holds, and

$$\lim_{n\to\infty}\frac{x_{n+1}-\alpha}{x_n-\alpha} = g'(\alpha). \tag{2.32}$$

Proof. The mean value theorem (Theorem 0.35) implies that, for all $x,y\in[a,b]$, $|g(x)-g(y)| \le \lambda|x-y|$. Theorem 2.30 yields all the results except (2.32), which follows from

$$x_{n+1}-\alpha = g(x_n)-g(\alpha) = g'(\xi)(x_n-\alpha),$$

$\lim x_n = \alpha$, and the fact that ξ is between x_n and α. □

Corollary 2.32. Let α be a fixed point of $g:\mathbb{R}\to\mathbb{R}$ with $|g'(\alpha)|<1$ and $g\in\mathcal{C}^1(\mathcal{B})$ on $\mathcal{B} = [\alpha-\delta,\alpha+\delta]$ with some $\delta>0$. If x_0 is chosen sufficiently close to α, then the results of Theorem 2.30 hold.

Proof. Choose λ so that $|g'(\alpha)|<\lambda<1$. Choose $\delta_0\le\delta$ so that $\max_{x\in\mathcal{B}_0}|g'(x)|\le\lambda<1$ on $\mathcal{B}_0 = [\alpha-\delta_0,\alpha+\delta_0]$. Then $g(\mathcal{B}_0)\subset\mathcal{B}_0$ and applying Theorem 2.31 completes the proof. □

Corollary 2.33. Consider $g:[a,b]\to[a,b]$ with a fixed point $g(\alpha)=\alpha\in[a,b]$. The fixed-point iteration (2.29) converges to α with pth-order accuracy ($p>1$, $p\in\mathbb{N}$) for any choice $x_0\in[a,b]$ if

$$\begin{cases} g\in\mathcal{C}^p[a,b],\\ \forall k=1,2,\ldots,p-1,\ g^{(k)}(\alpha)=0,\\ g^{(p)}(\alpha)\ne 0.\end{cases} \tag{2.33}$$

Proof. By Corollary 2.32, the fixed-point iteration converges uniquely to α because $g'(\alpha)=0$. By the Taylor expansion of g at α, we have

$$\begin{aligned}E_{\text{abs}}(x_{n+1}) &:= |x_{n+1}-\alpha| = |g(x_n)-g(\alpha)|\\ &= \left|\sum_{i=1}^{p-1}\frac{(x_n-\alpha)^i}{i!}g^{(i)}(\alpha) + \frac{(x_n-\alpha)^p}{p!}g^{(p)}(\xi)\right|\end{aligned}$$

for some $\xi\in[a,b]$. Since $g^{(p)}$ is continuous on $[a,b]$, Theorem 0.31 implies that $g^{(p)}$ is bounded on $[a,b]$. Hence there exists a constant M such that $E_{\text{abs}}(x_{n+1}) < ME^p_{\text{abs}}(x_n)$. □

Example 2.9. The following method has third-order convergence for computing $\sqrt{R}$:

$$x_{n+1} = \frac{x_n(x_n^2+3R)}{3x_n^2+R}.$$

First, $\sqrt{R}$ is the fixed point of $F(x) = \frac{x(x^2+3R)}{3x^2+R}$:

$$F\left(\sqrt{R}\right) = \frac{\sqrt{R}(R+3R)}{3R+R} = \sqrt{R}.$$

Second, the derivatives of $F(x)$ are

n	$F^{(n)}(x)$	$F^{(n)}(\sqrt{R})$
1	$\frac{3(x^2-R)^2}{(3x^2+R)^2}$	0
2	$\frac{48Rx(x^2-R)}{(3x^2+R)^3}$	0
3	$\frac{-48R(9x^4-18Rx^2+R^2)}{(3x^2+R)^4}$	$\frac{-48R(-8R^2)}{(4R)^4} = \frac{3}{2R} \ne 0$

The rest follows from Corollary 2.33.

2.8 Problems

2.8.1 Theoretical questions

I. Consider the bisection method starting with the initial interval $[1.5, 3.5]$. In the following questions "the interval" refers to the bisection interval whose width changes across different loops.

- What is the width of the interval at the nth step?
- What is the maximum possible distance between the root r and the midpoint of the interval?

II. In using the bisection algorithm with its initial interval as $[a_0,b_0]$, we want to determine the root with its *relative* error no greater than ϵ. Assume $a_0>0$. Prove that the number of steps n must satisfy

$$n \ge \frac{\log(b_0-a_0)-\log\epsilon-\log a_0}{\log 2} - 1.$$

III. If the bisection method is used in single precision FPNs of IEEE 754 starting with the interval $[128,129]$, can we compute the root with absolute accuracy $<10^{-6}$? Why?

IV. Perform four iterations of Newton's method for the polynomial equation $p(x) = 4x^3-2x^2+3 = 0$ with the starting point $x_0=-1$. Use a hand calculator and organize results of the iterations in a table.

V. Consider a variation of Newton's method in which only the derivative at x_0 is used,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}.$$

Find C and s such that

$$e_{n+1} = Ce_n^s.$$

VI. Within $(-\frac{\pi}{2}, \frac{\pi}{2})$, will the iteration $x_{n+1} = \tan^{-1} x_n$ converge?

VII. Let $p > 1$. What is the value of the following continued fraction?

$$x = \frac{1}{p + \frac{1}{p + \frac{1}{p + \dots}}}$$

Prove that the sequence of values converges. (Hint: this can be interpreted as $x = \lim_{n\to\infty} x_n$, where $x_1 = \frac{1}{p}$, $x_2 = \frac{1}{p+\frac{1}{p}}$, $x_3 = \frac{1}{p+\frac{1}{p+\frac{1}{p}}}$, ..., and so forth. Formulate x as a fixed point of some function.)

VIII. What happens in problem II if $a_0 < 0 < b_0$? Derive an inequality of the number of steps similar to that in II. In this case, is the relative error still an appropriate measure?

IX. (∗) Consider solving $f(x) = 0$ ($f \in \mathcal{C}^{k+1}$) by Newton's method with the starting point x_0 close to a root of multiplicity k. Note that α is a zero of multiplicity k of the function f iff

$$f^{(k)}(\alpha) \neq 0; \qquad \forall i < k, \quad f^{(i)}(\alpha) = 0.$$

- How can a multiple zero be detected by examining the behavior of the points $(x_n, f(x_n))$?
- Prove that if r is a zero of multiplicity k of the function f, then quadratic convergence in Newton's iteration will be restored by making this modification:

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

X. (∗) Analysis of the secant method for a root of multiplicity k by assuming it converges.

- Prove that if r is a zero of multiplicity $k > 1$ of the function f, the secant method only has linear convergence.
- Use the same argument to show that the convergence rate of the secant method is $\frac{\sqrt{5}+1}{2}$.

2.8.2 Programming assignments

A. Implement the bisection method and test your program on these functions and intervals.

- $x^{-1} - \tan x$ on $[0, \frac{\pi}{2}]$,
- $x^{-1} - 2^x$ on $[0, 1]$,
- $2^{-x} + e^x + 2\cos x - 6$ on $[1, 3]$,
- $(x^3 + 4x^2 + 3x + 5)/(2x^3 - 9x^2 + 18x - 2)$ on $[0, 4]$.

B. Implement Newton's method to solve the equation $x = \tan x$. Find the roots near 4.5 and 7.7.

C. Implement the secant method and test your program on the following functions and initial values.

- $\sin(x/2) - 1$ with $x_0 = 0, x_1 = \frac{\pi}{2}$,
- $e^x - \tan x$ with $x_0 = 1, x_1 = 1.4$,
- $x^3 - 12x^2 + 3x + 1$ with $x_0 = 0, x_1 = -0.5$.

You should play with other initial values and (if you get different results) think about the reasons.