1

Solution of equations by iteration

1.1 Introduction

Equations of various kinds arise in a range of physical applications and a substantial body of mathematical research is devoted to their study. Some equations are rather simple: in the early days of our mathematical education we all encountered the single *linear* equation $ax+b=0$, where a and b are real numbers and $a \neq 0$, whose solution is given by the formula $x = -b/a$. Many equations, however, are *nonlinear:* a simple example is $ax^2 + bx + c = 0$, involving a quadratic polynomial with real coefficients a, b, c, and $a \neq 0$. The two solutions to this equation, labelled x_1 and x_2, are found in terms of the coefficients of the polynomial from the familiar formulae

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \qquad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \tag{1.1}$$

It is less likely that you have seen the more intricate formulae for the solution of cubic and quartic polynomial equations due to the sixteenth century Italian mathematicians Niccolo Fontana Tartaglia (1499–1557) and Lodovico Ferrari (1522–1565), respectively, which were published by Girolamo Cardano (1501–1576) in 1545 in his *Artis magnae sive de regulis algebraicis liber unus.* In any case, if you have been led to believe that similar expressions involving radicals (roots of sums of products of coefficients) will supply the solution to any polynomial equation, then you should brace yourself for a surprise: no such closed formula exists for a general polynomial equation of degree n when $n \geq 5$. It transpires that for each $n \geq 5$ there exists a polynomial equation of degree n with

integer coefficients which cannot be solved in terms of radicals;[1] such is, for example, $x^5 - 4x - 2 = 0$.

Since there is no general formula for the solution of polynomial equations, no general formula will exist for the solution of an arbitrary nonlinear equation of the form $f(x) = 0$ where f is a continuous real-valued function. How can we then decide whether or not such an equation possesses a solution in the set of real numbers, and how can we find a solution?

The present chapter is devoted to the study of these questions. Our goal is to develop simple numerical methods for the approximate solution of the equation $f(x) = 0$ where f is a real-valued function, defined and continuous on a bounded and closed interval of the real line. Methods of the kind discussed here are iterative in nature and produce sequences of real numbers which, in favourable circumstances, converge to the required solution.

1.2 Simple iteration

Suppose that f is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line. It will be tacitly assumed throughout the chapter that $a < b$, so that the interval is nonempty. We wish to find a *real number* $\xi \in [a, b]$ such that $f(\xi) = 0$. If such ξ exists, it is called a **solution** to the equation $f(x) = 0$.

Even some relatively simple equations may fail to have a solution in the set of real numbers. Consider, for example,

$$f\colon\ x \mapsto x^2 + 1\,.$$

Clearly $f(x) = 0$ has no solution in any interval $[a, b]$ of the real line. Indeed, according to (1.1), the quadratic polynomial x^2+1 has two roots: $x_1 = \sqrt{-1} = \imath$ and $x_2 = -\sqrt{-1} = -\imath$. However, these belong to the set of imaginary numbers and are therefore excluded by our definition of solution which only admits *real* numbers. In order to avoid difficulties of this kind, we begin by exploring the existence of solutions to the equation $f(x) = 0$ in the set of real numbers. Our first result in this direction is rather simple.

[1] This result was proved in 1824 by the Norwegian mathematician Niels Henrik Abel (1802–1829), and was further refined in the work of Evariste Galois (1811–1832) who clarified the circumstances in which a closed formula may exist for the solution of a polynomial equation of degree n in terms of radicals.

Theorem 1.1 *Let f be a real-valued function, defined and continuous on a bounded closed interval $[a,b]$ of the real line. Assume, further, that $f(a)f(b) \leq 0$; then, there exists ξ in $[a,b]$ such that $f(\xi) = 0$.*

Proof If $f(a) = 0$ or $f(b) = 0$, then $\xi = a$ or $\xi = b$, respectively, and the proof is complete. Now, suppose that $f(a)f(b) \neq 0$. Then, $f(a)f(b) < 0$; in other words, 0 belongs to the open interval whose endpoints are $f(a)$ and $f(b)$. By the Intermediate Value Theorem (Theorem A.1), there exists ξ in the open interval (a,b) such that $f(\xi) = 0$. □

To paraphrase Theorem 1.1, if a continuous function f has opposite signs at the endpoints of the interval $[a,b]$, then the equation $f(x) = 0$ has a solution in (a,b). The converse statement is, of course, false. Consider, for example, a continuous function defined on $[a,b]$ which changes sign in the open interval (a,b) an even number of times, with $f(a)f(b) \neq 0$; then, $f(a)f(b) > 0$ even though $f(x) = 0$ has solutions inside $[a,b]$. Of course, in the latter case, there exist an even number of subintervals of (a,b) at the endpoints of each of which f *does* have opposite signs. However, finding such subintervals may not always be easy.

To illustrate this last point, consider the rather pathological function

$$f\colon x \mapsto \frac{1}{2} - \frac{1}{1 + M|x - 1.05|}\,, \tag{1.2}$$

depicted in Figure 1.1 for x in the closed interval $[0.8, 1.8]$ and $M = 200$. The solutions $x_1 = 1.05 - (1/M)$ and $x_2 = 1.05 + (1/M)$ to the equation $f(x) = 0$ are only a distance $2/M$ apart and, for large and positive M, locating them computationally will be a challenging task.

Remark 1.1 *If you have access to the mathematical software package Maple, plot the function f by typing*

```
plot(1/2-1/(1+200*abs(x-1.05)), x=0.8..1.8, y=-0.5..0.6);
```

at the Maple command line, and then repeat this experiment by choosing $M = 2000$, 20000, 200000, 2000000, and 20000000 in place of the number 200. What do you observe? For the last two values of M, replot the function f for x in the subinterval $[1.04999, 1.05001]$. ◇

An alternative sufficient condition for the existence of a solution to the equation $f(x) = 0$ is arrived at by rewriting it in the equivalent form $x - g(x) = 0$ where g is a certain real-valued function, defined

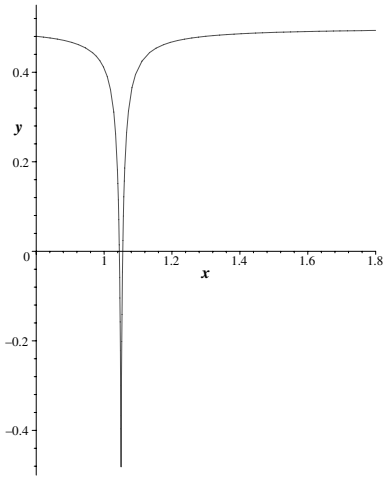

Fig. 1.1. Graph of the function f: $x \mapsto \frac{1}{2} - \frac{1}{1+200|x-1.05|}$ for $x \in [0.8, 1.8]$.

and continuous on $[a, b]$; the choice of g and its relationship with f will be clarified below through examples. Upon such a transformation the problem of solving the equation $f(x) = 0$ is converted into one of finding ξ such that $\xi - g(\xi) = 0$.

Theorem 1.2 (Brouwer's Fixed Point Theorem) *Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and let $g(x) \in [a, b]$ for all $x \in [a, b]$. Then, there exists ξ in $[a, b]$ such that $\xi = g(\xi)$; the real number ξ is called a* **fixed point of the function** *g.*

Proof Let $f(x) = x - g(x)$. Then, $f(a) = a - g(a) \leq 0$ since $g(a) \in [a, b]$ and $f(b) = b - g(b) \geq 0$ since $g(b) \in [a, b]$. Consequently, $f(a)f(b) \leq 0$, with f defined and continuous on the closed interval $[a, b]$. By Theorem 1.1 there exists $\xi \in [a, b]$ such that $0 = f(\xi) = \xi - g(\xi)$. □

Figure 1.2 depicts the graph of a function $x \mapsto g(x)$, defined and continuous on a closed interval $[a, b]$ of the real line, such that $g(x)$ belongs to $[a, b]$ for all x in $[a, b]$. The function g has three fixed points in the interval $[a, b]$: the x-coordinates of the three points of intersection of the graph of g with the straight line $y = x$.

Of course, any equation of the form $f(x) = 0$ can be rewritten in the

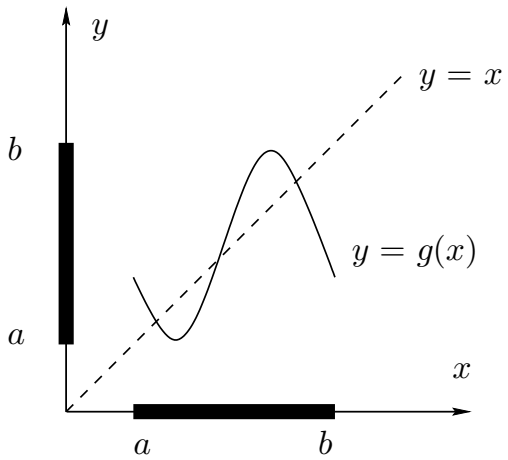

Fig. 1.2. Graph of a function g, defined and continuous on the interval $[a, b]$, which maps $[a, b]$ into itself; g has three fixed points in $[a, b]$: the x-coordinates of the three points of intersection of the graph of g with $y = x$.

equivalent form of $x = g(x)$ by letting $g(x) = x + f(x)$. While there is no guarantee that the function g, so defined, will satisfy the conditions of Theorem 1.2, there are many alternative ways of transforming $f(x) = 0$ into $x = g(x)$, and we only have to find one such rearrangement with g continuous on $[a, b]$ and such that $g(x) \in [a, b]$ for all $x \in [a, b]$. Sounds simple? Fine. Take a look at the following example.

Example 1.1 *Consider the function f defined by $f(x) = \mathrm{e}^x - 2x - 1$ for $x \in [1, 2]$. Clearly, $f(1) < 0$ and $f(2) > 0$. Thus we deduce from Theorem 1.1 the existence of ξ in $[1, 2]$ such that $f(\xi) = 0$.*

In order to relate this example to Theorem 1.2, let us rewrite the equation $f(x) = 0$ in the equivalent form $x - g(x) = 0$, where the function g is defined on the interval $[1, 2]$ by $g(x) = \ln(2x + 1)$; here (and throughout the book) ln means $\log_{\mathrm{e}}$. As $g(1) \in [1, 2]$, $g(2) \in [1, 2]$ and g is monotonic increasing, it follows that $g(x) \in [1, 2]$ for all $x \in [1, 2]$, showing that g satisfies the conditions of Theorem 1.2. Thus, again, we deduce the existence of $\xi \in [1, 2]$ such that $\xi - g(\xi) = 0$ or, equivalently, $f(\xi) = 0$.

We could have also rewritten our equation as $x = (\mathrm{e}^x - 1)/2$. However, the associated function g: $x \mapsto (\mathrm{e}^x - 1)/2$ does not map the interval $[1, 2]$ into itself, so Theorem 1.2 cannot then be applied. ◇

Although the ability to verify the existence of a solution to the equation $f(x) = 0$ is important, none of what has been said so far provides a *method* for solving this equation. The following definition is a first step in this direction: it will lead to the construction of an algorithm for computing an approximation to the fixed point ξ of the function g, and will thereby supply an approximate solution to the equivalent equation $f(x) = 0$.

Definition 1.1 *Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and assume that $g(x) \in [a, b]$ for all $x \in [a, b]$. Given that $x_0 \in [a, b]$, the recursion defined by*

$$x_{k+1} = g(x_k)\,, \qquad k = 0, 1, 2, \ldots\,, \tag{1.3}$$

is called a **simple iteration***; the numbers x_k, $k \geq 0$, are referred to as* **iterates***.*

If the sequence (x_k) defined by (1.3) converges, the limit must be a fixed point of the function g, since g is continuous on a closed interval. Indeed, writing $\xi = \lim_{k\to\infty} x_k$, we have that

$$\xi = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} g(x_k) = g\left(\lim_{k\to\infty} x_k\right) = g(\xi)\,, \tag{1.4}$$

where the second equality follows from (1.3) and the third equality is a consequence of the continuity of g.

A sufficient condition for the convergence of the sequence (x_k) is provided by our next result which represents a refinement of Brouwer's Fixed Point Theorem, under the additional assumption that the mapping g is a contraction.

Definition 1.2 (Contraction) *Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line. Then, g is said to be a* **contraction** *on $[a, b]$ if there exists a constant L such that $0 < L < 1$ and*

$$|g(x) - g(y)| \leq L|x - y| \quad \forall\, x, y \in [a, b]\,. \tag{1.5}$$

Remark 1.2 *The terminology 'contraction' stems from the fact that when (1.5) holds with $0 < L < 1$, the distance $|\,g(x) - g(y)\,|$ between the images of the points x, y is (at least $1/L$ times)* smaller *than the distance*

$|x - y|$ *between* x *and* y. *More generally, when* L *is any positive real number, (1.5) is referred to as a* **Lipschitz condition**.[1]

Armed with Definition 1.2, we are now ready to state the main result of this section.

Theorem 1.3 (Contraction Mapping Theorem) *Let* g *be a real-valued function, defined and continuous on a bounded closed interval* $[a, b]$ *of the real line, and assume that* $g(x) \in [a, b]$ *for all* $x \in [a, b]$. *Suppose, further, that* g *is a contraction on* $[a, b]$. *Then,* g *has a unique fixed point* ξ *in the interval* $[a, b]$. *Moreover, the sequence* (x_k) *defined by (1.3) converges to* ξ *as* $k \to \infty$ *for any starting value* x_0 *in* $[a, b]$.

Proof The existence of a fixed point ξ for g is a consequence of Theorem 1.2. The uniqueness of this fixed point follows from (1.5) by contradiction: for suppose that g has a second fixed point, η, in $[a, b]$. Then,

$$|\xi - \eta| = |g(\xi) - g(\eta)| \leq L|\xi - \eta|,$$

i.e., $(1 - L)|\xi - \eta| \leq 0$. As $1 - L > 0$, we deduce that $\eta = \xi$.

Let x_0 be any element of $[a, b]$ and consider the sequence (x_k) defined by (1.3). We shall prove that (x_k) converges to the fixed point ξ. According to (1.5) we have that

$$|x_k - \xi| = |g(x_{k-1}) - g(\xi)| \leq L|x_{k-1} - \xi|, \qquad k \geq 1,$$

from which we then deduce by induction that

$$|x_k - \xi| \leq L^k |x_0 - \xi|, \qquad k \geq 1. \tag{1.6}$$

As $L \in (0, 1)$, it follows that $\lim_{k\to\infty} L^k = 0$, and hence we conclude that $\lim_{k\to\infty} |x_k - \xi| = 0$. □

Let us illustrate the Contraction Mapping Theorem by an example.

Example 1.2 *Consider the equation* $f(x) = 0$ *on the interval* $[1, 2]$ *with* $f(x) = e^x - 2x - 1$, *as in Example 1.1. Recall from Example 1.1 that this equation has a solution,* ξ, *in the interval* $[1, 2]$, *and* ξ *is a fixed point of the function* g *defined on* $[1, 2]$ *by* $g(x) = \ln(2x + 1)$.

[1] Rudolf Otto Sigismund Lipschitz (14 May 1832, Königsberg, Prussia (now Kaliningrad, Russia) – 7 October 1903, Bonn, Germany) made important contributions to number theory, the theory of Bessel functions and Fourier series, the theory of ordinary and partial differential equations, and to analytical mechanics and potential theory.

Table 1.1. *The sequence* (x_k) *defined by (1.8).*

k	x_k
0	1.000000
1	1.098612
2	1.162283
3	1.201339
4	1.224563
5	1.238121
6	1.245952
7	1.250447
8	1.253018
9	1.254486
10	1.255323
11	1.255800

Now, the function g is defined and continuous on the interval $[1,2]$, and g is differentiable on $(1,2)$. Thus, by the Mean Value Theorem (Theorem A.3), for any x, y in $[1,2]$ we have that

$$|\, g(x) - g(y)\,| = |\, g'(\eta)(x-y)\,| = |g'(\eta)|\,|\, x-y\,| \tag{1.7}$$

for some η that lies between x and y and is therefore in the interval $[1,2]$. Further, $g'(x) = 2/(2x+1)$ and $g''(x) = -4/(2x+1)^2$. As $g''(x) < 0$ for all x in $[1,2]$, g' is monotonic decreasing on $[1,2]$. Hence $g'(1) \geq g'(\eta) \geq g'(2)$, *i.e.*, $g'(\eta) \in [2/5, 2/3]$. Thus we deduce from (1.7) that

$$|\, g(x) - g(y)\,| \leq L|\, x-y\,| \qquad \forall\, x, y \in [1,2]\,,$$

with $L = 2/3$. According to the Contraction Mapping Theorem, the sequence (x_k) defined by the simple iteration

$$x_{k+1} = \ln(2x_k + 1)\,, \qquad k = 0, 1, 2, \ldots, \tag{1.8}$$

converges to ξ for any starting value x_0 in $[1,2]$. Let us choose $x_0 = 1$, for example, and compute the next 11 iterates, say. The results are shown in Table 1.1. Even though we have carried six decimal digits, after 11 iterations only the first two decimal digits of the iterates x_k appear to have settled; thus it seems likely that $\xi = 1.26$ to two decimal digits. $\diamond$

You may now wonder how many iterations we should perform in (1.8)

to ensure that all six decimals have converged to their correct values. In order to answer this question, we need to carry out some analysis.

Theorem 1.4 *Consider the simple iteration (1.3) where the function g satisfies the hypotheses of the Contraction Mapping Theorem on the bounded closed interval $[a,b]$. Given $x_0 \in [a,b]$ and a certain tolerance $\varepsilon > 0$, let $k_0(\varepsilon)$ denote the smallest positive integer such that x_k is no more than ε away from the (unknown) fixed point ξ, i.e., $|x_k - \xi| \le \varepsilon$, for all $k \ge k_0(\varepsilon)$. Then,*

$$k_0(\varepsilon) \le \left[\frac{\ln |\, x_1 - x_0 \,| - \ln(\varepsilon(1-L))}{\ln(1/L)} \right] + 1 \,, \tag{1.9}$$

where, for a real number x, $[x]$ signifies the largest integer less than or equal to x.

Proof From (1.6) in the proof of Theorem 1.3 we know that

$$|x_k - \xi| \le L^k |x_0 - \xi| \,, \quad k \ge 1 \,.$$

Using this result with $k = 1$, we obtain

$$\begin{aligned} |x_0 - \xi| &= |x_0 - x_1 + x_1 - \xi| \\ &\le |x_0 - x_1| + |x_1 - \xi| \\ &\le |x_0 - x_1| + L|x_0 - \xi| \,. \end{aligned}$$

Hence

$$|x_0 - \xi| \le \frac{1}{1-L} |x_0 - x_1| \,.$$

By substituting this into (1.6) we get

$$|x_k - \xi| \le \frac{L^k}{1-L} |x_1 - x_0| \,. \tag{1.10}$$

Thus, in particular, $|x_k - \xi| \le \varepsilon$ provided that

$$L^k \frac{1}{1-L} |x_1 - x_0| \le \varepsilon \,.$$

On taking the (natural) logarithm of each side in the last inequality, we find that $|x_k - \xi| \le \varepsilon$ for all k such that

$$k \ge \frac{\ln |x_1 - x_0| - \ln(\varepsilon(1-L))}{\ln(1/L)} \,.$$

Therefore, the smallest integer $k_0(\varepsilon)$ such that $|x_k - \xi| \le \varepsilon$ for all

$k \geq k_0(\varepsilon)$ cannot exceed the expression on the right-hand side of the inequality (1.9). □

This result provides an upper bound on the maximum number of iterations required to ensure that the error between the kth iterate x_k and the (unknown) fixed point ξ is below the prescribed tolerance ε. Note, in particular, from (1.9), that if L is close to 1, then $k_0(\varepsilon)$ may be quite large for any fixed ε. We shall revisit this point later on in the chapter.

Example 1.3 *Now we can return to Example 1.2 to answer the question posed there about the maximum number of iterations required, with starting value $x_0 = 1$, to ensure that the last iterate computed is correct to six decimal digits.*

Letting $\varepsilon = 0.5 \times 10^{-6}$ and recalling from Example 1.2 that $L = 2/3$, the formula (1.9) yields $k_0(\varepsilon) \leq [32.778918] + 1$, so we have that $k_0(\varepsilon) \leq 33$. In fact, 33 is a somewhat pessimistic overestimate of the number of iterations required: computing the iterates x_k successively shows that already x_{25} is correct to six decimal digits, giving $\xi = 1.256431$. ◇

Condition (1.5) can be rewritten in the following equivalent form:

$$\left| \frac{g(x) - g(y)}{x - y} \right| \leq L \qquad \forall\, x, y \in [a, b]\,, \quad x \neq y\,,$$

with $L \in (0, 1)$, which can, in turn, be rephrased by saying that the absolute value of the slope of the function g does not exceed $L \in (0, 1)$. Assuming that g is a differentiable function on the open interval (a, b), the Mean Value Theorem (Theorem A.3) tells us that

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

for some η that lies between x and y and is therefore contained in the interval (a, b).

We shall therefore adopt the following assumption that is somewhat stronger than (1.5) but is easier to verify in practice:

$$\begin{aligned} &g \text{ is differentiable on } (a, b) \text{ and} \\ &\exists L \in (0, 1) \text{ such that } |g'(x)| \leq L \text{ for all } x \in (a, b)\,. \end{aligned} \tag{1.11}$$

Consequently, Theorem 1.3 still holds when (1.5) is replaced by (1.11).

We note that the requirement in (1.11) that g be differentiable is