An Introduction to Numerical Analysis

Endre Süli and David F. Mayers

University of Oxford

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface CMR 10/13 pt *System* LaTeX 2ε [TB]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

ISBN 0 521 81026 4 hardback
ISBN 0 521 00794 1 paperback

Contents

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

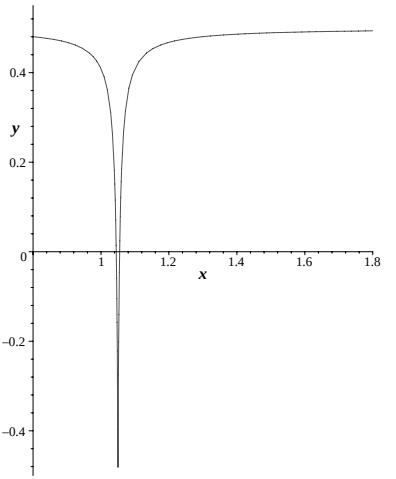

Fig. 1.1. Graph of the function $f\colon x \mapsto \frac{1}{2} - \frac{1}{1+200|x-1.05|}$ for $x \in [0.8, 1.8]$.

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

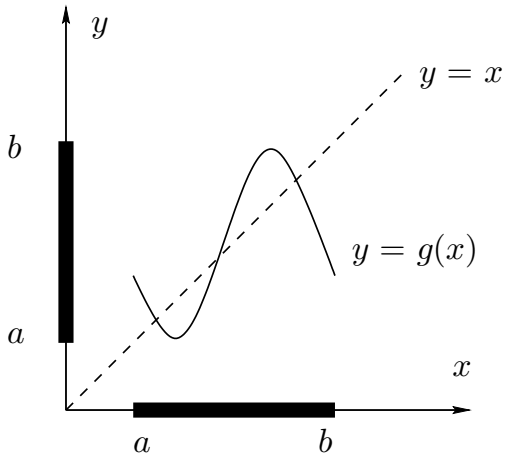

Fig. 1.2. Graph of a function g, defined and continuous on the interval $[a, b]$, which maps $[a, b]$ into itself; g has three fixed points in $[a, b]$: the x-coordinates of the three points of intersection of the graph of g with $y = x$.

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,

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

i.e., $(1-L)|\,\xi-\eta\,| \le 0$. 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

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

from which we then deduce by induction that

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

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* $f(x)=\mathrm{e}^x-2x-1$, *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. $\diamondsuit$

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., $|x_k - \xi| \leq \varepsilon$, for all $k \geq k_0(\varepsilon)$. Then,*

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

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

$$|x_k - \xi| \leq L^k \, |x_0 - \xi| \,, \quad k \geq 1 \,.$$

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

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

Hence

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

By substituting this into (1.6) we get

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

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

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

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

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

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

indeed more demanding than the Lipschitz condition (1.5): for example, $g(x) = |x|$ satisfies the Lipschitz condition on any closed interval of the real line, with $L = 1$, yet g is not differentiable at $x = 0$.[1]

Next we discuss a local version of the Contraction Mapping Theorem, where (1.11) is only assumed in a neighbourhood of the fixed point ξ rather than over the entire interval $[a, b]$.

Theorem 1.5 *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 $x \in [a, b]$. Let $\xi = g(\xi) \in [a, b]$ be a fixed point of g (whose existence is ensured by Theorem 1.2), and assume that g has a continuous derivative in some neighbourhood of ξ with $|g'(\xi)| < 1$. Then, the sequence (x_k) defined by $x_{k+1} = g(x_k)$, $k \geq 0$, converges to ξ as $k \to \infty$, provided that x_0 is sufficiently close to ξ.*

Proof By hypothesis, there exists $h > 0$ such that g' is continuous in the interval $[\xi - h, \xi + h]$. Since $|g'(\xi)| < 1$ we can find a smaller interval $I_\delta = [\xi - \delta, \xi + \delta]$, where $0 < \delta \leq h$, such that $|g'(x)| \leq L$ in this interval, with $L < 1$. To do so, take $L = \frac{1}{2}(1 + |g'(\xi)|)$ and then choose $\delta \leq h$ such that

$$|g'(x) - g'(\xi)| \leq \tfrac{1}{2}(1 - |g'(\xi)|)$$

for all x in I_δ; this is possible since g' is continuous at ξ. Hence,

$$|g'(x)| \leq |g'(x) - g'(\xi)| + |g'(\xi)| \leq \tfrac{1}{2}(1 - |g'(\xi)|) + |g'(\xi)| = L$$

for all $x \in I_\delta$. Now, suppose that x_k lies in the interval I_δ. Then,

$$x_{k+1} - \xi = g(x_k) - \xi = g(x_k) - g(\xi) = (x_k - \xi)g'(\eta_k)$$

by the Mean Value Theorem (Theorem A.3), where η_k lies between x_k and ξ, and therefore also belongs to I_δ. Hence $|g'(\eta_k)| \leq L$, and

$$|x_{k+1} - \xi| \leq L|x_k - \xi| \,. \tag{1.12}$$

This shows that x_{k+1} also lies in I_δ, and a simple argument by induction shows that if x_0 belongs to I_δ, then all x_k, $k \geq 0$, are in I_δ, and also

$$|x_k - \xi| \leq L^k |x_0 - \xi| \,, \qquad k \geq 0 \,. \tag{1.13}$$

Since $0 < L < 1$ this implies that the sequence (x_k) converges to ξ. □

[1] If you are familiar with the concept of Lebesgue measure, you will find the following result, known as **Rademacher's Theorem**, revealing. *A function f satisfying the Lipschitz condition (1.5) on an interval $[a, b]$ is differentiable on $[a, b]$, except, perhaps, at the points of a subset of zero Lebesgue measure.*

If the conditions of Theorem 1.5 are satisfied in the vicinity of a fixed point ξ, then the sequence (x_k) defined by the iteration $x_{k+1} = g(x_k)$, $k \geq 0$, will converge to ξ for any starting value x_0 that is sufficiently close to ξ. If, on the other hand, the conditions of Theorem 1.5 are violated, there is no guarantee that any sequence (x_k) defined by the iteration $x_{k+1} = g(x_k)$, $k \geq 0$, will converge to the fixed point ξ for any starting value x_0 near ξ. In order to distinguish between these two cases, we introduce the following definition.

Definition 1.3 *Suppose that g is a real-valued function, defined and continuous on the bounded closed interval $[a, b]$, such that $g(x) \in [a, b]$ for all $x \in [a, b]$, and let ξ denote a fixed point of g. We say that ξ is a* **stable fixed point** *of g, if the sequence (x_k) defined by the iteration $x_{k+1} = g(x_k)$, $k \geq 0$, converges to ξ whenever the starting value x_0 is sufficiently close to ξ. Conversely, if no sequence (x_k) defined by this iteration converges to ξ for any starting value x_0 close to ξ, except for $x_0 = \xi$, then we say that ξ is an* **unstable fixed point** *of g.*

We note that, with this definition, a fixed point may be neither stable nor unstable (see Exercise 2).

As will be demonstrated below in Example 1.5, even some very simple functions may possess both stable and unstable fixed points. Theorem 1.5 shows that if g' is continuous in a neighbourhood of ξ, then the condition $|g'(\xi)| < 1$ is sufficient to ensure that ξ is a stable fixed point. The case of an unstable fixed point will be considered later, in Theorem 1.6.

Now, assuming that ξ is a stable fixed point of g, we may also be interested in the speed at which the sequence (x_k) defined by the iteration $x_{k+1} = g(x_k)$, $k \geq 0$, converges to ξ. Under the hypotheses of Theorem 1.5, it follows from the proof of that theorem that

$$\lim_{k\to\infty} \frac{|x_{k+1} - \xi|}{|x_k - \xi|} = \lim_{k\to\infty} \left| \frac{g(x_k) - g(\xi)}{x_k - \xi} \right| = |g'(\xi)| \,. \tag{1.14}$$

Consequently, we can regard $|g'(\xi)| \in (0, 1)$ as a measure of the speed of convergence of the sequence (x_k) to the fixed point ξ.

Definition 1.4 *Suppose that $\xi = \lim_{k\to\infty} x_k$. We say that the sequence (x_k) converges to ξ* **at least linearly** *if there exist a sequence (ε_k) of positive real numbers converging to 0, and $\mu \in (0, 1)$, such that*

$$|x_k - \xi| \leq \varepsilon_k \,, \quad k = 0, 1, 2, \ldots \,, \qquad \text{and} \qquad \lim_{k\to\infty} \frac{\varepsilon_{k+1}}{\varepsilon_k} = \mu \,. \tag{1.15}$$

If (1.15) holds with $\mu = 0$, then the sequence (x_k) is said to converge to ξ **superlinearly**.

If (1.15) holds with $\mu \in (0,1)$ and $\varepsilon_k = |x_k - \xi|$, $k = 0, 1, 2, \ldots$, then (x_k) is said to converge to ξ **linearly**, *and the number $\rho = -\log_{10} \mu$ is then called the* **asymptotic rate of convergence** *of the sequence. If (1.15) holds with $\mu = 1$ and $\varepsilon_k = |x_k - \xi|$, $k = 0, 1, 2, \ldots$, the rate of convergence is slower than linear and we say that the sequence converges to ξ* **sublinearly**.

The words 'at least' in this definition refer to the fact that we only have inequality in $|x_k - \xi| \leq \varepsilon_k$, which may be all that can be ascertained in practice. Thus, it is really the sequence of bounds ε_k that converges linearly.

For a linearly convergent sequence the asymptotic rate of convergence ρ measures the number of correct decimal digits gained in one iteration; in particular, the number of iterations required in order to gain one more correct decimal digit is at most $[1/\rho] + 1$. Here $[1/\rho]$ denotes the largest integer that is less than or equal to $1/\rho$.

Under the hypotheses of Theorem 1.5, the equalities (1.14) will hold with $\mu = |g'(\xi)| \in [0,1)$, and therefore the sequence (x_k) generated by the simple iteration will converge to the fixed point ξ linearly or superlinearly.

Example 1.4 *Given that α is a fixed positive real number, consider the function g defined on the interval $[0,1]$ by*

$$g(x) = \begin{cases} 2^{-\left\{1+(\log_2(1/x))^{1/\alpha}\right\}^{\alpha}} & \text{for } 0 < x \leq 1\,, \\ 0 & \text{for } x = 0\,. \end{cases}$$

As $\lim_{x \to 0+} g(x) = 0$, the function g is continuous on $[0,1]$. Moreover, g is strictly monotonic increasing on $[0,1]$ and $g(x) \in [0, 1/2] \subset [0,1]$ for all x in $[0,1]$. We note that $\xi = 0$ is a fixed point of g (cf. Figure 1.3).

Consider the sequence (x_k) defined by $x_{k+1} = g(x_k)$, $k \geq 0$, with $x_0 = 1$. It is a simple matter to show by induction that $x_k = 2^{-k^{\alpha}}$, $k \geq 0$. Thus we deduce that (x_k) converges to $\xi = 0$ as $k \to \infty$. Since

$$\lim_{k \to \infty} \left| \frac{x_{k+1}}{x_k} \right| = \mu = \begin{cases} 1 & \text{for } 0 < \alpha < 1\,, \\ \frac{1}{2} & \text{for } \alpha = 1\,, \\ 0 & \text{for } \alpha > 1\,, \end{cases}$$

we conclude that for $\alpha \in (0,1)$ the sequence (x_k) converges to $\xi = 0$ sublinearly. For $\alpha = 1$ it converges to $\xi = 0$ linearly with asymptotic rate

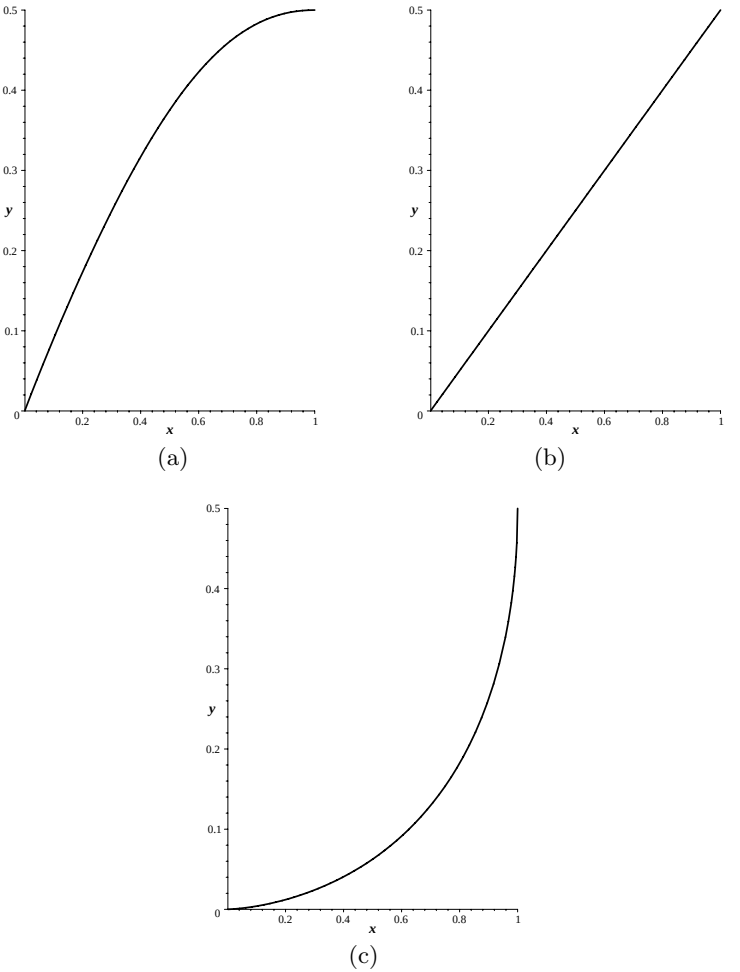

Fig. 1.3. Graph of the function g from Example 1.4 on the interval $x \in [0, 1]$ for (a) $\alpha = 1/2$, (b) $\alpha = 1$, (c) $\alpha = 2$.

$\rho = -\log_{10} \mu = \log_{10} 2$. When $\alpha > 1$, the sequence converges to the fixed point $\xi = 0$ superlinearly. The same conclusions could have been reached by showing (through tedious differentiation) that $\lim_{x \to 0+} g'(x) = \mu$, with μ as defined above for the various values of the parameter α. $\diamond$

For a linearly convergent simple iteration $x_{k+1} = g(x_k)$, where g' is continuous in a neighbourhood of the fixed point ξ and $0 < |g'(\xi)| < 1$, Definition 1.4 and (1.14) imply that the asymptotic rate of convergence

of the sequence (x_k) is $\rho = -\log_{10}|g'(\xi)|$. Evidently, a small value of $|g'(\xi)|$ corresponds to a large positive value of ρ and will result in more rapid convergence, while if $|g'(\xi)| < 1$ but $|g'(\xi)|$ is very close to 1, ρ will be a small positive number and the sequence will converge very slowly.[1]

Next, we discuss the behaviour of the iteration (1.3) in the vicinity of an *unstable fixed point* ξ. If $|g'(\xi)| > 1$, then the sequence (x_k) defined by (1.3) does not converge to ξ from any starting value x_0; the next theorem gives a rigorous proof of this fact.

Theorem 1.6 *Suppose that $\xi = g(\xi)$, where the function g has a continuous derivative in some neighbourhood of ξ, and let $|g'(\xi)| > 1$. Then, the sequence (x_k) defined by $x_{k+1} = g(x_k)$, $k \geq 0$, does not converge to ξ from any starting value x_0, $x_0 \neq \xi$.*

Proof Suppose that $x_0 \neq \xi$. As in the proof of Theorem 1.5, we can see that there is an interval $I_\delta = [\xi-\delta, \xi+\delta]$, $\delta > 0$, in which $|g'(x)| \geq L > 1$ for some constant L. If x_k lies in this interval, then

$$|x_{k+1} - \xi| = |g(x_k) - g(\xi)| = |(x_k - \xi)\, g'(\eta_k)| \geq L|x_k - \xi|\,,$$

for some η_k between x_k and ξ. If x_{k+1} lies in I_δ the same argument shows that

$$|x_{k+2} - \xi| \geq L|x_{k+1} - \xi| \geq L^2|x_k - \xi|\,,$$

and so on. Evidently, after a finite number of steps some member of the sequence $x_{k+1}, x_{k+2}, x_{k+3}, \ldots$ must be outside the interval I_δ, since $L > 1$. Hence there can be no value of $k_0 = k_0(\delta)$ such that $|x_k - \xi| \leq \delta$ for all $k \geq k_0$, and the sequence therefore does not converge to ξ. □

Example 1.5 *In this example we explore the simple iteration (1.3) for g defined by*

$$g(x) = \tfrac{1}{2}(x^2 + c)$$

where $c \in \mathbb{R}$ is a fixed constant.

The fixed points of the function g are the solutions of the quadratic equation $x^2 - 2x + c = 0$, which are $1 \pm \sqrt{(1-c)}$. If $c > 1$ there are no solutions (in the set $\mathbb{R}$ of real numbers, that is!), if $c = 1$ there is one solution in $\mathbb{R}$, and if $c < 1$ there are two.

[1] Thus $0 < \rho \ll 1$ corresponds to slow linear convergence and $\rho \gg 1$ to fast linear convergence. It is for this reason that we defined the asymptotic rate of convergence ρ, for a linearly convergent sequence, as $-\log_{10}\mu$ (or $-\log_{10}|g'(\xi)|$) rather than μ (or $|g'(\xi)|$).

Suppose now that $c < 1$; we denote the solutions by $\xi_1 = 1 - \sqrt{(1-c)}$ and $\xi_2 = 1 + \sqrt{(1-c)}$, so that $\xi_1 < 1 < \xi_2$. We see at once that $g'(x) = x$, so the fixed point ξ_2 is unstable, but that the fixed point ξ_1 is stable provided that $-3 < c < 1$. In fact, it is easy to see that the sequence (x_k) defined by the iteration $x_{k+1} = g(x_k)$, $k \geq 0$, will converge to ξ_1 if the starting value x_0 satisfies $-\xi_2 < x_0 < \xi_2$. (See Exercise 1.) If c is close to 1, $g'(\xi_1)$ will also be close to 1 and convergence will be slow. When $c = 0$, $\xi_1 = 0$ so that convergence is superlinear. This is an example of quadratic convergence which we shall meet later. ◇

The purpose of our next example is to illustrate the concept of asymptotic rate of convergence. According to Definition 1.4, the asymptotic rate of convergence of a sequence describes the relative closeness of successive terms in the sequence to the limit ξ as $k \to \infty$. Of course, for small values of k the sequence may behave in quite a different way, and since in practical computation we are interested in approximating the limit of the sequence by using just a small number of terms, the asymptotic rate of convergence may sometimes give a misleading impression.

Example 1.6 *In this example we study the convergence of the sequences* (u_k) *and* (v_k) *defined by*

$$\begin{aligned} u_{k+1} &= g_1(u_k), \qquad k = 0, 1, 2, \ldots, \qquad u_0 = 1\,, \\ v_{k+1} &= g_2(v_k), \qquad k = 0, 1, 2, \ldots, \qquad v_0 = 1\,, \end{aligned}$$

where

$$g_1(x) = 0.99x \qquad \textit{and} \qquad g_2(x) = \frac{x}{(1 + x^{1/10})^{10}}\,.$$

Each of the two functions has a fixed point at $\xi = 0$, and we easily find that $g_1'(0) = 0.99$, $g_2'(0) = 1$. Hence the sequence (u_k) is linearly convergent to zero with asymptotic rate of convergence $\rho = -\log_{10} 0.99 \approx 0.004$, while Theorem 1.5 does not apply to the sequence (v_k). It is quite easy to show by induction that $v_k = (k+1)^{-10}$, so the sequence (v_k) also converges to zero, but since $\lim_{k\to\infty}(v_{k+1}/v_k) = 1$ the convergence is sublinear. This means that, in the limit, (u_k) will converge faster than (v_k). However, this is not what happens for small k, as Table 1.2 shows very clearly.

The sequence (v_k) has converged to zero correct to 6 decimal digits when $k = 4$, and to 10 decimal digits when $k = 10$, at which stage u_k

Table 1.2. *The sequences* (u_k) *and* (v_k) *in Example 1.6.*

k	u_k	v_k
0	1.000000	1.000000
1	0.990000	0.000977
2	0.980100	0.000017
3	0.970299	0.000001
4	0.960596	0.000000
5	0.950990	0.000000
6	0.941480	0.000000
7	0.932065	0.000000
8	0.922745	0.000000
9	0.913517	0.000000
10	0.904382	0.000000

is still larger than 0.9. Although (u_k) eventually converges faster than v_k, we find that $u_k = (0.99)^k$ becomes smaller than $v_k = (k+1)^{-10}$ when

$$k > \frac{10}{\ln(1/0.99)} \ln(k+1)\,.$$

This first happens when $k = 9067$, at which point u_k and v_k are both roughly 10^{-40}. In this rather extreme example the concept of asymptotic rate of convergence is not useful, since for any practical purposes (v_k) converges faster than (u_k). ◇

1.3 Iterative solution of equations

In this section we apply the idea of simple iteration to the solution of equations. Given a real-valued continuous function f, we wish to construct a sequence (x_k), using iteration, which converges to a solution of $f(x) = 0$. We begin with an example where it is easy to derive various such sequences; in the next section we shall describe a more general approach.

Example 1.7 *Consider the problem of determining the solutions of the equation* $f(x) = 0$*, where* $f\colon x \mapsto \mathrm{e}^x - x - 2$*.*

Since $f'(x) = \mathrm{e}^x - 1$ the function f is monotonic increasing for positive x and monotonic decreasing for negative values of x. Moreover,

$$\left.\begin{aligned} f(1) &= \mathrm{e} - 3 < 0\,, \\ f(2) &= \mathrm{e}^2 - 4 > 0\,, \\ f(-1) &= \mathrm{e}^{-1} - 1 < 0\,, \\ f(-2) &= \mathrm{e}^{-2} > 0\,. \end{aligned}\right\} \tag{1.16}$$

Hence the equation $f(x) = 0$ has exactly one positive solution, which lies in the interval $(1, 2)$, and exactly one negative solution, which lies in the interval $(-2, -1)$. This is illustrated in Figure 1.4, which shows the graphs of the functions $x \mapsto \mathrm{e}^x$ and $x \mapsto x + 2$ on the same axes. We shall write ξ_1 for the positive solution and ξ_2 for the negative solution.

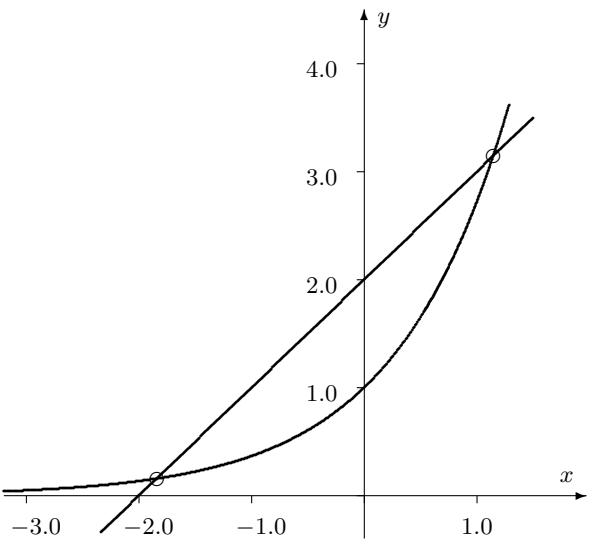

Fig. 1.4. Graphs of $y = \mathrm{e}^x$ and $y = x + 2$.

The equation $f(x) = 0$ may be written in the equivalent form

$$x = \ln(x + 2)\,,$$

which suggests a simple iteration defined by $g(x) = \ln(x + 2)$. We shall show that the positive solution ξ_1 is a stable fixed point of g, while ξ_2 is an unstable fixed point of g.

Clearly, $g'(x) = 1/(x + 2)$, so $0 < g'(\xi_1) < 1$, since ξ_1 is the positive solution. Therefore, by Theorem 1.5, the sequence (x_k) defined by the iteration

$$x_{k+1} = \ln(x_k + 2)\,, \qquad k = 0, 1, 2, \ldots\,, \tag{1.17}$$

will converge to the positive solution, ξ_1, provided that the starting value x_0 is sufficiently close to it.[1] As $0 < g'(\xi_1) < 1/3$, the asymptotic rate of convergence of (x_k) to ξ_1 is certainly greater than $\log_{10} 3$.

On the other hand, $g'(\xi_2) > 1$ since $-2 < \xi_2 < -1$, so the sequence (x_k) defined by (1.17) cannot converge to the solution ξ_2. It is not difficult to prove that for $x_0 > \xi_2$ the sequence (x_k) converges to ξ_1 while if $x_0 < \xi_2$ the sequence will decrease monotonically until $x_k \leq -2$ for some k, and then the iteration breaks down as $g(x_k)$ becomes undefined.

The equation $f(x) = 0$ may also be written in the form $x = \mathrm{e}^x - 2$, suggesting the sequence (x_k) defined by the iteration

$$x_{k+1} = \mathrm{e}^{x_k} - 2\,, \qquad k = 0, 1, 2, \ldots\,.$$

In this case $g(x) = \mathrm{e}^x - 2$ and $g'(x) = \mathrm{e}^x$. Hence $g'(\xi_1) > 1$, $g'(\xi_2) < \mathrm{e}^{-1}$, showing that the sequence (x_k) may converge to ξ_2, but cannot converge to ξ_1. It is quite straightforward to show that the sequence converges to ξ_2 for any $x_0 < \xi_1$, but diverges to $+\infty$ when $x_0 > \xi_1$.

As a third alternative, consider rewriting the equation $f(x) = 0$ as $x = g(x)$ where the function g is defined by $g(x) = x(\mathrm{e}^x - x)/2$; the fixed points of the associated iteration $x_{k+1} = g(x_k)$ are the solutions ξ_1 and ξ_2 of $f(x) = 0$, and also the point 0. For this iteration neither of the fixed points, ξ_1 or ξ_2, is stable, and the sequence (x_k) either converges to 0 or diverges to $\pm\infty$.

Evidently the given equation may be written in many different forms, leading to iterations with different properties. ◇

1.4 Relaxation and Newton's method

In the previous section we saw how various ingenious devices lead to iterations which may or may not converge to the desired solutions of a given equation $f(x) = 0$. We would obviously benefit from a more generally applicable iterative method which would, except possibly in special cases, produce a sequence (x_k) that always converges to a required solution. One way of constructing such a sequence is by relaxation.

[1] In fact, by applying the Contraction Mapping Theorem on an arbitrary bounded closed interval $[0, M]$ where $M > \xi_1$, we conclude that the sequence (x_k) defined by the iteration (1.17) will converge to ξ_1 from any positive starting value x_0.

Definition 1.5 *Suppose that f is a real-valued function, defined and continuous in a neighbourhood of a real number ξ.* **Relaxation** *uses the sequence (x_k) defined by*

$$x_{k+1} = x_k - \lambda f(x_k)\,, \qquad k = 0, 1, 2, \ldots\,, \tag{1.18}$$

where $\lambda \neq 0$ is a fixed real number whose choice will be made clear below, and x_0 is a given starting value near ξ.

If the sequence (x_k) defined by (1.18) converges to ξ, then ξ is a solution of the equation $f(x) = 0$, as we assume that f is continuous.

It is clear from (1.18) that relaxation is a simple iteration of the form $x_{k+1} = g(x_k)$, $k = 0, 1, 2, \ldots$, with $g(x) = x - \lambda f(x)$. Suppose now, further, that f is differentiable in a neighbourhood of ξ. It then follows that $g'(x) = 1 - \lambda f'(x)$ for all x in this neighbourhood; hence, if $f(\xi) = 0$ and $f'(\xi) \neq 0$, the sequence (x_k) defined by the iteration $x_{k+1} = g(x_k)$, $k = 0, 1, 2, \ldots$, will converge to ξ if we choose λ to have the same sign as $f'(\xi)$, to be not too large, and take x_0 sufficiently close to ξ. This idea is made more precise in the next theorem.

Theorem 1.7 *Suppose that f is a real-valued function, defined and continuous in a neighbourhood of a real number ξ, and let $f(\xi) = 0$. Suppose further that f' is defined and continuous in some neighbourhood of ξ, and let $f'(\xi) \neq 0$. Then, there exist positive real numbers λ and δ such that the sequence (x_k) defined by the relaxation iteration (1.18) converges to ξ for any x_0 in the interval $[\xi - \delta, \xi + \delta]$.*

Proof Suppose that $f'(\xi) = \alpha$, and that α is positive. If $f'(\xi)$ is negative, the proof is similar, with appropriate changes of sign. Since f' is continuous in some neighbourhood of ξ, we can find a positive real number δ such that $f'(x) \geq \frac{1}{2}\alpha$ in the interval $[\xi - \delta, \xi + \delta]$. Let M be an upper bound for $f'(x)$ in this interval. Hence $M \geq \frac{1}{2}\alpha$. In order to fix the value of the real number λ, we begin by noting that, for any $\lambda > 0$,

$$1 - \lambda M \leq 1 - \lambda f'(x) \leq 1 - \tfrac{1}{2}\lambda\alpha\,, \qquad x \in [\xi - \delta, \xi + \delta]\,.$$

We now choose λ so that these extreme values are equal and opposite, *i.e.*, $1 - \lambda M = -\vartheta$ and $1 - \frac{1}{2}\lambda\alpha = \vartheta$ for a suitable nonnegative real number ϑ. There is a unique value of ϑ for which this holds; it is given by the formula

$$\vartheta = \frac{2M - \alpha}{2M + \alpha}\,,$$