Chapter 3

The Solution of Nonlinear Equations

Introduction - Sketching $y = f(x)$

The most rudimentary method of sketching $y = f(x)$ consists of selecting a few points x_i in the domain of f, evaluating $y_i = f(x_i)$, plotting the coordinates (x_i, y_i), and then tracing a curve through these points. In regions where f changes rapidly a heavier concentration of points (x_i, y_i) is necessary to capture the behavior of the function.

A more sophisticated approach to curve tracing utilizes the following concepts.

a. The x intercepts of f; that is, the values ξ such that $f(\xi) = 0$.

b. Vertical asumptotes; that is, lines $x = a$, $|a| < \infty$, such that f is unbounded as $x \to a$.

c. Asymptotes of infinity; that is, curves that describe the behavior of f as x becomes unbounded.

d. Critical values of f; that is, values η in the domain of f where either $f'(\eta) = 0$ or $f'(\eta)$ fails to exist. In the former case the tangent line at $(\eta, f(\eta))$ is horizontal while in the latter the tangent line is vertical. If in addition the sign of $f'(x)$ changes on open intervals to the left and right of η then the point $(\eta, f(\eta))$ corresponds to a relative maximum or minimum.

e. Inflection points of f; that is, points $(\zeta, f(\zeta))$ such that either $f''(\zeta) = 0$ or $f''(\zeta)$ does not exist and, further, $f''(x)$ changes sign on open intervals to the left and right of ζ.

Example 3.0.1 Consider $y = f(x)$ where

$$
\begin{aligned}
f(x) &= \frac{1}{2}x^2 - e^{-x} \\
f'(x) &= x + e^{-x} \\
f''(x) &= 1 - e^{-x}.
\end{aligned}
$$

We first note that $f(0)f(1) < 0$ hence by the Intermediate Value Theorem f has a root $\xi \in (0,1)$. Since $f'(x) > 0$ for all x (why?) this value ξ is the unique root of f and, moreover, there are no relative maxima or minima. From $f''(x)$ we glean the information that $(0,-1)$ is an inflection point. Finally, note that the behavior of f as $x \to \infty$ is dominated by the term $\frac{1}{2}x^2$ while as $x \to -\infty$ f closely follows $-e^{-x}$. Combining this information yields the rough sketch of f shown in Figure 3.0.1 below. If we were now to numerically sketch the graph we would concentrate points near $x = 0$ to detect the change in concavity and concentrate points on the interval $(0,1)$ to locate the root ξ.

Figure 3.0.1: The graph $f(x) = \frac{1}{2}x^2 - e^{-x}$.

This example called on two concepts. The first is the notion of solving $f(x) = 0$ (and possibly $f'(x) = 0$, $f''(x) = 0$). Finding roots of functions is the subject of the present chapter. The second notion is that of asymptotics. In all but relatively trivial examples we defer this topic to Chapter 6.

Problems

3.0.1 Sketch the graph of each of the following equations.

a. $y = x(x^2 - 1)$ (This has 3 real roots.)

b. $1000y = 1000x(x^2 - 1) - 89$ (This has 3 real roots, 2 close together.)

c. $1000y = 1000x(x^2 - 1) - 90$ (This has one real root.)

3.0.2 Sketch the graph of the equation

$$xy^2 - 2xy + y - 1 = 0$$

and use it to show that the equation

$$x^7 - 2x^4 + x^3 - 1 = 0$$

only has one real root.

3.0.3 Descartes rule of signs states that a polynomial with real coefficients can only have as many positive real roots as there are changes of signs in the coefficients, although it may have $2k$ fewer roots than the number of changes of sign, where $k \geq 0$ is an integer. How many real roots can the polynomial

$$p(x) = x^7 - c^5x^2 + c^7$$

have if (a) $c > 0$, (b) $c < 0$?

3.1 Elementary One-Dimensional Methods

Finding a satisfactory value of ξ such that $f(\xi) = 0$ involves a miriad of considerations. For instance, a rough sketch of $y = f(x)$ may reveal that the root we seek corresponds to a change in sign of $f(x)$. If so and if only a crude approximation to the root, accurate to two or three significant digits, suffices then the simple method of bisections outlined below adequately suits the setting. If greater accuracy is sought we turn to the more sophisticated secant method or Newton's method. In exchange for greater accuracy, these methods have the disadvantage that they must be started sufficiently close to ξ before they converge.

For each of the methods we describe we sketch the algorithm (not code) for obtaining an aproximate value of ξ. In each algorithm, the user must specify the value of the positive number ε that is the desired accuracy of the computed root. In order to avoid difficulties due to round off error, ε should be larger than the magnitude of the error in the newly computed approximation in each cycle, assuming all values are exact at the start of the cycle. In other words, it is unreasonable to insist that ε be smaller than the precision of machine calculations. Finally, the user must also supply an upper limit on the number of iterations the algorithm performs.

3.1.1 The Method of Bisections

The bisection method is based on successive application of the Intermediate Value Theorem.

Example 3.1.1 Consider finding an approximate value of ξ such that

$$f(x) = x^3 - 2x - 3 = 0.$$

Either by means of the sketch in Figure 3.1.1 or by noting that $f(1)f(2) < 0$, it is clear that f must be zero for some $\xi \in (1, 2)$. Since $f'(x) > 0$ on $(1, 2)$, $f(x) = 0$ exactly once on $(1, 2)$.

We evaluate $f(x)$ at the midpoint of $[1, 2]$ to get

$$f(1.5) = -2.625 < 0 < 1 = f(2)$$

hence, the root ξ lies in the smaller interval $(1.5, 2)$. Next evaluating f at the midpoint 1.75 we have

$$f(1.75) \approx -1.14 < 0 < 1 = f(2)$$

and we now have that $1.75 < \xi < 2$; that is, the approximation $\xi = 1.875$ is in error by less than .125.

In general if f is continuous on some interval $[a, b]$ and $f(a)f(b) < 0$ then in the absence of round off error the bisection method is guaranteed to converge to within ε of a root $\xi \in (a, b)$.

Algorithm 3.1.2 Bisection Algorithm

$c = (a + b)/2$
if $b - c \leq \varepsilon$ then accept $c \to$ exit
    otherwise
    if $f(c)f(a) < 0$
        then $b = c$
        otherwise $a = c$

Figure 3.1.1: Sketch of $f(x) = x^3 - 2x - 3$.

Example 3.1.3 Referring to the function $f(x) = \frac{1}{2}x^2 - e^{-x}$ in Example 3.0.1 we tabulate the results in Table 3.1.1.

We remark that the bisection method converges even if more than one root of $f(x)$ is present in the interval. Of course, a rough sketch of $f(x)$ is usually sufficient for the user to select an interval $[a, b]$ containing a unique root.

There are disadvantages associated with the bisection method. Chief amongst them is that the algorithm converges very slowly when compared to the methods discussed next. Further, because the only information the algorithm uses is the sign of f at various points, good intermediate approximations are ignored. Not to abandon the algorithm altogether, bisection is often used to narrow down the location of a root before switching to a faster iterative technique. These faster techniques must, among other restrictions, be started sufficiently close to a root to guarantee convergence.

3.1.2 The Secant Method

In order to numerically find a root ξ of $f(x)$ faster than the bisection method more information about the values of f must be used at each step. For example, if by the Intermediate Value Theorem we know a unique root ξ of $f(x)$ exists on $[a, b] = [x_0, x_1]$ then it is possible to estimate ξ with the x-intercept of the secant line joining $(x_0, f(x_0))$ and $(x_1, f(x_1))$. See Figure

Table 3.1.1: The Bisection Algorithm applied to $f(x) = \frac{1}{2}x^2 - e^{-x}$.

n	c_n	Error Bound	$f(c_n)$
	$a = 0.5$		-.4815 3066
	$b = 1.0$		.1321 2055
1	.75	.25	-.1911 1655
2	.875	.125	-.0340 0495
3	.9375	.0625	.0478 4749
4	.9062 5	.0312 5	.0066 0800
5	.8906 25	.0156 25	-.0137 9272
6	.8984 375	.0078 125	-.0036 1045
7	.9023 4375	.0039 0625	.0014 9424
8	.9003 9062	.0019 5312	-.0010 5923
9	.9013 6718	.0009 7656	.0020 1722
10	.9008 7890	.0004 8828	-.0004 2107

3.1.2 below. Using the points $(x_0, f(x_0))$, $(x_1, f(x_1))$, and $(x_2, 0)$ we can equate two expressions for the slope of the secant line, namely,

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{0 - f(x_1)}{x_2 - x_1}.$$

Now solving for x_2 gives

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1). \tag{3.1.1}$$

One interpretation of this expression is that x_2 is x_1 plus a correction term. Another explanation follows by rewriting (3.1.1) as a weighted average between x_0 and x_1 the weights being $|f(x_0)|$ and $|f(x_1)|$, that is,

$$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)} \tag{3.1.2}$$

Intuitively from this explanation and the graph below, in general we expect faster convergence of the secant method than the bisection method.

Algorithm 3.1.4 Secant Algorithm

Figure 3.1.2: The first iterate of the secant method.

$F = f(x_0)$
$G = f(x_1)$
for $n = 1, 2, \ldots$

$$x_{n+1} = \frac{x_n - x_{n-1}}{G - F} G$$

if $|x_{n+1} - x_n| \leq \varepsilon$ then accept $x_{n+1} \rightarrow$ exit
otherwise
$F = G$
$G = f(x_{n+1})$

Example 3.1.5 Again we have $f(x) = \frac{1}{2}x^2 - e^{-x}$, $x_0 = 0.5$, and $x_1 = 1.0$. Unlike the bisection algorithm applied to the same problem each iterate of $f(x_n)$ gains at least a decimal point.

Several remarks are in order. First we choose to iterate using the expression

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) \tag{3.1.3}$$

rather than

$$x_{n+1} = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} \,. \tag{3.1.4}$$

Although (3.1.3) and (3.1.4) are algebraically equivalent, in practice (3.1.4) is more prone to round off error than (3.1.3). Our second remark is reminiscent

Figure 3.1.3: The secant method.

Table 3.1.2: The secant method applied to $f(x) = \frac{1}{2}x^2 - e^{-x}$.

n	x_{n+1}	$\lvert x_{n+1} - x_n \rvert$	$f(x_{n+1})$
	$x_0 = 0.5$		
	$x_1 = 1.0$		
1	.8923 4882		-.0115 4911
2	.9010 0252	.0086 7652	-.0002 5948
3	.9012 0142	.0001 9889	.0000 0050

of the conditioning discussion in Chapter 2. That is the secant method may converge to a point such that $|f(x)|$ is small and the iterates are "close", however, this may not approximate the root accurately. The figure below attempts to illustrate the situation. Notice that Table 3.1.1 includes an error bound column while the best we can do with Table table 3.1.2 is include a column tabulating $|x_{n+1} - x_n|$. Fortunately there are circumstances that guarantee the iterates converge to the true solution ξ of $f(x) = 0$. It can be shown that when f' exists and is non-zero in a neighborhood of ξ then by selecting x_0 and x_1 sufficiently close to ξ the secant method converges to ξ.

We close our discussion of the secant method by noting that some authors refer to the algorithm as the method of false position, or *regula falsi* method, while for others the method of false position and the secant method are

Figure 3.1.4:

distinct. A third variant of the secant method is the modified *regula falsi* method. These variants are discussed in the problems at the end of the section.

3.1.3 Newton's Method

The student of calculus knows a limiting process replaces the secant line by the tangent line. Therefore a natural adaptation of the secant method is to begin with a single point x_0 and plot its tangent line. As before, using the points $(x_0, f(x_0))$ and $(x_1, 0)$ along with $f'(x_0)$ we can calculate the x-intercept of the tangent line by equating slopes. That is,

$$f'(x_0) = \frac{x_0 - x_1}{f(x_0) - 0}$$

yields

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \ .$$

An alternative derivation uses the Taylor series development of $f(x)$ about $x = x_n$,

$$f(x) = f(x_n) + (x - x_n)f'(x_n) + \frac{(x - x_n)^2}{2!}f''(\eta)$$

Figure 3.1.5: The first iterate of Newton's Method.

where η is between x and x_n. If we let $x = \xi$ and use $f(\xi) = 0$ then

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

Hence the last term represents error and we iterate using

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

Algorithm 3.1.6 Newton's Algorithm

for $n = 1, 2, \ldots$
  $x_n = x_{n-1} - f(x_{n-1})/f'(x_{n-1})$
  if $|x_n - x_{n-1}| < \varepsilon$ then accept $x_n \rightarrow$ exit

Example 3.1.7 For $f(x) = \frac{1}{2}x^2 - e^{-x}$ Table 3.1.3 indicates that Newton's method outperforms the already rapid secant method.

In Section 3.4 we establish that if $f''(x)$ exists in a neighborhood N of a solution ξ of $f(x) = 0$, if $f'(x) \neq 0$ in N, and if x_0 is sufficiently close to ξ then the iterates of Newton's method converge to ξ.

Table 3.1.3: Newton's method applied to $f(x) = \frac{1}{2}x^2 - e^{-x}$.

n	x_n	$\lvert x_n - x_{n-1}\rvert$	$f(x_n)$
	$x_0 = 1$		
1	.9034 1213	.0965 8786	.0028 9198
2	.9012 0214	.0022 0999	.0000 0145
3	.9012 0103	.0000 0111	-1.4×10^{-12}

We note that among the bisection, secant, and Newton's methods, Newton's method converged fastest for the example discussed. However, Newton's method requires two function evaluations per step, that of $f(x)$ and $f'(x)$, while the other two methods require only one function evaluation per iterate. Therefore, an iterate of Newton's method is more costly besides requiring explicit knowledge of $f'(x)$. Indeed, it is shown in Section 3.5 that the rate of convergence per function evaluation is fastest for the secant method. Of the three, only Newton's method is applicable for complex roots.

Problems

3.1.1 a. By sketching a graph or otherwise, find an interval containing the real positive root of the function $x^2 - 3x - 1$.

b. Compute the real positive root of $x^2 - 3x - 1 = 0$ correct to 4 decimals by the method of bisections.

3.1.2 In the method of bisections let $x_1, x_2, x_3, \ldots$ represent successive approximations.

a. Show that

$$|x_{i+1} - x_i| \le \frac{|x_1 - x_2|}{2^{i-1}} .$$

b. Show that the number I of evaluations of f required to guarantee an approximation that is accurate to within ε is

$$I > \frac{\log\left[\frac{|x_1 - x_2|}{\varepsilon}\right]}{\log 2} + 1 .$$

3.1.3 The bisection method is applicable when $f(a)f(b) < 0$. If $f(x)$ has more than one zero in the interval $[a, b]$ which zero, if any, does the method locate?

3.1.4 Given the following functions

$$f(x) = (3x-1)^{-1} \qquad g(x) = \cos 10x \qquad h(x) = \begin{cases} 1 & x \geq 0.3 \\ -1 & x < 0.3 \end{cases}$$

that change sign in $[0, 1]$

a. What point does the bisection algorithm locate?

b. Is this a true zero of the function?

3.1.5 Carry out 3 steps of the secant method for the problem $x^3 - 3 = 0$, with $x_0 = 0$,$x_1 = 1$.

3.1.6 Use Newton's method to calculate the unique root of

$$x + e^{-Bx^2} \cos x = 0$$

with $B > 0$ a parameter to be chosen. Use a variety of increasing values of B, for example, $B = 1, 5, 10, 25, 50$. Among the choices of x_0 used, choose $x_0 = 0$ and explain any anomalous behavior. Theoretically, the Newton method converges for any value of x_0 and B. Compare this with actual computations for larger values of B.

3.1.7 Apply the Secant method to the equation in Problem 3.1.6

3.2 The Fixed Point Method

From the point of view of understanding the convergence properties of iterative methods it is more convenient to study the equation

$$x = g(x) \tag{3.2.1}$$

rather than

$$f(x) = 0. \tag{3.2.2}$$

Figure 3.2.1: Fixed points of $y = g(x)$.

When we solve the latter equation we are said to find the roots of $f(x)$ while in the nomenclature of the former equation we are finding the fixed points of $g(x)$. Figure3.2.1 below illustrates a function with three fixed points.

It is natural to solve the equation (3.2.1) by starting with some initial approximate x_0 and forming the sequence $\{x_n\}$ using the recurrence relation

$$x_{n+1} = g(x_n) \ , \ n = 0, 1, 2, \ldots \ . \tag{3.2.3}$$

Figure 3.2.2 illustrates that if $|g'(x)| < 1$ in a neighborhood N of the fixed point ξ and $x_0 \in N$ then $x_n \to \xi$ as $n \to \infty$. On the other hand Figure 3.2.3 shows that when $|g'(x)| > 1$ in general the sequence $\{x_n\}$ does not converge to ξ. We say "in general" because of course if $x_0 = \xi$ then convergence is trivial.

Example 3.2.1 Referring to the function $f(x) = \frac{1}{2}x^2 - e^{-x}$ in Example 3.0.1 one obvious way to convert the problem of finding ξ such that $f(\xi) = 0$ to a fixed point problem is to solve for x; that is,

$$g(x) = \sqrt{2e^{-x}} \, . \tag{3.2.4}$$

Selecting $x_0 = 1.0$ and iterating to a tolerance of $tol = .00005$ gives convergence after 11 iterates as shown in Table 3.2.1.

Another method of converting $f(x) = 0$ to a fixed point problem is simply to add x to both sides of the equation. For the particular $f(x)$ we are given

$$g(x) = x + \frac{1}{2}x^2 + e^{-x} \tag{3.2.5}$$

Figure 3.2.2: $|g'(x)| < 1 \Longrightarrow$ convergence.

Figure 3.2.3: $|g'(x)| > 1 \Longrightarrow$ possible divergence.

this proves unsuccessful as evidenced by Table 3.2.2 . The difficulty lies with the constraint on the magnitude of $g'(x)$ near $x = .9$.

Theorem 3.2.2 *Let g be defined on the interval $I \equiv [x_0 - a, x_0 + a]$, let*

$$|g(x) - g(y)| \leq \alpha|x - y| \tag{3.2.6}$$

for all x and y on this interval, where $0 \leq alpha < 1$, and let

$$|x_0 - g(x_0)| = |x_0 - x_1| \leq (1 - \alpha)a\,. \tag{3.2.7}$$

Then the sequence $\{x_p\}$ defined by $x_p = g(x_{p-1})$ converges to a number $\xi \in I$. Moreover, $\xi = g(\xi)$ is the unique solution in I of $x = g(x)$ in I and

$$|x_p - \xi| \leq \frac{\alpha}{1 - \alpha}|x_p - x_{p-1}| \leq \alpha^p a. \tag{3.2.8}$$

Table 3.2.1: The fixed point method applied to $g(x) = (2e^{-x})^{1/2}$.

n	x_{n-1}	$x_n = g(x_{n-1})$	$\lvert x_n - x_{n-1}\rvert$
1	1.000 000	.857 764	.142 236
2	.857 764	.920 988	.063 224
3	.920 988	.892 329	.028 659
4	.892 329	.905 208	.012 879
5	.905 208	.899 397	.005 810
6	.899 397	.902 014	.002 617
7	.902 014	.900 835	.001 179
8	.900 835	.901 366	.000 531
9	.901 366	.901 127	.000 239
10	.901 127	.901 235	.000 108
11	.901 235	.901 186	.000 049

Table 3.2.2: The fixed point method applied to $g(x) = x + \frac{1}{2}x^2 + e^{-x}$.

n	x_{n-1}	$x_n = g(x_{n-1})$	$\lvert x_n - x_{n-1}\rvert$
1	1.000 000	1.132 121	.132 121
2	1.132 121	1.450 620	.318 500
3	1.450 620	2.268 345	.817 724
4	2.268 345	4.737 555	2.469 210
5	4.737 555	15.951 008	11.213 453

Proof: We first show that $x_p \in I$ for all $n \geq 0$. This is clearly the case for $n = 0$ and $n = 1$. Now if $x_0, x_1, \ldots, x_k \in I$ where $k \geq 1$ then

$$\begin{aligned} |x_{k+1} - x_k| &= |g(x_k) - g(x_{k-1})| \\ &\leq \alpha|x_k - x_{k-1}| \\ &\leq \\ &\vdots \\ &\leq \alpha|x_1 - x_0| \\ &\leq \alpha^k(1-\alpha)a. \end{aligned} \tag{3.2.9}$$

Hence,

$$\begin{aligned}
|x_{k+1} - x_0| &= \left| \sum_{j=0}^{k} (x_{j+1} - x_j) \right| \\
&\le \sum_{j=0}^{k} |x_{j+1} - x_j| \\
&\le \sum_{j=0}^{k} \alpha^j (1-\alpha) a \\
&= (1 - \alpha^{k+1}) a < a
\end{aligned} \tag{3.2.10}$$

so that $x_{k+1} \in I$. Hence $x_p \in I$ for all $p \ge 0$.

Next we note by (3.2.10) that for $p \ge 1$ and $m \ge 10$

$$\begin{aligned}
|a_{m+p} - x_m| &= \sum_{k=0}^{p-1} (x_{m+k+1} - x_{m+k}) \\
&\le \sum_{k=0}^{p-1} |x_{m+k+1} - x_{m+k}| \\
&\le \left[\sum_{k=0}^{p-1} \alpha^k \right] |x_{m+1} - x_m| \\
&\le \frac{1-\alpha^p}{1-\alpha} \alpha^m (1-\alpha) |x_1 - x_0| \\
&\le \alpha^m a.
\end{aligned} \tag{3.2.11}$$

This estimate shows that the sequence $\{x_n\}$ is a Cauchy sequence that converges to some point $\xi \in I$. Letting $p \to \infty$ on the left hand side of (3.2.11) we get the error bounds

$$|\xi - x_m| \le \frac{\alpha}{1-\alpha} |x_m - x_{m-1}| \le \frac{\alpha^m}{1-\alpha} |x_1 - x_0| . \tag{3.2.12}$$

In order to show that $\xi = g(\xi)$, we note that

$$\begin{aligned}
|\xi - g(\xi)| &\le |\xi - x_p| + |x_p - g(\xi)| \\
&= |\xi - s_p| + |g(x_{p-1}) - g(\xi)| \\
&\le |\xi - x_p| + \alpha |x_{p-1} - \xi| \\
&\le \alpha^p a + \alpha^p a ,
\end{aligned} \tag{3.2.13}$$

and letting $p \to \infty$ on the right hand side, it follows that

$$|\xi - g(\xi)| = 0 \tag{3.2.14}$$

or, equivalently, that $\xi = g(\xi)$.

In order to prove uniqueness, let us assume that there exists an $\eta \in I$ such that $\eta = g(\eta)$ but $\eta \neq \xi$. Then, since $g(\xi) = \xi$,

$$0 < |\xi - \eta| = |g(\xi) - g(\eta)| < \alpha|\xi - \eta| < |\xi - \eta|\,,$$

which is a contradiction. Hence there is no $\eta \in I$ such that $\eta = g(\eta)$ and $\eta \neq \xi$. This completes the proof of Theorem 3.2.2.

The above theorem and proof may be used to describe several different numerical methods. It also gives an insight regarding the type of function g we want in order to achieve rapid convergence, how close to the solution ξ we should start the method, and how fast the method converges. We also obtain an *a priori* estimate of the error of x_n that depends only on x_{n-1} , x_n, and α.

In numerical analysis we sometimes hear the following statements regarding iterative methods such as Newton's method: "The method converges if it is started close enough to the solution." The following corollary states that this is true for the function g of Theorem 3.2.2

Theorem 3.2.3 *Let $\xi = g(\xi)$, let $|g'(\xi)| < 1$, and let $g'(x)$ be continuous in a neighborhood of ξ. If x_0 is sufficiently close to ξ, then the sequence $\{x_n\}$ defined by $x_n = g(x_{n-1})$ converges to ξ.*

Proof:

Problems

3.2.1 For each of the following equations determine an iteration function (and an interval I) so that the conditions of Theorem 3.2.3 are satisfied with respect to the smallest positive root.

a. $x^3 - x - 1 = 0$

b. $x - \tan x = 0$

c. $e^{-x} - \cos x = 0$

3.2.2 Determine the largest interval I with the property that for all $x_0 \in I$ fixed point iteration with the iteration function

$$g(x) = x(2 - ax)$$

converges when started with x_0. Is Theorem 3.2.3 satisfied for your choice of I? What numbers are possible limits of this iteration? Can you think of a good reason for using this particular iteration? Note that the interval depends on the constant a.

3.2.3 Same as Problem 3.2.2 but with

$$g(x) = \frac{x + a/x}{2}.$$

3.2.4 The equation $e^x - 4x^2 = 0$ has a root between $x = 4$ and $x = 5$. Show that we cannot find this root using fixed point iteration with the"natural" iteration function

$$x = \frac{1}{2}e^{x/2}.$$

Can you find an iteration function that correctly locates this root?

3.2.5 The equation $e^x - 4x^2 = 0$ also has a root between $x = 0$ and $x = 1$. Show that the iteration function $x = \frac{1}{2}e^{x/2}$ converges to this root if x_0 is chosen in the iterval $[0, 1]$.

3.3 Rate of Convergence of Iterative Methods

Consider finding a solution ξ of the equation $x = g(x)$ by generating a sequence $x_{p+1} = g(x_p)$, $p = 0, 1, 2, \ldots$. If g has a continuous derivative in a neighborhood of ξ then a Taylor series expansion about ξ gives

$$\begin{aligned} g(x) &= g(\xi) + (x - \xi)g'(\eta) && (3.3.1) \\ &= \xi + (x - \xi)g'(\eta) && (3.3.2) \end{aligned}$$

for some η between x and ξ. Substituting $x = x_{p-1}$ into (3.3.2) the equation becomes

$$x_p = \xi + (x_{p-1} - \xi)g'(\eta) \qquad (3.3.3)$$

or

$$|x_p - \xi| = |g'(\eta)|\, |x_{p-1} - \xi|\,. \tag{3.3.4}$$

Now for x_n sufficiently close to ξ (3.3.4) admits the approximate

$$|x_p - \xi| \approx |g'(\xi)|\, |x_{p-1} - \xi|\ ,\ p = 0, 1, \ldots \tag{3.3.5}$$

hence if $0 < |g'(\xi)| < 1$ then error decreases with each successive iterate.

More generally, at the solution ξ of the equation $x = g(x)$, the following may occur

$$g'(\xi) = g''(\xi) = \cdots = g^{(q-1)}(\xi) = 0 \tag{3.3.6}$$

$$g^{(q)}(\xi) \neq 0 \tag{3.3.7}$$

where q is a positive integer. In this case, the Taylor series expansion about ξ yields

$$g(x) = g(\xi) + \frac{(x-\xi)^q g^{(q)}(\eta)}{q!} \tag{3.3.8}$$

where again η is between x and ξ. Proceeding as above we have the estimate

$$|x_p - \xi| \approx \frac{g^{(q)}(\xi)|}{q!} |x_{p-1} - \xi|^q\ ,\ p = 1, 2, \ldots \tag{3.3.9}$$

Therefore the method is said to be of order q since for x_p sufficiently close to ξ x_p has roughly q times as many correct significant figures as x_{p-1}.

3.3.1 The Order of Newton's Method

Let us find the order q of Newton's method where recall the expression

$$x_{p+1} = x_p - \frac{f(x_p)}{f'(x_p)} \tag{3.3.10}$$

implies that the $g(x)$ we are considering is

$$g(x) = x - \frac{f(x)}{f'(x)}\,. \tag{3.3.11}$$

Differentiating $g(x)$

$$g'(x) = \frac{f(x)\, f''(x)}{(f'(x))^2} \tag{3.3.12}$$

we see that $g'(x)$ vanishes when $x = \xi$ assuming that $f'(x)$ as well as $f''(x)$ exist and are non-zero in a neighborhood N of ξ. Now checking the second derivative of $g(x)$,

$$\begin{aligned} g''(x) &= \frac{f(x)\, f^{(3)}(x)}{(f'(x))^2} + \frac{f''(x)}{f'(x)} - \frac{2f(x)(f''(x))^2}{(f'(x))^3} && (3.3.13) \\ &= \frac{f''(\xi)}{f'(\xi)} \neq 0 \; . && (3.3.14) \end{aligned}$$

Hence if $x_p \to \xi$ then

$$|x_p - \xi| \approx \left| \frac{f''(\xi)}{2f'(\xi)} \right| |x_{p-1} - \xi|^2 \; . \qquad (3.3.15)$$

In this setting, Newton's method is said to converge quadratically.

When $f'(\xi) = 0$ but $f'' \neq 0$ in a neighborhood N of ξ it is still possible to demonstrate the convergence of Newton's method. Using the Taylor series of $f(x)$ and $f'(x)$ we have

$$\begin{aligned} f(x) &= \frac{(x-\xi)^2}{x} f''(\xi) + \frac{(x-\xi)^3}{3!} f^{(3)}(\eta) && (3.3.16) \\ f'(x) &= (x-\xi) f''(\xi) + \frac{(x-\xi)^2}{2} f^{(3)}(\zeta) && (3.3.17) \end{aligned}$$

so that

$$g'(x) = \frac{f(x)\, f''(x)}{(f'(x))^2} \to \frac{1}{2} \text{ as } x \to \xi \; . \qquad (3.3.18)$$

Therefore referring to (3.3.9), when $f(\xi) = f'(\xi) = 0$, $f''(\xi) \neq 0$ Newton's method is apparently a first order method.

In general, the order of a method need not be an integer.

Definition 3.3.1 *The* order of an iterative method *that converges to ξ is defined to be the largest number q such that*

$$\limsup_{p\to\infty} \frac{|x_p - \xi|^{q'}}{|x_{p+1} - \xi|} = \infty \qquad (3.3.19)$$

for all $q' < q$.

3.3.2 The Order of the Secant Method

To deduce the order of the secant method we replace x_0, x_1, and x_2 in (3.1.1) by x_{p-1}, x_p, and x_{p+1} respectively. Then subtracting ξ from each side of (3.1.1) gives

$$\begin{aligned} \xi - x_{p+1} &= \xi - x_p + f(x_p)\,\frac{x_p - x_{p-1}}{f(x_p) - f(x_{p-1})} && (3.3.20) \\ &= (\xi - x_p)\,\frac{f[x_{p-1}, x_p] - f[x_{p-1}, \xi]}{f[x_{p-1}, x_p]} && (3.3.21) \end{aligned}$$

where we use the notation

$$f[a, b] = \frac{f(a) - f(b)}{a - b} \tag{3.3.22}$$

from Chapter ?. The right hand side of (3.3.22) further simplifies to the form

$$\xi - x_{p+1} = -(\xi - x_p)(\xi - x_{p-1})\,\frac{f[x_{p-1}, x_p, \xi]}{f[x_{p-1}, x_p]} \tag{3.3.23}$$

where

$$f[x_{p-1}, x_p, \xi] = \frac{f[x_{p-1}, x_p] - f[x_p, \xi]}{x_{p-1} - \xi}\,. \tag{3.3.24}$$

If we anticipate some results of Chapter ? and assume that $f''(x)$ exists in a neighborhood N of ξ then (3.3.21) yields

$$\xi - x_{p+1} = -\frac{f''(\eta_p)}{2f'(\zeta)}\,(\xi - x_p)(\xi - x_{p-1})\,, \tag{3.3.25}$$

for $p = 1, 2, \ldots$, where η_p and ζ_p are between the largest and smallest values of x_{p-1}, x_p, and ξ. If x_{p-1} and x_p are sufficiently close to ξ, if $f'(\xi) \neq 0$, and if $f''(x)$ is bounded in the neighborhood N, then we may assume that

$$\left|\frac{f''(\eta_p)}{2f'(\zeta_p)}\right| \leq M < \infty\,. \tag{3.3.26}$$

Hence, by setting

$$M|\xi - x_P| = \varepsilon_p \tag{3.3.27}$$

we obtain the inequalities

$$\varepsilon_{p+1} \leq \varepsilon_p \varepsilon_{p-1}\ ,\ p = 1, 2, \ldots\,. \tag{3.3.28}$$

Defining $\delta = \max\{\varepsilon_0, \varepsilon_1\}$, ? yields the following inequalities

$$\begin{aligned} \varepsilon_0 &\le \delta \\ \varepsilon_1 &\le \delta \\ \varepsilon_2 &\le \delta^2 \\ \varepsilon_3 &\le \delta^3 \\ \varepsilon_4 &\le \delta^5 \\ &\vdots \\ \varepsilon_p &\le \delta^{m_p} \end{aligned}$$

where

$$\begin{aligned} m_0 &= 1 \\ m_1 &= 1 \\ m_{p+1} &= m_p + m_{p-1} \ , \ p = 1, 2, \ldots \ . \end{aligned}$$

The sequence $\{m_p\}_{p=0}^{\infty}$ is called the Fibonacci sequence. It may be shown (see ?) that

$$m_p = \frac{1}{\sqrt{5}} \left(\alpha^{p+1} - \beta^{p+1} \right) \tag{3.3.29}$$

where

$$\alpha = \frac{1 + \sqrt{5}}{2} \ , \ \beta = \frac{1 - \sqrt{5}}{2} \, . \tag{3.3.30}$$

Thus for large p,

$$m_p \approx \frac{1}{\sqrt{5}} \alpha^{p+1} \approx .44721(1.61803)^{p+1} \tag{3.3.31}$$

and therefore, roughly,

$$|x_{p+1} - \xi| \approx C |x_p - \xi|^{(1+\sqrt{5})/2} \, . \tag{3.3.32}$$

The order of the secant method is thus $q = (1 + \sqrt{5})/2 \approx 1.61803$. The number $(1 + \sqrt{5})/2$ is called the *golden number*.

The definition of the order q of convergence of a method applies to a system of equations is the same as for scalar equations, except that ? is now replaced by

$$\limsup_{p \to \infty} \frac{||\vec{x}_p - \vec{\xi}||^{q'}}{||\vec{x}_{p+1} - \vec{\xi}||} = \infty \ \text{ for all } q' < q \, . \tag{3.3.33}$$

That is, the order is q if (3.3.33) is satisfied for all $q' < q$.

3.4 Roots of Polynomials

In applications the problem of obtaining a solution for a nonlinear equation often takes the polynomial form

$$p(z) \equiv a_m z^m + a_{m-1} z^{m-1} + \cdots + a_0 = 0 \tag{3.4.1}$$

For reasons associated with a particular application it may be known that the roots of this polynomial are real. For instance this occurs in the case of Gaussian quadrature. At other times the number of real roots is not known a priori. Aside from this question, some applications require knowledge of all the roots while others are only concerned with the root that is largest or smallest in magnitude.

Example 3.4.1 Suppose that an electrical network has a response Y that satisfies the integral equation

$$Y(t) = \int_0^t K(t-u)Y(u)\,du + G(t) \tag{3.4.2}$$

where K and G are given polynomials. Taking the Laplace transform of each side leads to an expression for $y(s)$, the Laplace transform of Y, that is the ratio of two polynomials. Once the roots of the denominator, including multiplicities, are known we can invert $y(s)$ thereby deriving an explicit expression for Y in terms of t.

In this section we consider the approximate solution of (3.4.1) for the variety circumstances mentioned above. The method of bisections usually suffices to quickly produce a rough approximation to a real root. If more accuracy for a real root is required then the secant method is an alternative. Newton's method is appropriate for accurate estimates of complex roots.

3.4.1 Newton's Method for Roots of Polynomials

A positive feature of Newton's method is its easy adaptation to finding both real and complex roots of polynomials. If $p(z)$ is given in the form (3.4.1) we can evaluate $p(z)$ and $q(z) = p'(z)$ simultaneously be means of the following algorithm.

Algorithm 3.4.2

$$\begin{aligned} &q \leftarrow 0 \\ &p \leftarrow a_m \\ &\left.\begin{aligned} &q \leftarrow zq + p \\ &p \leftarrow zp + a_{m-n} \end{aligned}\right\} \quad n = 1, 2, \ldots, m \end{aligned}$$

Consider the sequence

$$z_{n+1} = z_n - \frac{p(z_n)}{q(z_n)} \;,\; n = 0, 1, 2, \ldots \tag{3.4.3}$$

generated by Newton's method. If the coefficients of $p(z)$ are real and the initial guess z_0 is real then necessarily all successive iterates are real. Hence under this scenario it is not possible to find a complex root ξ of $p(z)$. However, for z_0 complex and sufficiently close to ξ the method does indeed converge to ξ.

Example 3.4.3 ?Frank has an example here of finding a real root of a polynomial. Doesn't seem to make sense given the previous paragraph.

3.4.2 Bernoulli's Method

Bernoulli's method is particularly suited for getting the root of largest modulus as well as the root of smallest modulus for (3.4.1). The general solution of the difference equation ? can be written in the form

$$y_n = \sum_{i=1}^{k} \sum_{j=0}^{\mu_i - 1} c_{i,j} n^j z_i^n \tag{3.4.4}$$

where the z_i are the k distinct roots of $p(z)$ with multiplicity μ_i. Hence if $|z_1| > |z_i|$ for $i = 2, 3, \ldots, k$, and if at least one of the $c_{1,j}$ is non-zero, then

$$\frac{y_{n+1}}{y_n} = \frac{\sum_{i=0}^{\mu_1 - 1} c_{i,j}(n+1)^j z_1^{n+1}[1 + o(1)]}{\sum_{j=0}^{\mu_1 - 1} c_{i,j} n^j z_1^n [1 + o(1)]} \tag{3.4.5}$$

approaches z_1 as $n \to \infty$. Thus in this case the difference equation ? is a convenient tool for computing the root of largest modulus.

Since (3.4.4) holds for negative as well as positive integers n, it follows that if $0 < |z_k| < |z_i|$ for $i = k-1, k-2, \ldots, 1$ and if at least one of the $c_{k,j}$ in (3.4.4) is non-zero then by proceeding as in (3.4.5)

$$\frac{y_{n+1}}{y_n} \to \frac{1}{z_k} \text{ as } n \to -\infty\,. \tag{3.4.6}$$

Now suppose that the coefficients of $p(z)$ are real and that the two roots of largest modulus of $p(z)$ are complex conjugate roots. If real values are taken for y_n in (3.4.1) and if $z_1 = \rho e^{i\theta}$, $z_2 = \rho e^{-i\theta}$, then in the case when the roots z_1 and z_2 are simple, $c_1 = ae^{i\alpha}$,$c_2 = ae^{-i\alpha}$, $a > 0$ (3.4.4) yields

$$y_n = 2a\rho^n[\cos(n\theta + \phi) + o(1)] \tag{3.4.7}$$

as $n \to \infty$. Hence for large n,

$$[y_{n+1} - 2\rho\cos\theta y_n + \rho^2 y_{n-1}][1 + o(1)] = 0 \tag{3.4.8}$$

$$[y_{n+2} - 2\rho\cos\theta y_{n+1} + \rho^2 y_n][1 + o(1)] = 0\,. \tag{3.4.9}$$

Solving the pair of equations (3.4.8) and (3.4.9) for ρ and $\cos\theta$ yields

$$\frac{y_n^2 - y_{n+1}y_{n-1}}{y_{n-1}^2 - y_n y_{n-2}} \to \rho^2 \tag{3.4.10}$$

$$\frac{y_{n+1}y_{n-2} - y_n y_{n-1}}{y_{n-2}^2 - y_n y_{n-2}} \to -2\rho\cos\theta \tag{3.4.11}$$

as $n \to \infty$.

Example 3.4.4 Suppose that we want to find the largest root of the cubic polynomial $p(x) = x^3 - 3x^2 + x - 10$ by Bernoulli's method. Then the difference equation we use is

$$y_{n+1} = 3y_n - y_{n-1} + 10y_{n-2}\ ,\ n = 2, 3, \ldots, \tag{3.4.12}$$

where we take $y_0 = y_1 = y_2 = 1$. Table 3.4.1 includes selected values of y_n as well as the ratios y_{n+1}/y_n. Apparently the convergence is very slow. The exact answer, to nine decimals, is 3.522 167 383. Fewer iterates whould have sufficed to yield the same accuracy if we had used Atkin's δ^2 process.

Example 3.4.5 If we apply the Bernoulli method to solve the cubic $p(x) = x^3 - x^2 + 2x + 6 = 0$ then the difference equation becomes

$$y_{n+1} = y_n - 2Y_{n-1} - 6y_{n-2},\ ,\ n = 2, 3, \ldots\ . \tag{3.4.13}$$

Taking $y_0 = y_1 = 0$ and $y_2 = 1$ we calculate the entries in Table 3.4.2 below. In this case $y_{n=1}/y_n$ does not approach a limit, and it is necessary to use (3.4.10) and (3.4.11). The exact answer, to 4 decimals is $\rho = 2.1924$, $\theta = 1.0324$. Notice the very slow convergence.

Table 3.4.1: Bernoulli's method for solving $p(x) = x^3 - 3x^2 + x - 10$.

n	y_n	y_{n+1}/y_n
0	1	1
1	1	1
2	1	12
3	12	3.75
4	45	2.955 555 555
5	133	3.563 909 774
10	74 977	3.526 961 602
15	40 662 788	3.521 973 678
20	22 040 879 700	3.522 171 187
25	11 194 756 442 000	3.522 167 371
26	42 081 321 560 000	

3.4.3 Deflation

If ξ is a root of the polynomial $p(z)$ in (3.4.1) then there is a polynomial $q(z)$ of degree $m-1$ such that

$$p(z) = (z-\xi)q(z)\,. \tag{3.4.14}$$

Setting

$$q(z) = b_{m-1}z^{m-1} + b_{m-2}z^{m-2} + \cdots + b_0 \tag{3.4.15}$$

in (3.4.14) and using (3.4.1), we get

$$\sum_{k=0}^{m} a_k z^k = \sum_{k=0}^{m} (b_k - \xi b_{k-1}) z^k \tag{3.4.16}$$

where $b_m = b_{-1} = 0$. Equating coefficients of equal powers of z on each side yields the recurence relation

$$b_m \;=\; 0 \tag{3.4.17}$$

$$b_k \;=\; a_{k+1} + \xi b_{k=1}\ ,\ k = m-1, m-2, \ldots, 0$$

or, if we compute the coefficients in reverse order assuming that $\xi \neq 0$,

$$b_{-1} \;=\; 0 \tag{3.4.18}$$

Table 3.4.2: Bernoulli's method for solving $p(x) = x^3 - x^2 + 2x + 6 = 0$.

n	y_n	ρ_n^2	$-2\rho_n \cos\theta$	ρ_n	θ
0	0				
1	0				
2	1				
3	1				
4	-1				
5	-9	6.8	-2.2	2.6077	1.1353
10	-101	4.82	-2.30	2.1954	1.0484
15	17 747	4.8947	-2.2463	2.1919	1.0328
16	-7005	4.8114	-2.2500	2.1935	1.0323
17	-101 149	4.8124	-2.2475	2.1937	1.0330
18	-193 621				

$$b_k = \frac{b_{k-1} - a_k}{\xi}, \quad k = 0, 1, \ldots, m-1 .$$

Example 3.4.6 Let $p(z)$ be defined as in Example 3.4.1. Take $\xi = z_4$ in Table 3.4.1 and find $q(z)$ be means of (3.4.19). The values of a_k and the compute b_k are given in Table 3.4.3. Thus $q(x) \approx x^2 - 2.248298x + 4.806545$. Notice that since $b_{m-1} = a_m$ the computed b_{m-1} in (3.4.19) provides a check on the accuracy of the previously computed b_k.

Table 3.4.3: The coefficients of $q(x)$ in Example 3.6.5??.

k	a_k	b_k
0	6	4.806 545 277
1	2	-2.248 297 824
2	-1	1.000 000 002
3	1	

Let us next consider the problem of finding all the zeros of the polynomial (3.4.1). In the textbook [12] J. H. Wilkinson has studied the effect of errors in the coefficients of $q(z)$ on the roots of $q(z)$. A small error in a computed root of $p(z)$ may cause a large change in some of the roots of $q(z)$. For example, the polynomial

$$p(z) \equiv (z-1)(z-2)(z-3)\cdots(z-19)(z-20) \tag{3.4.19}$$

obviously has the roots $1, 2, 3, \cdots, 20$. The coefficient of z^{19} in $p(z)$ is -190. If we leave all coefficients of z^k fixed, except for that of z^{19} from which we subtract $2^{-23} \approx 1.19 \times 10^{-7}$, then the exact roots of the resulting polynomial to nine decimals are shown in Table 3.4.4.

Table 3.4.4: Accurate zeros of $(z-1)(z-2)\cdots(z-20) - z^{-23}z^{19}$.

1.00000 0000		6.00000 6944
2.00000 0000		6.99969 7234
3.00000 0000		8.00726 7603
4.00000 0000		8.91725 0249
4.99999 9928		20.84690 8101
10.09526 6145	$\pm$	0.64350 0904i
11.79363 3881	$\pm$	1.65232 9728i
13.99235 8137	$\pm$	2.51883 0070i
16.73073 7466	$\pm$	2.81262 4894i
19.50243 9400	$\pm$	1.94033 0347i

J. H. Wilkinson has shown moreover in [12] that if the relations (3.4.18) are used to repeatedly deflate $p(z)$ then for best numerical accuracy one should compute the zeros in ascending order of magnitude. Similarly if one computes the zero of $p(z)$ that has largest magnitude, then for greatest accuracy in the resulting polynomial $q(z)$ one should use (3.4.19) rather than (3.4.18).

Other direct methods of getting all the roots of $p(z)$ are described in the problems that follow.

It can be shown (see Problem **??**) that if $p(z)$ has roots $z_1, z_2, \ldots, z_m$ and if z_k is a root of p with multiplicity μ, then the polynomial

$$p(x, \varepsilon) \equiv p(z) + \varepsilon r(z) \tag{3.4.20}$$

where

$$r(z) \equiv r_m z^m + r_{m-1} z^{m-1} + \cdots + r_0 \tag{3.4.21}$$

has roots $z_k(\varepsilon)$ where

$$z_k(\varepsilon) - z_k - \left[-\frac{\mu! r(z_k)}{p^{\mu}(z_k)}\right]^{1/\mu} \varepsilon^{1/\mu} = O(\varepsilon^{2/\mu}) \tag{3.4.22}$$

as $\varepsilon \to 0$. This relation show that a small change in the coefficients of $p(z)$ may cause a relatively large change in the zeros of $p(z)$. For example, if $\mu = 3$ and $\varepsilon = 10^{-9}$ then $varepsilon^{1/\mu} = 10^{-3} = 10^{6} \times 10^{-9}$, i. e. ; the change in z_k may be 10^6 times as large as the change in a coefficient.