Chapter 2

Solutions of Equations of One Variable

2.1 Introduction

In this chapter we consider one of the most basic problems of numerical approximation, the root-finding problem. This process involves finding a root, or solution, of an equation of the form \( f(x) = 0 \). A root of this equation is also called a zero of the function \( f \). This is one of the oldest known approximation problems, yet research continues in this area at the present time.

The problem of finding an approximation to the root of an equation can be traced at least as far back as 1700 B.C. A cuneiform table in the Yale Babylonian Collection dating from that period gives a sexagesimal (base-60) number equivalent to 1.414222 as an approximation to \( \sqrt{2} \), a result that is accurate to within \( 10^{-5} \). This approximation can be found by applying a technique given in Section 2.4.

2.2 The Bisection Method

The first and most elementary technique we consider is the Bisection, or Binary-Search, method. The Bisection method is used to determine, to any specified accuracy that your computer will permit, a solution to \( f(x) = 0 \) on an interval \([a, b]\), provided that \( f \) is continuous on the interval and that \( f(a) \) and \( f(b) \) are of opposite sign. Although the method will work for the case when more than one root is contained in the interval \([a, b]\), we assume for simplicity of our discussion that the root in this interval is unique.

To begin the Bisection method, set \( a_1 = a \) and \( b_1 = b \), as shown in Figure 2.1, and let \( p_1 \) be the midpoint of the interval \([a, b]\):

\[
p_1 = a_1 + \frac{b_1 - a_1}{2}.
\]
If \( f(p_1) = 0 \), then the root \( p \) is given by \( p = p_1 \); if \( f(p_1) \neq 0 \), then \( f(p_1) \) has the same sign as either \( f(a_1) \) or \( f(b_1) \).

**Figure 2.1**

If \( f(p_1) \) and \( f(a_1) \) have the same sign, then \( p \) is in the interval \((p_1, b_1)\), and we set

\[
a_2 = p_1 \quad \text{and} \quad b_2 = b_1.
\]

If, on the other hand, \( f(p_1) \) and \( f(a_1) \) have opposite signs, then \( p \) is in the interval \((a_1, p_1)\), and we set

\[
a_2 = a_1 \quad \text{and} \quad b_2 = p_1.
\]

We reapply the process to the interval \([a_2, b_2]\), and continue forming \([a_3, b_3]\), \([a_4, b_4]\), \ldots. Each new interval will contain \( p \) and have length one half of the length of the preceding interval.

**[Bisection Method]** An interval \([a_{n+1}, b_{n+1}]\) containing an approximation to a root of \( f(x) = 0 \) is constructed from an interval \([a_n, b_n]\) containing the root by first letting

\[
p_n = a_n + \frac{b_n - a_n}{2}.
\]

Then set

\[
a_{n+1} = a_n \quad \text{and} \quad b_{n+1} = p_n \quad \text{if} \quad f(a_n)f(p_n) < 0,
\]

and

\[
a_{n+1} = p_n \quad \text{and} \quad b_{n+1} = b_n \quad \text{otherwise.}
\]
There are three stopping criteria commonly incorporated in the Bisection method. First, the method stops if one of the midpoints happens to coincide with the root. It also stops when the length of the search interval is less than some prescribed tolerance we call \( TOL \). The procedure also stops if the number of iterations exceeds a preset bound \( N_0 \).

To start the Bisection method, an interval \([a, b]\) must be found with \( f(a) \cdot f(b) < 0 \). At each step, the length of the interval known to contain a zero of \( f \) is reduced by a factor of 2. Since the midpoint \( p_1 \) must be within \((b - a)/2\) of the root \( p \), and each succeeding iteration divides the interval under consideration by 2, we have

\[
|p_n - p| \leq \frac{b - a}{2^n}.
\]

As a consequence, it is easy to determine a bound for the number of iterations needed to ensure a given tolerance. If the root needs to be determined within the tolerance \( TOL \), we need to determine the number of iterations, \( n \), so that

\[
\frac{b - a}{2^n} < TOL.
\]

Solving for \( n \) in this inequality gives

\[
\frac{b - a}{TOL} < 2^n, \quad \text{which implies that} \quad \log_2 \left( \frac{b - a}{TOL} \right) < n.
\]

Since the number of required iterations to guarantee a given accuracy depends on the length of the initial interval \([a, b]\), we want to choose this interval as small as possible. For example, if \( f(x) = 2x^3 - x^2 + x - 1 \), we have both

\[
f(-4) \cdot f(4) < 0 \quad \text{and} \quad f(0) \cdot f(1) < 0,
\]

so the Bisection method could be used on either \([-4, 4]\) or \([0, 1]\). Starting the Bisection method on \([0, 1]\) instead of \([-4, 4]\) reduces by 3 the number of iterations required to achieve a specified accuracy.

**EXAMPLE 1**

The equation \( f(x) = x^3 + 4x^2 - 10 = 0 \) has a root in \([1, 2]\) since \( f(1) = -5 \) and \( f(2) = 14 \). It is easily seen from a sketch of the graph of \( f \) in Figure 2.2 that there is only one root in \([1, 2]\).
To use Maple to approximate the root, we define the function $f$ by the command

\[ f := x \rightarrow x^3 + 4x^2 - 10; \]

The values of $a_1$ and $b_1$ are given by

\[ a_1 := 1; b_1 := 2; \]

We next compute $f(a_1) = -5$ and $f(b_1) = 14$ by

\[ f(a_1) := f(a_1); f(b_1) := f(b_1); \]

and the midpoint $p_1 = 1.5$ and $f(p_1) = 2.375$ by

\[ p_1 := a_1 + 0.5(b_1 - a_1); \]
\[ f(p_1) := f(p_1); \]

Since $f(a_1)$ and $f(p_1)$ have opposite signs, we reject $b_1$ and let $a_2 = a_1$ and $b_2 = p_1$. This process is continued to find $p_2$, $p_3$, and so on.

As discussed in the Preface, each of the methods we consider in the book has an accompanying set of programs contained on the CD that is in the back of the book. The programs are given for the programming languages C, FORTRAN, and Pascal, and also in Maple V, Mathematica, and MATLAB. The program BISECT21, provided with the inputs $a = 1$, $b = 2$, $TOL = 0.0005$, and $N_0 = 20$, gives the values in Table 2.1. The actual root $p$, to 10 decimal places, is $p = 1.3652300134$, and $|p - p_{11}| < 0.0005$. Since the expected number of iterations is $\log_2((2 - 1)/0.0005) \approx 10.96$, the bound $N_0$ was certainly sufficient.
2.2. THE BISECTION METHOD

The Bisection method, though conceptually clear, has serious drawbacks. It is slow to converge relative to the other techniques we will discuss, and a good intermediate approximation may be inadvertently discarded. This happened, for example, with $p_9$ in Example 1. However, the method has the important property that it always converges to a solution and it is easy to determine a bound for the number of iterations needed to ensure a given accuracy. For these reasons, the Bisection method is frequently used as a dependable starting procedure for the more efficient methods presented later in this chapter.

The bound for the number of iterations for the Bisection method assumes that the calculations are performed using infinite-digit arithmetic. When implementing the method on a computer, consideration must be given to the effects of round-off error. For example, the computation of the midpoint of the interval $[a_n, b_n]$ should be found from the equation

$$p_n = a_n + \frac{b_n - a_n}{2}$$

instead of from the algebraically equivalent equation

$$p_n = \frac{a_n + b_n}{2}.$$ 

The first equation adds a small correction, $(b_n - a_n)/2$, to the known value $a_n$. When $b_n - a_n$ is near the maximum precision of the machine, this correction might be in error, but the error would not significantly affect the computed value of $p_n$. However, in the second equation, if $b_n - a_n$ is near the maximum precision of the machine, it is possible for $p_n$ to return a midpoint that is not even in the interval $[a_n, b_n]$.

A number of tests can be used to see if a root has been found. We would normally use a test of the form

$$|f(p_n)| < \varepsilon,$$
where $\varepsilon > 0$ would be a small number related in some way to the tolerance. However, it is also possible for the value $f(p_n)$ to be small when $p_n$ is not near the root $p$.

As a final remark, to determine which subinterval of $[a_n, b_n]$ contains a root of $f$, it is better to make use of the *signum* function, which is defined as

$$
\text{sgn}(x) = \begin{cases} 
-1, & \text{if } x < 0, \\
0, & \text{if } x = 0, \\
1, & \text{if } x > 0.
\end{cases}
$$

The test

$$
\text{sgn}(f(a_n)) \text{sgn}(f(p_n)) < 0 \quad \text{instead of} \quad f(a_n)f(p_n) < 0
$$

gives the same result but avoids the possibility of overflow or underflow in the multiplication of $f(a_n)$ and $f(p_n)$. 
EXERCISE SET 2.2

1. Use the Bisection method to find \( p_3 \) for \( f(x) = \sqrt{x} - \cos x \) on \([0, 1]\).

2. Let \( f(x) = 3(x + 1)(x - \frac{1}{2})(x - 1) \). Use the Bisection method on the following intervals to find \( p_3 \).
   
   (a) \([-2, 1.5]\]   
   (b) \([-1.25, 2.5]\]

3. Use the Bisection method to find solutions accurate to within \( 10^{-2} \) for \( x^3 - 7x^2 + 14x - 6 = 0 \) on each interval.
   
   (a) \([0, 1]\]   
   (b) \([1, 3.2]\]   
   (c) \([3.2, 4]\]

4. Use the Bisection method to find solutions accurate to within \( 10^{-2} \) for \( x^4 - 2x^3 - 4x^2 + 4x + 4 = 0 \) on each interval.
   
   (a) \([-2, -1]\]   
   (b) \([0, 2]\]   
   (c) \([2, 3]\]   
   (d) \([-1, 0]\]

5. (a) Sketch the graphs of \( y = x \) and \( y = 2 \sin x \).

   (b) Use the Bisection method to find an approximation to within \( 10^{-2} \) to the first positive value of \( x \) with \( x = 2 \sin x \).

6. (a) Sketch the graphs of \( y = x \) and \( y = \tan x \).

   (b) Use the Bisection method to find an approximation to within \( 10^{-2} \) to the first positive value of \( x \) with \( x = \tan x \).

7. Let \( f(x) = (x + 2)(x + 1)x(x - 1)^3(x - 2) \). To which zero of \( f \) does the Bisection method converge for the following intervals?
   
   (a) \([-3, 2.5]\]   
   (b) \([-2.5, 3]\]   
   (c) \([-1.75, 1.5]\]   
   (d) \([-1.5, 1.75]\]

8. Let \( f(x) = (x + 2)(x + 1)^2x(x - 1)^3(x - 2) \). To which zero of \( f \) does the Bisection method converge for the following intervals?
   
   (a) \([-1.5, 2.5]\]   
   (b) \([-0.5, 2.4]\]   
   (c) \([-0.5, 3]\]   
   (d) \([-3, -0.5]\]

9. Use the Bisection method to find an approximation to \( \sqrt{3} \) correct to within \( 10^{-4} \). [Hint: Consider \( f(x) = x^2 - 3 \).]
10. Use the Bisection method to find an approximation to \( \sqrt[3]{25} \) correct to within \( 10^{-4} \).

11. Find a bound for the number of Bisection method iterations needed to achieve an approximation with accuracy \( 10^{-3} \) to the solution of \( x^3 + x - 4 = 0 \) lying in the interval \([1, 4]\). Find an approximation to the root with this degree of accuracy.

12. Find a bound for the number of Bisection method iterations needed to achieve an approximation with accuracy \( 10^{-4} \) to the solution of \( x^3 - x - 1 = 0 \) lying in the interval \([1, 2]\). Find an approximation to the root with this degree of accuracy.

13. The function defined by \( f(x) = \sin \pi x \) has zeros at every integer. Show that when \( -1 < a < 0 \) and \( 2 < b < 3 \), the Bisection method converges to

   \[
   \begin{align*}
   &\text{(a) } 0, \quad \text{if } a + b < 2 \\
   &\text{(b) } 2, \quad \text{if } a + b > 2 \\
   &\text{(c) } 1, \quad \text{if } a + b = 2
   \end{align*}
   \]
2.3 The Secant Method

Although the Bisection method always converges, the speed of convergence is usually too slow for general use. Figure 2.3 gives a graphical interpretation of the Bisection method that can be used to discover how improvements on this technique can be derived. It shows the graph of a continuous function that is negative at \( a_1 \) and positive at \( b_1 \). The first approximation \( p_1 \) to the root \( p \) is found by drawing the line joining the points \((a_1, \text{sgn}(f(a_1))) \) = \((a_1, -1)\) and \((b_1, \text{sgn}(f(b_1))) \) = \((b_1, 1)\) and letting \( p_1 \) be the point where this line intersects the \( x \)-axis. In essence, the line joining \((a_1, -1)\) and \((b_1, 1)\) has been used to approximate the graph of \( f \) on the interval \([a_1, b_1]\). Successive approximations apply this same process on subintervals of \([a_1, b_1]\), \([a_2, b_2]\), and so on. Notice that the Bisection method uses no information about the function \( f \) except the fact that \( f(x) \) is positive and negative at certain values of \( x \).

Figure 2.3

Suppose that in the initial step we know that \(|f(a_1)| < |f(b_1)|\). Then we would expect the root \( p \) to be closer to \( a_1 \) than to \( b_1 \). Alternatively, if \(|f(b_1)| < |f(a_1)|\), \( p \) is likely to be closer to \( b_1 \) than to \( a_1 \). Instead of choosing the intersection of the line through \((a_1, \text{sgn}(f(a_1))) = (a_1, -1)\) and \((b_1, \text{sgn}(f(b_1))) = (b_1, 1)\) as the approximation to the root \( p \), the \textbf{Secant method} chooses the \( x \)-intercept of the secant line to the curve, the line through \((a_1, f(a_1))\) and \((b_1, f(b_1))\). This places the approximation closer to the endpoint of the interval for which \( f \) has smaller absolute value, as shown in Figure 2.4.
The sequence of approximations generated by the Secant method is started by setting \( p_0 = a \) and \( p_1 = b \). The equation of the secant line through \((p_0, f(p_0))\) and \((p_1, f(p_1))\) is
\[
y = f(p_1) + \frac{f(p_1) - f(p_0)}{p_1 - p_0} (x - p_1).
\]
The \( x \)-intercept \((p_2, 0)\) of this line satisfies
\[
0 = f(p_1) + \frac{f(p_1) - f(p_0)}{p_1 - p_0} (p_2 - p_1)
\]
and solving for \( p_2 \) gives
\[
p_2 = p_1 - \frac{f(p_1)(p_1 - p_0)}{f(p_1) - f(p_0)}.
\]

The Secant method does not have the root-bracketing property of the Bisection method. As a consequence, the method does not always converge, but when it does converge, it generally does so much faster than the Bisection method.

We use two stopping conditions in the Secant method. First, we assume that \( p_n \) is sufficiently accurate when \(|p_n - p_{n-1}|\) is within a given tolerance. Also, a safeguard exit based upon a maximum number of iterations is given in case the method fails to converge as expected.
The iteration equation should not be simplified algebraically to
\[ p_{n+1} = p_n - \frac{f(p_n)(p_n - p_{n-1})}{f(p_n) - f(p_{n-1})} = \frac{f(p_{n-1})p_n - f(p_n)p_{n-1}}{f(p_{n-1}) - f(p_n)}. \]

Although this is algebraically equivalent to the iteration equation, it could increase the significance of rounding error if the nearly equal numbers \( f(p_{n-1})p_n \) and \( f(p_n)p_{n-1} \) are subtracted.

**EXAMPLE 1**

In this example we will approximate a root of the equation \( x^3 + 4x^2 - 10 = 0 \). To use Maple we first define the function \( f(x) \) and the numbers \( p_0 \) and \( p_1 \) with the commands

\[
> f := x -> x^3 + 4*x^2 - 10;
> p0 := 1; p1 := 2;
\]

The values of \( f(p_0) = -5 \) and \( f(p_1) = 14 \) are computed by

\[
> fp0 := f(p0); fp1 := f(p1);
\]

and the first secant approximation, \( p_2 = \frac{24}{19} \), by

\[
> p2 := p1 - fp1*(p1-p0)/(fp1-fp0);
\]

The next command forces a floating-point representation for \( p_2 \) instead of an exact rational representation.

\[
> p2 := evalf(p2);
\]

We compute \( f(p_2) = -1.602274379 \) and continue to compute \( p_3 = 1.338827839 \) by

\[
> fp2 := f(p2);
> p3 := p2 - fp2*(p2-p1)/(fp2-fp1);
\]

The program SECANT22 with inputs \( p_0 = 1, p_1 = 2, TOL = 0.0005 \), and \( N_0 = 20 \) produces the results in Table 2.2. About half the number of iterations are needed, compared to the Bisection method in Example 1 of Section 2.2. Further, \( |p - p_6| = |1.3652300134 - 1.3652300011| < 1.3 \times 10^{-8} \) is much smaller than the tolerance 0.0005.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( p_n )</th>
<th>( f(p_n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.2631578947</td>
<td>-1.6022743840</td>
</tr>
<tr>
<td>3</td>
<td>1.3388278388</td>
<td>-0.4303647480</td>
</tr>
<tr>
<td>4</td>
<td>1.3666163947</td>
<td>0.0229094308</td>
</tr>
<tr>
<td>5</td>
<td>1.3652119026</td>
<td>-0.0002990679</td>
</tr>
<tr>
<td>6</td>
<td>1.3652300011</td>
<td>-0.0000002032</td>
</tr>
</tbody>
</table>
There are other reasonable choices for generating a sequence of approximations based on the intersection of an approximating line and the $x$-axis. The method of False Position (or Regula Falsi) is a hybrid bisection-secant method that constructs approximating lines similar to those of the Secant method but always brackets the root in the manner of the Bisection method. As with the Bisection method, the method of False Position requires that an initial interval $[a, b]$ first be found, with $f(a)$ and $f(b)$ of opposite sign. With $a_1 = a$ and $b_1 = b$, the approximation, $p_2$, is given by

$$p_2 = a_1 - \frac{f(a_1)(b_1 - a_1)}{f(b_1) - f(a_1)}.$$  

If $f(p_2)$ and $f(a_1)$ have the same sign, then set $a_2 = p_2$ and $b_2 = b_1$. Alternatively, if $f(p_2)$ and $f(b_1)$ have the same sign, set $a_2 = a_1$ and $b_2 = p_2$. (See Figure 2.5.)

---

[Method of False Position] An interval $[a_{n+1}, b_{n+1}]$, for $n > 1$, containing an approximation to a root of $f(x) = 0$ is found from an interval $[a_n, b_n]$ containing the root by first computing

$$p_{n+1} = a_n - \frac{f(a_n)(b_n - a_n)}{f(b_n) - f(a_n)}.$$  

Then set

$$a_{n+1} = a_n \quad \text{and} \quad b_{n+1} = p_{n+1} \quad \text{if} \quad f(a_n)f(p_{n+1}) < 0,$$

and

$$a_{n+1} = p_{n+1} \quad \text{and} \quad b_{n+1} = b_n \quad \text{otherwise.}$$

---

Figure 2.5

<table>
<thead>
<tr>
<th>Secant method</th>
<th>Method of False Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$y = f(x)$</td>
</tr>
<tr>
<td>$x$</td>
<td></td>
</tr>
<tr>
<td>$0$</td>
<td>$p_0$</td>
</tr>
<tr>
<td>$p_1$</td>
<td></td>
</tr>
<tr>
<td>$p_2$</td>
<td></td>
</tr>
<tr>
<td>$p_3$</td>
<td></td>
</tr>
<tr>
<td>$p_4$</td>
<td></td>
</tr>
<tr>
<td>$p$</td>
<td>$x$</td>
</tr>
</tbody>
</table>
2.3. THE SECANT METHOD

Although the method of False Position may appear superior to the Secant method, it generally converges more slowly, as the results in Table 2.3 indicate for the problem we considered in Example 1. In fact, the method of False Position can converge even more slowly than the Bisection method (as the problem given in Exercise 14 shows), although this is not usually the case. The program FALPOS23 implements the method of False Position.

Table 2.3

<table>
<thead>
<tr>
<th>n</th>
<th>$a_n$</th>
<th>$b_n$</th>
<th>$p_{n+1}$</th>
<th>$f(p_{n+1})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00000000</td>
<td>2.00000000</td>
<td>1.26315789</td>
<td>−1.60227438</td>
</tr>
<tr>
<td>2</td>
<td>1.26315789</td>
<td>2.00000000</td>
<td>1.33882784</td>
<td>−0.43036475</td>
</tr>
<tr>
<td>3</td>
<td>1.33882784</td>
<td>2.00000000</td>
<td>1.35854634</td>
<td>−0.11000879</td>
</tr>
<tr>
<td>4</td>
<td>1.35854634</td>
<td>2.00000000</td>
<td>1.36354744</td>
<td>−0.02776209</td>
</tr>
<tr>
<td>5</td>
<td>1.36354744</td>
<td>2.00000000</td>
<td>1.36480703</td>
<td>−0.00698342</td>
</tr>
<tr>
<td>6</td>
<td>1.36480703</td>
<td>2.00000000</td>
<td>1.36512372</td>
<td>−0.00175521</td>
</tr>
<tr>
<td>7</td>
<td>1.36512372</td>
<td>2.00000000</td>
<td>1.36520330</td>
<td>−0.00044106</td>
</tr>
</tbody>
</table>
EXERCISE SET 2.3

1. Let \( f(x) = x^2 - 6 \), \( p_0 = 3 \), and \( p_1 = 2 \). Find \( p_3 \) using each method.
   (a) Secant method
   (b) method of False Position

2. Let \( f(x) = -x^3 - \cos x \), \( p_0 = -1 \), and \( p_1 = 0 \). Find \( p_3 \) using each method.
   (a) Secant method
   (b) method of False Position

3. Use the Secant method to find solutions accurate to within \( 10^{-4} \) for the following problems.
   (a) \( x^3 - 2x^2 - 5 = 0 \), on \([1, 4]\)
   (b) \( x^3 + 3x^2 - 1 = 0 \), on \([-3, -2]\)
   (c) \( x - \cos x = 0 \), on \([0, \pi/2]\)
   (d) \( x - 0.8 - 0.2 \sin x = 0 \), on \([0, \pi/2]\)

4. Use the Secant method to find solutions accurate to within \( 10^{-5} \) for the following problems.
   (a) \( 2x \cos 2x - (x - 2)^2 = 0 \) on \([2, 3]\) and on \([3, 4]\)
   (b) \( (x - 2)^2 - \ln x = 0 \) on \([1, 2]\) and on \([e, 4]\)
   (c) \( e^x - 3x^2 = 0 \) on \([0, 1]\) and on \([3, 5]\)
   (d) \( \sin x - e^{-x} = 0 \) on \([0, 1]\), on \([3, 4]\) and on \([6, 7]\)

5. Repeat Exercise 3 using the method of False Position.


7. Use the Secant method to find all four solutions of \( 4x \cos(2x) - (x - 2)^2 = 0 \) in \([0, 8]\) accurate to within \( 10^{-5} \).

8. Use the Secant method to find all solutions of \( x^2 + 10 \cos x = 0 \) accurate to within \( 10^{-5} \).

9. Use the Secant method to find an approximation to \( \sqrt{3} \) correct to within \( 10^{-4} \), and compare the results to those obtained in Exercise 9 of Section 2.2.

10. Use the Secant method to find an approximation to \( \sqrt[3]{25} \) correct to within \( 10^{-6} \), and compare the results to those obtained in Exercise 10 of Section 2.2.

11. Approximate, to within \( 10^{-4} \), the value of \( x \) that produces the point on the graph of \( y = x^2 \) that is closest to \((1, 0)\). [Hint: Minimize \( [d(x)]^2 \), where \( d(x) \) represents the distance from \((x, x^2)\) to \((1, 0)\).]
12. Approximate, to within $10^{-4}$, the value of $x$ that produces the point on the graph of $y = 1/x$ that is closest to $(2, 1)$.

13. The fourth-degree polynomial

$$f(x) = 230x^4 + 18x^3 + 9x^2 - 221x - 9$$

has two real zeros, one in $[-1, 0]$ and the other in $[0, 1]$. Attempt to approximate these zeros to within $10^{-6}$ using each method.

(a) method of False Position  
(b) Secant method

14. The function $f(x) = \tan \pi x - 6$ has a zero at $(1/\pi) \arctan 6 \approx 0.447431543$. Let $p_0 = 0$ and $p_1 = 0.48$ and use 10 iterations of each of the following methods to approximate this root. Which method is most successful and why?

(a) Bisection method

(b) method of False Position

(c) Secant method

15. Use Maple to determine how many iterations of the Secant method with $p_0 = \frac{1}{7}$ and $p_1 = \pi/4$ are needed to find a root of $f(x) = \cos x - x$ to within $10^{-100}$.

16. The sum of two numbers is 20. If each number is added to its square root, the product of the two sums is 155.55. Determine the two numbers to within $10^{-4}$.

17. A trough of length $L$ has a cross section in the shape of a semicircle with radius $r$. (See the accompanying figure.) When filled with water to within a distance $h$ of the top, the volume, $V$, of water is

$$V = L \left[ 0.5\pi r^2 - r^2 \arcsin \left( \frac{h}{r} \right) - h(r^2 - h^2)^{1/2} \right]$$

Suppose $L = 10$ ft, $r = 1$ ft, and $V = 12.4$ ft$^3$. Find the depth of water in the trough to within 0.01 ft.

18. A particle starts at rest on a smooth inclined plane whose angle $\theta$ is changing at a constant rate

$$\frac{d\theta}{dt} = \omega < 0.$$
At the end of $t$ seconds, the position of the object is given by

$$x(t) = \frac{g}{2\omega^2} \left( e^{\omega t} - \frac{e^{-\omega t}}{2} - \sin \omega t \right).$$

Suppose the particle has moved 1.7 ft in 1 s. Find, to within $10^{-5}$, the rate $\omega$ at which $\theta$ changes. Assume that $g = -32.17$ ft/s$^2$. 
2.4 Newton’s Method

The Bisection and Secant methods both have geometric representations that use the zero of an approximating line to the graph of a function \( f \) to approximate the solution to \( f(x) = 0 \). The increase in accuracy of the Secant method over the Bisection method is a consequence of the fact that the secant line to the curve better approximates the graph of \( f \) than does the line used to generate the approximations in the Bisection method.

The line that best approximates the graph of the function at a point on its graph is the tangent line to the graph at that point. Using this line instead of the secant line produces Newton’s method (also called the Newton–Raphson method), the technique we consider in this section.

Suppose that \( p_0 \) is an initial approximation to the root \( p \) of the equation \( f(x) = 0 \) and that \( f' \) exists in an interval containing all the approximations to \( p \). The slope of the tangent line to the graph of \( f \) at the point \( (p_0, f(p_0)) \) is \( f'(p_0) \), so the equation of this tangent line is

\[
y - f(p_0) = f'(p_0)(x - p_0).
\]

Since this line crosses the \( x \)-axis when the \( y \)-coordinate of the point on the line is zero, the next approximation, \( p_1 \), to \( p \) satisfies

\[
0 - f(p_0) = f'(p_0)(p_1 - p_0),
\]

which implies that

\[
p_1 = p_0 - \frac{f(p_0)}{f'(p_0)},
\]

provided that \( f'(p_0) \neq 0 \). Subsequent approximations are found for \( p \) in a similar manner, as shown in Figure 2.6.

Figure 2.6
[Newton’s Method] The approximation $p_{n+1}$ to a root of $f(x) = 0$ is computed from the approximation $p_n$ using the equation

$$p_{n+1} = p_n - \frac{f(p_n)}{f'(p_n)}.$$

**EXAMPLE 1** In this example we use Newton’s method to approximate the root of the equation $x^3 + 4x^2 - 10 = 0$. Maple is used to find the first iteration of Newton’s method with $p_0 = 1$. We define $f(x)$ and compute $f'(x)$ by

```maple
> f := x -> x^3 + 4*x^2 - 10;
f := x -> x^3 + 4*x^2 - 10;
> fp := x -> D(f)(x);
f := x -> x^3 + 4*x^2 - 10
> fp := x -> D(f)(x);
> p0 := 1;
p0 := 1;
```

The first iteration of Newton’s method gives $p_1 = \frac{16}{11}$, which is obtained with

```maple
> p1 := p0 - f(p0)/fp(p0);
p1 := p0 - f(p0)/fp(p0);
```

A decimal representation of $1.454545455$ for $p_1$ is given by

```maple
> p1 := evalf(p1);
p1 := evalf(p1)
```

The process can be continued to generate the entries in Table 2.4.

<table>
<thead>
<tr>
<th>Table 2.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

We use $p_0 = 1$, $TOL = 0.0005$, and $N_0 = 20$ in the program NEWTON24 to compare the convergence of this method with those applied to this problem previously. The number of iterations needed to solve the problem by Newton’s method is less than the number needed for the Secant method, which, in turn, required less than half the iterations needed for the Bisection method. In addition, for Newton’s method we have $|p - p_4| \approx 10^{-10}$. 

\[\square\]
Newton’s method generally produces accurate results in just a few iterations. With the aid of Taylor polynomials we can see why this is true. Suppose $p$ is the solution to $f(x) = 0$ and that $f''$ exists on an interval containing both $p$ and the approximation $p_n$. Expanding $f$ in the first Taylor polynomial at $p_n$ and evaluating at $x = p$ gives

$$0 = f(p) = f(p_n) + f'(p_n)(p - p_n) + \frac{f''(\xi)}{2}(p - p_n)^2,$$

where $\xi$ lies between $p_n$ and $p$. Consequently, if $f'(p_n) \neq 0$, we have

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

Since

$$p_{n+1} = p_n - \frac{f(p_n)}{f'(p_n)},$$

this implies that

$$p - p_{n+1} = -\frac{f''(\xi)}{2f'(p_n)}(p - p_n)^2.$$

If a positive constant $M$ exists with $|f''(x)| \leq M$ on an interval about $p$, and if $p_n$ is within this interval, then

$$|p - p_{n+1}| \leq \frac{M}{2|f'(p_n)|}|p - p_n|^2.$$

The important feature of this inequality is that the error $|p - p_{n+1}|$ of the $(n + 1)$st approximation is bounded by approximately the square of the error of the $n$th approximation, $|p - p_n|$. This implies that Newton’s method has the tendency to approximately double the number of digits of accuracy with each successive approximation. Newton’s method is not, however, infallible, as the equation in Exercise 12 shows.

**EXAMPLE 2** Find an approximation to the solution of the equation $x = 3^{-x}$ that is accurate to within $10^{-8}$.

A solution of this equation corresponds to a solution of

$$0 = f(x) = x - 3^{-x}.$$

Since $f$ is continuous with $f(0) = -1$ and $f(1) = \frac{2}{3}$, a solution of the equation lies in the interval $(0, 1)$. We have chosen the initial approximation to be the midpoint of this interval, $p_0 = 0.5$. Succeeding approximations are generated by applying the formula

$$p_{n+1} = p_n - \frac{f(p_n)}{f'(p_n)} = p_n - \frac{p_n - 3^{-p_n}}{1 + 3^{-p_n} \ln 3}.$$
These approximations are listed in Table 2.5, together with differences between successive approximations. Since Newton’s method tends to double the number of decimal places of accuracy with each iteration, it is reasonable to suspect that $p_3$ is correct at least to the places listed.

Table 2.5

| $n$ | $p_n$ | $|p_n - p_{n-1}|$ |
|-----|-------|-----------------|
| 0   | 0.500000000 |              |
| 1   | 0.547329757 | 0.047329757    |
| 2   | 0.547808574 | 0.000478817    |
| 3   | 0.547808622 | 0.000000048    |

The success of Newton’s method is predicated on the assumption that the derivative of $f$ is nonzero at the approximations to the root $p$. If $f'$ is continuous, this means that the technique will be satisfactory provided that $f'(p) \neq 0$ and that a sufficiently accurate initial approximation is used. The condition $f'(p) \neq 0$ is not trivial; it is true precisely when $p$ is a simple root. A simple root of a function $f$ occurs at $p$ if a function $q$ exists with the property that, for $x \neq p$,

$$f(x) = (x - p)q(x), \quad \text{where} \quad \lim_{x \to p} q(x) \neq 0.$$  

When the root is not simple, Newton’s method may converge, but not with the speed we have seen in our previous examples.

**EXAMPLE 3**  
The root $p = 0$ of the equation $f(x) = e^x - x - 1 = 0$ is not simple, since both $f(0) = e^0 - 0 - 1 = 0$ and $f'(0) = e^0 - 1 = 0$. The terms generated by Newton’s method with $p_0 = 0$ are shown in Table 2.6 and converge slowly to zero. The graph of $f$ is shown in Figure 2.7.
2.4. NEWTON’S METHOD

\[ f(x) = e^x - x - 1 \]

Diagram showing the graph of \( f(x) = e^x - x - 1 \) with points \((-1, e^{-1})\) and \((1, e - 2)\).
EXERCISE SET 2.4

1. Let \( f(x) = x^2 - 6 \) and \( p_0 = 1 \). Use Newton’s method to find \( p_2 \).

2. Let \( f(x) = -x^3 - \cos x \) and \( p_0 = -1 \). Use Newton’s method to find \( p_2 \). Could \( p_0 = 0 \) be used for this problem?

3. Use Newton’s method to find solutions accurate to within \( 10^{-4} \) for the following problems.
   \[(a) \quad x^3 - 2x^2 - 5 = 0, \quad \text{on } [1, 4]\]
   \[(b) \quad x^3 + 3x^2 - 1 = 0, \quad \text{on } [-3, -2]\]
   \[(c) \quad x - \cos x = 0, \quad \text{on } [0, \pi/2]\]
   \[(d) \quad x - 0.8 - 0.2\sin x = 0, \quad \text{on } [0, \pi/2]\]

4. Use Newton’s method to find solutions accurate to within \( 10^{-5} \) for the following problems.
   \[(a) \quad 2x \cos 2x - (x - 2)^2 = 0, \quad \text{on } [2, 3] \text{ and } [3, 4]\]
   \[(b) \quad (x - 2)^2 - \ln x = 0, \quad \text{on } [1, 2] \text{ and } [e, 4]\]
   \[(c) \quad e^x - 3x^2 = 0, \quad \text{on } [0, 1] \text{ and } [3, 5]\]
   \[(d) \quad \sin x - e^{-x} = 0, \quad \text{on } [0, 1], [3, 4], \text{ and } [6, 7]\]

5. Use Newton’s method to find all four solutions of \( 4x \cos(2x) - (x - 2)^2 = 0 \) in \([0, 8]\) accurate to within \( 10^{-5} \).

6. Use Newton’s method to find all solutions of \( x^2 + 10\cos x = 0 \) accurate to within \( 10^{-5} \).

7. Use Newton’s method to approximate the solutions of the following equations to within \( 10^{-5} \) in the given intervals. In these problems the convergence will be slower than normal since the roots are not simple roots.
   \[(a) \quad x^2 - 2xe^{-x} + e^{-2x} = 0, \quad \text{on } [0, 1]\]
   \[(b) \quad \cos(x + \sqrt{2}) + x (x/2 + \sqrt{2}) = 0, \quad \text{on } [-2, -1]\]
   \[(c) \quad x^3 - 3x^2(2^{-x}) + 3x(4^{-x}) + 8^{-x} = 0, \quad \text{on } [0, 1]\]
   \[(d) \quad e^{6x} + 3(\ln 2)^2e^{2x} - (\ln 8)e^{4x} - (\ln 2)^3, \quad \text{on } [-1, 0]\]

8. The numerical method defined by
   \[
p_n = p_{n-1} - \frac{f(p_{n-1})f'(p_{n-1})}{[f'(p_{n-1})]^2 - f(p_{n-1})f''(p_{n-1})},
   \]
   for \( n = 1, 2, \ldots \), can be used instead of Newton’s method for equations having multiple roots. Repeat Exercise 7 using this method.
9. Use Newton’s method to find an approximation to $\sqrt{3}$ correct to within $10^{-4}$, and compare the results to those obtained in Exercise 9 of Sections 2.2 and 2.3.

10. Use Newton’s method to find an approximation to $\sqrt{25}$ correct to within $10^{-6}$, and compare the results to those obtained in Exercise 10 of Section 2.2 and 2.3.

11. In Exercise 14 of Section 2.3 we found that for $f(x) = \tan \pi x - 6$, the Bisection method on $[0, 0.48]$ converges more quickly than the method of False Position with $p_0 = 0$ and $p_1 = 0.48$. Also, the Secant method with these values of $p_0$ and $p_1$ does not give convergence. Apply Newton’s method to this problem with (a) $p_0 = 0$, and (b) $p_0 = 0.48$. (c) Explain the reason for any discrepancies.

12. Use Newton’s method to determine the first positive solution to the equation $\tan x = x$, and explain why this problem can give difficulties.

13. Use Newton’s method to solve the equation

$$0 = \frac{1}{2} + \frac{1}{4} x^2 - x \sin x - \frac{1}{2} \cos 2x,$$

with $p_0 = \frac{\pi}{2}$.

Iterate using Newton’s method until an accuracy of $10^{-5}$ is obtained. Explain why the result seems unusual for Newton’s method. Also, solve the equation with $p_0 = 5\pi$ and $p_0 = 10\pi$.

14. Use Maple to determine how many iterations of Newton’s method with $p_0 = \pi/4$ are needed to find a root of $f(x) = \cos x - x$ to within $10^{-100}$.

15. Player A will shut out (win by a score of 21–0) player B in a game of racquetball with probability

$$P = \frac{1 + p}{2} \left( \frac{p}{1 - p + p^2} \right)^{21},$$

where $p$ denotes the probability A will win any specific rally (independent of the server). (See [K,J], p. 267.) Determine, to within $10^{-3}$, the minimal value of $p$ that will ensure that A will shut out B in at least half the matches they play.

16. The function described by $f(x) = \ln(x^2 + 1) - e^{0.4x} \cos \pi x$ has an infinite number of zeros.

(a) Determine, within $10^{-6}$, the only negative zero.

(b) Determine, within $10^{-6}$, the four smallest positive zeros.

(c) Determine a reasonable initial approximation to find the $n$th smallest positive zero of $f$. [Hint: Sketch an approximate graph of $f$.]
(d) Use part (c) to determine, within $10^{-6}$, the 25th smallest positive zero of $f$.

17. The accumulated value of a savings account based on regular periodic payments can be determined from the *annuity due equation*,

$$A = \frac{P}{i}[(1 + i)^n - 1].$$

In this equation $A$ is the amount in the account, $P$ is the amount regularly deposited, and $i$ is the rate of interest per period for the $n$ deposit periods. An engineer would like to have a savings account valued at $750,000 upon retirement in 20 years and can afford to put $1500 per month toward this goal. What is the minimal interest rate at which this amount can be invested, assuming that the interest is compounded monthly?

18. Problems involving the amount of money required to pay off a mortgage over a fixed period of time involve the formula

$$A = \frac{P}{i}[1 - (1 + i)^{-n}],$$

known as an *ordinary annuity equation*. In this equation $A$ is the amount of the mortgage, $P$ is the amount of each payment, and $i$ is the interest rate per period for the $n$ payment periods. Suppose that a 30-year home mortgage in the amount of $135,000 is needed and that the borrower can afford house payments of at most $1000 per month. What is the maximal interest rate the borrower can afford to pay?

19. A drug administered to a patient produces a concentration in the bloodstream given by $c(t) = Ate^{-t/3}$ milligrams per milliliter $t$ hours after $A$ units have been injected. The maximum safe concentration is 1 mg/ml.

(a) What amount should be injected to reach this maximum safe concentration and when does this maximum occur?

(b) An additional amount of this drug is to be administered to the patient after the concentration falls to 0.25 mg/ml. Determine, to the nearest minute, when this second injection should be given.

(c) Assuming that the concentration from consecutive injections is additive and that 75% of the amount originally injected is administered in the second injection, when is it time for the third injection?

20. Let $f(x) = 3^{3x+1} - 7 \cdot 5^{2x}$.

(a) Use the Maple commands `solve` and `fsolve` to try to find all roots of $f$.

(b) Plot $f(x)$ to find initial approximations to roots of $f$. 
(c) Use Newton’s method to find roots of $f$ to within $10^{-16}$.

(d) Find the exact solutions of $f(x) = 0$ algebraically.
2.5 Error Analysis and Accelerating Convergence

In the previous section we found that Newton’s method generally converges very rapidly if a sufficiently accurate initial approximation has been found. This rapid speed of convergence is due to the fact that Newton’s method produces \( \text{quadratically} \) convergent approximations.

A method that produces a sequence \( \{p_n\} \) of approximations that converge to a number \( p \) converges \textbf{linearly} if, for large values of \( n \), a constant \( 0 < M < 1 \) exists with

\[
|p - p_{n+1}| \leq M|p - p_n|.
\]

The sequence converges \textbf{quadratically} if, for large values of \( n \), a constant \( 0 < M \) exists with

\[
|p - p_{n+1}| \leq M|p - p_n|^2.
\]

The following example illustrates the advantage of quadratic over linear convergence.

**EXAMPLE 1** Suppose that \( \{p_n\} \) converges linearly to \( p = 0 \), \( \hat{p}_n \) converges quadratically to \( p = 0 \), and the constant \( M = 0.5 \) is the same in each case. Then

\[
|p_1| \leq M|p_0| \leq (0.5) \cdot |p_0| \quad \text{and} \quad |\hat{p}_1| \leq M|\hat{p}_0|^2 \leq (0.5) \cdot |\hat{p}_0|^2.
\]

Similarly,

\[
|p_2| \leq M|p_1| \leq 0.5(0.5) \cdot |p_0| = (0.5)^2|p_0|
\]

and

\[
|\hat{p}_2| \leq M|\hat{p}_1|^2 \leq 0.5(0.5)|\hat{p}_0|^2 \leq (0.5)^3|p_0|^4.
\]

Continuing,

\[
|p_3| \leq M|p_2| \leq 0.5((0.5)^2|p_0|) = (0.5)^3|p_0|
\]

and

\[
|\hat{p}_3| \leq M|\hat{p}_2|^2 \leq 0.5((0.5)^3|\hat{p}_0|^4) = (0.5)^7|p_0|^8.
\]

In general,

\[
|p_n| \leq 0.5^n|p_0|, \quad \text{whereas} \quad |\hat{p}_n| \leq (0.5)^{2^n-1}|p_0|^{2^n}
\]

for each \( n = 1, 2, \ldots \). Table 2.7 illustrates the relative speed of convergence of these error bounds to zero, assuming that \( |p_0| = |\hat{p}_0| = 1 \).
2.5. ERROR ANALYSIS AND ACCELERATING CONVERGENCE

Table 2.7

<table>
<thead>
<tr>
<th>$n$</th>
<th>Linear Convergence Sequence Bound $p_n = (0.5)^n$</th>
<th>Quadratic Convergence Sequence Bound $\hat{p}_n = (0.5)^{2^n-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$5.0000 \times 10^{-1}$</td>
<td>$5.0000 \times 10^{-1}$</td>
</tr>
<tr>
<td>2</td>
<td>$2.5000 \times 10^{-1}$</td>
<td>$1.2500 \times 10^{-1}$</td>
</tr>
<tr>
<td>3</td>
<td>$1.2500 \times 10^{-1}$</td>
<td>$7.8125 \times 10^{-3}$</td>
</tr>
<tr>
<td>4</td>
<td>$6.2500 \times 10^{-2}$</td>
<td>$3.0518 \times 10^{-5}$</td>
</tr>
<tr>
<td>5</td>
<td>$3.1250 \times 10^{-2}$</td>
<td>$4.6566 \times 10^{-10}$</td>
</tr>
<tr>
<td>6</td>
<td>$1.5625 \times 10^{-2}$</td>
<td>$1.0842 \times 10^{-19}$</td>
</tr>
<tr>
<td>7</td>
<td>$7.8125 \times 10^{-3}$</td>
<td>$5.8775 \times 10^{-39}$</td>
</tr>
</tbody>
</table>

The quadratically convergent sequence is within $10^{-38}$ of zero by the seventh term. At least 126 terms are needed to ensure this accuracy for the linearly convergent sequence. If $|\hat{p}_0| < 1$, the bound on the sequence $\{\hat{p}_n\}$ will decrease even more rapidly. No significant change will occur, however, if $|p_0| < 1$.

As illustrated in Example 1, quadratically convergent sequences generally converge much more quickly than those that converge only linearly. However, linearly convergent methods are much more common than those that converge quadratically. Aitken’s $\Delta^2$ method is a technique that can be used to accelerate the convergence of a sequence that is linearly convergent, regardless of its origin or application.

Suppose $\{p_n\}_{n=0}^\infty$ is a linearly convergent sequence with limit $p$. To motivate the construction of a sequence $\{q_n\}$ that converges more rapidly to $p$ than does $\{p_n\}$, let us first assume that the signs of $p_n - p$, $p_{n+1} - p$, and $p_{n+2} - p$ agree and that $n$ is sufficiently large that

$$\frac{p_{n+1} - p}{p_n - p} \approx \frac{p_{n+2} - p}{p_{n+1} - p}.$$ 

Then

$$(p_{n+1} - p)^2 \approx (p_{n+2} - p)(p_n - p),$$ 

so

$$p_{n+1}^2 - 2p_{n+1}p + p^2 \approx p_{n+2}p_n - (p_n + p_{n+2})p + p^2$$ 

and

$$(p_{n+2} + p_n - 2p_{n+1})p \approx p_{n+2}p_n - p_{n+1}^2.$$ 

Solving for $p$ gives

$$p \approx \frac{p_{n+2}p_n - p_{n+1}^2}{p_{n+2} - 2p_{n+1} + p_n}.$$
Adding and subtracting the terms $p_n^2$ and $2p_np_{n+1}$ in the numerator and grouping terms appropriately gives

$$p \approx \frac{p_n p_{n+2} - 2p_n p_{n+1} + p_n^2 - p_{n+1}^2 + 2p_n p_{n+1} - p_n^2}{p_{n+2} - 2p_{n+1} + p_n} = \frac{p_n (p_{n+2} - 2p_{n+1} + p_n) - (p_{n+1}^2 - 2p_n p_{n+1} + p_n^2)}{p_{n+2} - 2p_{n+1} + p_n} = p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}.$$  

Aitken’s $\Delta^2$ method uses the sequence $\{q_n\}_{n=0}^\infty$ defined by this approximation to $p$.

### [Aitken’s $\Delta^2$ Method]

If $\{p_n\}_{n=0}^\infty$ is a sequence that converges linearly to $p$, and if

$$q_n = p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n},$$

then $\{q_n\}_{n=0}^\infty$ also converges to $p$, and, in general, more rapidly.

#### EXAMPLE 2

The sequence $\{p_n\}_{n=1}^\infty$, where $p_n = \cos(1/n)$, converges linearly to $p = 1$. The first few terms of the sequences $\{p_n\}_{n=1}^\infty$ and $\{q_n\}_{n=1}^\infty$ are given in Table 2.8. It certainly appears that $\{q_n\}_{n=1}^\infty$ converges more rapidly to $p = 1$ than does $\{p_n\}_{n=1}^\infty$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p_n$</th>
<th>$q_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.54030</td>
<td>0.96178</td>
</tr>
<tr>
<td>2</td>
<td>0.87758</td>
<td>0.98213</td>
</tr>
<tr>
<td>3</td>
<td>0.94496</td>
<td>0.98979</td>
</tr>
<tr>
<td>4</td>
<td>0.96891</td>
<td>0.99342</td>
</tr>
<tr>
<td>5</td>
<td>0.98007</td>
<td>0.99541</td>
</tr>
<tr>
<td>6</td>
<td>0.98614</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.98981</td>
<td></td>
</tr>
</tbody>
</table>

For a given sequence $\{p_n\}_{n=0}^\infty$, the **forward difference**, $\Delta p_n$ (read “delta $p_n$”), is defined as

$$\Delta p_n = p_{n+1} - p_n, \quad \text{for} \ n \geq 0.$$  

Higher powers of the operator $\Delta$ are defined recursively by

$$\Delta^k p_n = \Delta(\Delta^{k-1} p_n), \quad \text{for} \ k \geq 2.$$
2.5. ERROR ANALYSIS AND ACCELERATING CONVERGENCE

The definition implies that

\[ \Delta^2 p_n = \Delta(p_{n+1} - p_n) = \Delta p_{n+1} - \Delta p_n = (p_{n+2} - p_{n+1}) - (p_{n+1} - p_n), \]

so

\[ \Delta^2 p_n = p_{n+2} - 2p_{n+1} + p_n. \]

Thus, the formula for \( q_n \) given in Aitken’s \( \Delta^2 \) method can be written as

\[ q_n = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n}, \quad \text{for all } n \geq 0. \]

The sequence \( \{q_n\}_{n=1}^\infty \) converges to \( p \) more rapidly than does the original sequence \( \{p_n\}_{n=0}^\infty \) in the following sense:

\[ \text{[Aitken’s } \Delta^2 \text{ Convergence] If } \{p_n\} \text{ is a sequence that converges linearly to the limit } p \text{ and } (p_n - p)(p_{n+1} - p) > 0 \text{ for large values of } n, \text{ and} \]

\[ q_n = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n}, \quad \text{then} \quad \lim_{n \to \infty} \frac{q_n - p}{p_n - p} = 0. \]

We will find occasion to apply this acceleration technique at various times in our study of approximation methods.
EXERCISE SET 2.5

1. The following sequences are linearly convergent. Generate the first five terms of the sequence \( \{q_n\} \) using Aitken’s \( \Delta^2 \) method.

   (a) \( p_0 = 0.5 \), \( p_n = (2 - e^{p_{n-1}} + p_{n-1}^2)/3 \), for \( n \geq 1 \)

   (b) \( p_0 = 0.75 \), \( p_n = (e^{p_{n-1}}/3)^{1/2} \), for \( n \geq 1 \)

   (c) \( p_0 = 0.5 \), \( p_n = 3 - p_{n-1} \), for \( n \geq 1 \)

   (d) \( p_0 = 0.5 \), \( p_n = \cos p_{n-1} \), for \( n \geq 1 \)

2. Newton’s method does not converge quadratically for the following problems. Accelerate the convergence using the Aitken’s \( \Delta^2 \) method. Iterate until \( |q_n - q_{n-1}| < 10^{-4} \).

   (a) \( x^2 - 2xe^{-x} + e^{-2x} = 0 \), \([-1, 0]\)

   (b) \( \cos(x + \sqrt{2}) + x(x/2 + \sqrt{2}) = 0 \), \([-2, -1]\)

   (c) \( x^3 - 3x^2(2^{-x}) + 3x(4^{-x}) - 8^{-x} = 0 \), \([0, 1]\)

   (d) \( e^{6x} + 3(\ln 2)^2 e^{2x} - (\ln 8)e^{4x} - (2)^3 = 0 \), \([-1, 0]\)

3. Consider the function \( f(x) = e^{6x} + 3(\ln 2)^2 e^{2x} - (\ln 8)e^{4x} - (2)^3 \). Use Newton’s method with \( p_0 = 0 \) to approximate a zero of \( f \). Generate terms until \( |p_n - p_{n-1}| < 0.0002 \). Construct the Aitken’s \( \Delta^2 \) sequence \( \{q_n\} \). Is the convergence improved?

4. Repeat Exercise 3 with the constants in \( f(x) \) replaced by their four-digit approximations, that is, with \( f(x) = e^{6x} + 1.441e^{2x} - 2.079e^{4x} - 0.3330 \), and compare the solutions to the results in Exercise 3.

5. (i) Show that the following sequences \( \{p_n\} \) converge linearly to \( p = 0 \). (ii) How large must \( n \) be before \( |p_n - p| \leq 5 \times 10^{-2} \)? (iii) Use Aitken’s \( \Delta^2 \) method to generate a sequence \( \{q_n\} \) until \( |q_n - p| \leq 5 \times 10^{-2} \).

   (a) \( p_n = \frac{1}{n} \), for \( n \geq 1 \)

   (b) \( p_n = \frac{1}{n^2} \), for \( n \geq 1 \)

6. (a) Show that for any positive integer \( k \), the sequence defined by \( p_n = 1/n^k \) converges linearly to \( p = 0 \).

   (b) For each pair of integers \( k \) and \( m \), determine a number \( N \) for which \( 1/N^k < 10^{-m} \).

7. (a) Show that the sequence \( p_n = 10^{-2^n} \) converges quadratically to zero.
2.5. ERROR ANALYSIS AND ACCELERATING CONVERGENCE

(b) Show that the sequence \( p_n = 10^{-n^k} \) does not converge to zero quadratically, regardless of the size of the exponent \( k > 1 \).

8. A sequence \( \{ p_n \} \) is said to be **superlinearly convergent** to \( p \) if a sequence \( \{ c_n \} \) converging to zero exists with

\[
|p_{n+1} - p| \leq c_n |p_n - p|.
\]

(a) Show that if \( \{ p_n \} \) is superlinearly convergent to \( p \), then \( \{ p_n \} \) is linearly convergent to \( p \).

(b) Show that \( p_n = 1/n^n \) is superlinearly convergent to zero but is not quadratically convergent to zero.
2.6 Müller’s Method

There are a number of root-finding problems for which the Secant, False Position, and Newton’s methods will not give satisfactory results. They will not give rapid convergence, for example, when the function and its derivative are simultaneously close to zero. In addition, these methods cannot be used to approximate complex roots unless the initial approximation is a complex number whose imaginary part is nonzero. This often makes them a poor choice for use in approximating the roots of polynomials, which, even with real coefficients, commonly have complex roots occurring in conjugate pairs.

In this section we consider Müller’s method, which is a generalization of the Secant method. The Secant method finds the zero of the line passing through points on the graph of the function that corresponds to the two immediately previous approximations, as shown in Figure 2.8(a). Müller’s method uses the zero of the parabola through the three immediately previous points on the graph as the new approximation, as shown in part (b) of Figure 2.8.

Figure 2.8

Suppose that three initial approximations, \( p_0, p_1, \) and \( p_2, \) are given for a solution of \( f(x) = 0. \) The derivation of Müller’s method for determining the next approximation \( p_3 \) begins by considering the quadratic polynomial

\[
P(x) = a(x - p_2)^2 + b(x - p_2) + c
\]

that passes through \( (p_0, f(p_0)), (p_1, f(p_1)), \) and \( (p_2, f(p_2)). \) The constants \( a, b, \) and \( c \) can be determined from the conditions

\[
f(p_0) = a(p_0 - p_2)^2 + b(p_0 - p_2) + c,
\]
\[
f(p_1) = a(p_1 - p_2)^2 + b(p_1 - p_2) + c,
\]
2.6. MÜLLER’S METHOD

and

\[ f(p_2) = a \cdot 0^2 + b \cdot 0 + c. \]

To determine \( p_3 \), the root of \( P(x) = 0 \), we apply the quadratic formula to \( P(x) \). Because of round-off error problems caused by the subtraction of nearly equal numbers, however, we apply the formula in the manner prescribed in Example 1 of Section 1.4:

\[ p_3 - p_2 = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}. \]

This gives two possibilities for \( p_3 \), depending on the sign preceding the radical term. In Müller’s method, the sign is chosen to agree with the sign of \( b \). Chosen in this manner, the denominator will be the largest in magnitude, which avoids the possibility of subtracting nearly equal numbers and results in \( p_3 \) being selected as the closest root of \( P(x) = 0 \) to \( p_2 \).

[Müller’s Method] Given initial approximations \( p_0, p_1, \) and \( p_2 \), generate

\[ p_3 = p_2 - \frac{2c}{b + \text{sgn}(b) \sqrt{b^2 - 4ac}}, \]

where

\[ c = f(p_2), \]

\[ b = \frac{(p_0 - p_2)^2 [f(p_1) - f(p_2)] - (p_1 - p_2)^2 [f(p_0) - f(p_2)]}{(p_0 - p_2)(p_1 - p_2)(p_0 - p_1)}, \]

and

\[ a = \frac{(p_1 - p_2)[f(p_0) - f(p_2)] - (p_0 - p_2)[f(p_1) - f(p_2)]}{(p_0 - p_2)(p_1 - p_2)(p_0 - p_1)}. \]

Then continue the iteration, with \( p_1, p_2, \) and \( p_3 \) replacing \( p_0, p_1, \) and \( p_2 \).

The method continues until a satisfactory approximation is obtained. Since the method involves the radical \( \sqrt{b^2 - 4ac} \) at each step, the method approximates complex roots when \( b^2 - 4ac < 0 \), provided, of course, that complex arithmetic is used.

EXAMPLE 1 Consider the polynomial \( f(x) = 16x^4 - 40x^3 + 5x^2 + 20x + 6 \). Using the program MULLER25 with accuracy tolerance \( 10^{-5} \) and various inputs for \( p_0, p_1, \) and \( p_2 \) produces the results in Tables 2.9, 2.10, and 2.11.

Table 2.9

\[ p_0 = 0.5, \ p_1 = -0.5, \ p_2 = 0 \]
CHAPTER 2. SOLUTIONS OF EQUATIONS OF ONE VARIABLE

Table 2.10

\[
p_0 = 0.5, \ p_1 = 1.0, \ p_2 = 1.5
\]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( p_n )</th>
<th>( f(p_n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>-0.555556 + 0.598352i</td>
<td>-29.4007 - 3.89872i</td>
</tr>
<tr>
<td>4</td>
<td>-0.435450 + 0.102101i</td>
<td>1.33223 - 1.19309i</td>
</tr>
<tr>
<td>5</td>
<td>-0.390631 + 0.141852i</td>
<td>0.375057 - 0.670164i</td>
</tr>
<tr>
<td>6</td>
<td>-0.357699 + 0.169926i</td>
<td>-0.146746 - 0.00744629i</td>
</tr>
<tr>
<td>7</td>
<td>-0.356051 + 0.162856i</td>
<td>-0.183868 \times 10^{-2} + 0.539780 \times 10^{-3}i</td>
</tr>
<tr>
<td>8</td>
<td>-0.356062 + 0.162758i</td>
<td>0.286102 \times 10^{-5} + 0.953674 \times 10^{-6}i</td>
</tr>
</tbody>
</table>

Table 2.11

\[
p_0 = 2.5, \ p_1 = 2.0, \ p_2 = 2.25
\]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( p_n )</th>
<th>( f(p_n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.28785</td>
<td>-1.37624</td>
</tr>
<tr>
<td>4</td>
<td>1.23746</td>
<td>0.126941</td>
</tr>
<tr>
<td>5</td>
<td>1.24160</td>
<td>0.219440 \times 10^{-2}</td>
</tr>
<tr>
<td>6</td>
<td>1.24168</td>
<td>0.257492 \times 10^{-4}</td>
</tr>
<tr>
<td>7</td>
<td>1.24168</td>
<td>0.257492 \times 10^{-4}</td>
</tr>
</tbody>
</table>

To use Maple to generate the first entry in Table 2.9 we define \( f(x) \) and the

<table>
<thead>
<tr>
<th>Maple statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;f := x-&gt;16<em>x^4-40</em>x^3+5<em>x^2+20</em>x+6;</td>
</tr>
<tr>
<td>&gt;p0:=0.5; p1:=-0.5; p2:=0.0;</td>
</tr>
</tbody>
</table>

We evaluate the polynomial at the initial values

<table>
<thead>
<tr>
<th>Maple statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;f0:=f(p0); f1:=f(p1); f2:=f(p2);</td>
</tr>
</tbody>
</table>

and we compute \( c = 6, \ b = 10, \ a = 9 \), and \( p_3 = -0.5555555558 + 0.5983516452i \)

using the Müller’s method formulas:
2.6. MÜLLER’S METHOD

\[ c := f_2; \]
\[ b := ((p_0 - p_2)^2 (f_1 - f_2) - (p_1 - p_2)^2 (f_0 - f_2)) / ((p_0 - p_2) (p_1 - p_2) (p_0 - p_1)); \]
\[ a := ((p_1 - p_2) (f_0 - f_2) - (p_0 - p_2) (f_1 - f_2)) / ((p_0 - p_2) (p_1 - p_2) (p_0 - p_1)); \]
\[ p_3 := p_2 - (2c) / (b + (b/abs(b)) * sqrt(b^2 - 4*a*c)); \]

The value \( p_3 \) was generated using complex arithmetic, as is the calculation

\[ f_3 := f(p_3); \]

which gives \( f_3 = -29.40070112 - 3.898724738i \).

The actual values for the roots of the equation are \(-0.356062 \pm 0.162758i\), \(1.241677\), and \(1.970446\), which demonstrate the accuracy of the approximations from Müller’s method.

Example 1 illustrates that Müller’s method can approximate the roots of polynomials with a variety of starting values. In fact, the technique generally converges to the root of a polynomial for any initial approximation choice. General-purpose software packages using Müller’s method request only one initial approximation per root and, as an option, may even supply this approximation.

Although Müller’s method is not quite as efficient as Newton’s method, it is generally better than the Secant method. The relative efficiency, however, is not as important as the ease of implementation and the likelihood that a root will be found. Any of these methods will converge quite rapidly once a reasonable initial approximation is determined.

When a sufficiently accurate approximation \( p^* \) to a root has been found, \( f(x) \) is divided by \( x - p^* \) to produce what is called a deflated equation. If \( f(x) \) is a polynomial of degree \( n \), the deflated polynomial will be of degree \( n - 1 \), so the computations are simplified. After an approximation to the root of the deflated equation has been determined, either Müller’s method or Newton’s method can be used in the original function with this root as the initial approximation. This will ensure that the root being approximated is a solution to the true equation, not to the less accurate deflated equation.
EXERCISE SET 2.6

1. Find the approximations to within $10^{-4}$ to all the real zeros of the following polynomials using Newton’s method.

(a) $P(x) = x^3 - 2x^2 - 5$
(b) $P(x) = x^3 + 3x^2 - 1$
(c) $P(x) = x^4 + 2x^2 - x - 3$
(d) $P(x) = x^5 - x^4 + 2x^3 - 3x^2 + x - 4$

2. Find approximations to within $10^{-5}$ to all the zeros of each of the following polynomials by first finding the real zeros using Newton’s method and then reducing to polynomials of lower degree to determine any complex zeros.

(a) $P(x) = x^4 + 5x^3 - 9x^2 - 85x - 136$
(b) $P(x) = x^4 - 2x^3 - 12x^2 + 16x - 40$
(c) $P(x) = x^4 + x^3 + 3x^2 + 2x + 2$
(d) $P(x) = x^5 + 11x^4 - 21x^3 - 10x^2 - 21x - 5$

3. Repeat Exercise 1 using Müller’s method.

4. Repeat Exercise 2 using Müller’s method.

5. Find, to within $10^{-3}$, the zeros and critical points of the following functions. Use this information to sketch the graphs of $P$.

(a) $P(x) = x^3 - 9x^2 + 12$
(b) $P(x) = x^4 - 2x^3 - 5x^2 + 12x - 5$

6. $P(x) = 10x^3 - 8.3x^2 + 2.295x - 0.21141 = 0$ has a root at $x = 0.29$.

(a) Use Newton’s method with $p_0 = 0.28$ to attempt to find this root.

(b) Use Müller’s method with $p_0 = 0.275$, $p_1 = 0.28$, and $p_2 = 0.285$ to attempt to find this root.

(c) Explain any discrepancies in parts (a) and (b).

7. Use Maple to find the exact roots of the polynomial $P(x) = x^3 + 4x - 4$.

8. Use Maple to find the exact roots of the polynomial $P(x) = x^3 - 2x - 5$.

9. Use each of the following methods to find a solution accurate to within $10^{-4}$ for the problem

$$600x^4 - 550x^3 + 200x^2 - 20x - 1 = 0, \quad \text{for } 0.1 \leq x \leq 1.$$
2.6. MÜLLER’S METHOD

(a) Bisection method
(b) Newton’s method
(c) Secant method
(d) method of False Position
(e) Müller’s method

10. Two ladders crisscross an alley of width $W$. Each ladder reaches from the base of one wall to some point on the opposite wall. The ladders cross at a height $H$ above the pavement. Find $W$ given that the lengths of the ladders are $x_1 = 20$ ft and $x_2 = 30$ ft and that $H = 8$ ft. (See the figure on page 58.)

11. A can in the shape of a right circular cylinder is to be constructed to contain 1000 cm$^3$. The circular top and bottom of the can must have a radius of 0.25 cm more than the radius of the can so that the excess can be used to form a seal with the side. The sheet of material being formed into the side of the can must also be 0.25 cm longer than the circumference of the can so that a seal can be formed. Find, to within $10^{-4}$, the minimal amount of material needed to construct the can.

12. In 1224 Leonardo of Pisa, better known as Fibonacci, answered a mathematical challenge of John of Palermo in the presence of Emperor Frederick II. His challenge was to find a root of the equation $x^3 + 2x^2 + 10x = 20$. He first showed that the equation had no rational roots and no Euclidean irrational root—that is, no root in one of the forms $a \pm \sqrt{b}$, $\sqrt{a \pm \sqrt{b}}$, or $\sqrt[3]{a \pm \sqrt{b}}$, where $a$ and $b$ are rational numbers. He then approximated the only real root, probably using an algebraic technique of Omar Khayyam.
involving the intersection of a circle and a parabola. His answer was given in the base-60 number system as

$$1 + 22 \left( \frac{1}{60} \right) + 7 \left( \frac{1}{60} \right)^2 + 42 \left( \frac{1}{60} \right)^3 + 33 \left( \frac{1}{60} \right)^4 + 4 \left( \frac{1}{60} \right)^5 + 40 \left( \frac{1}{60} \right)^6.$$ 

How accurate was his approximation?
2.7 Survey of Methods and Software

In this chapter we have considered the problem of solving the equation \( f(x) = 0 \), where \( f \) is a given continuous function. All the methods begin with an initial approximation and generate a sequence that converges to a root of the equation, if the method is successful. If \( [a, b] \) is an interval on which \( f(a) \) and \( f(b) \) are of opposite sign, then the Bisection method and the method of False Position will converge. However, the convergence of these methods may be slow. Faster convergence is generally obtained using the Secant method or Newton’s method. Good initial approximations are required for these methods, two for the Secant method and one for Newton’s method, so the Bisection or the False Position method can be used as starter methods for the Secant or Newton’s method.

Müller’s method will give rapid convergence without a particularly good initial approximation. It is not quite as efficient as Newton’s method, but it is better than the Secant method, and it has the added advantage of being able to approximate complex roots.

Deflation is generally used with Müller’s method once an approximate root of a polynomial has been determined. After an approximation to the root of the deflated equation has been determined, use either Müller’s method or Newton’s method in the original polynomial with this root as the initial approximation. This procedure will ensure that the root being approximated is a solution to the true equation, not to the deflated equation. We recommended Müller’s method for finding all the zeros of polynomials, real or complex. Müller’s method can also be used for an arbitrary continuous function.

Other high-order methods are available for determining the roots of polynomials. If this topic is of particular interest, we recommend that consideration be given to Laguerre’s method, which gives cubic convergence and also approximates complex roots (see [Ho, pp. 176–179] for a complete discussion), the Jenkins-Traub method (see [JT]), and Brent’s method. (see [Bre]). Both IMSL and NAG supply subroutines based on Brent’s method. This technique uses a combination of linear interpolation, an inverse quadratic interpolation similar to Müller’s method, and the bisection method.

The netlib FORTRAN subroutine fzero.f uses a combination of the Bisection and Secant method developed by T. J. Dekker to approximate a real zero of \( f(x) = 0 \) in the interval \([a, b]\). It requires specifying an interval \([a, b]\) that contains a root and returns an interval with a width that is within a specified tolerance. The FORTRAN subroutine sdzro.f uses a combination of the bisection method, interpolation, and extrapolation to find a real zero of \( f(x) = 0 \) in a given interval \([a, b]\). The routines rpzero and cpzero can be used to approximate all zeros of a real polynomial or complex polynomial, respectively. Both methods use Newton’s method for systems, which will be considered in Chapter 10. All routines are given in single and double precision. These methods are available on the Internet from netlib at http://www.netlib.org/slatec/src.

Within MATLAB, the function ROOTS is used to compute all the roots, both real and complex, of a polynomial. For an arbitrary function, FZERO computes a root near a specified initial approximation to within a specified tolerance.
Maple has the procedure \texttt{fsolve} to find roots of equations. For example,
\begin{verbatim}
>fsolve(x^2 - x - 1, x);
\end{verbatim}
returns the numbers \(-0.6180339887\) and \(1.618033989\). You can also specify a particular variable and interval to search. For example,
\begin{verbatim}
>fsolve(x^2 - x - 1, x, 1..2);
\end{verbatim}
returns only the number \(1.618033989\). The command \texttt{fsolve} uses a variety of specialized techniques that depend on the particular form of the equation or system of equations.

Notice that in spite of the diversity of methods, the professionally written packages are based primarily on the methods and principles discussed in this chapter. You should be able to use these packages by reading the manuals accompanying the packages to better understand the parameters and the specifications of the results that are obtained.

There are three books that we consider to be classics on the solution of nonlinear equations, those by Traub [Tr], by Ostrowski [Os], and by Householder [Ho]. In addition, the book by Brent [Bre] served as the basis for many of the currently used root-finding methods.