CHAPTER

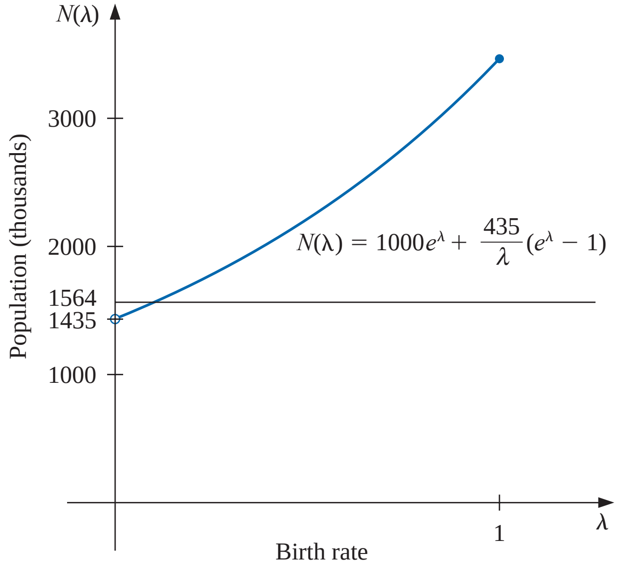
2 Solutions of Equations in One Variable

Introduction

The growth of a population can often be modeled over short periods of time by assuming that the population grows continuously with time at a rate proportional to the number present at that time. Suppose that $N(t)$ denotes the number in the population at time t and λ denotes the constant birth rate of the population. Then the population satisfies the differential equation

$$\frac{dN(t)}{dt} = \lambda N(t),$$

whose solution is $N(t) = N_0 e^{\lambda t}$, where N_0 denotes the initial population.



This exponential model is valid only when the population is isolated, with no immigration. If immigration is permitted at a constant rate v, then the differential equation becomes

$$\frac{dN(t)}{dt} = \lambda N(t) + v,$$

whose solution is

$$N(t) = N_0 e^{\lambda t} + \frac{v}{\lambda}(e^{\lambda t} - 1).$$

Suppose a certain population contains $N(0) = 1{,}000{,}000$ individuals initially, that 435,000 individuals immigrate into the community in the first year, and that $N(1) = 1{,}564{,}000$ individuals are present at the end of one year. To determine the birth rate of this population, we need to find λ in the equation

$$1{,}564{,}000 = 1{,}000{,}000e^{\lambda} + \frac{435{,}000}{\lambda}(e^{\lambda} - 1).$$

It is not possible to solve explicitly for λ in this equation, but numerical methods discussed in this chapter can be used to approximate solutions of equations of this type to an arbitrarily high accuracy. The solution to this particular problem is considered in Exercise 24 of Section 2.3.

2.1 The Bisection Method

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 $f(x) = 0$, for a given function f. A root of this equation is also called a **zero** of the function f.

The problem of finding an approximation to the root of an equation can be traced back at least to 1700 B.C.E. A cuneiform table in the Yale Babylonian Collection dating from that period gives a sexigesimal (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 described in Exercise 19 of Section 2.2.

Bisection Technique

In computer science, the process of dividing a set continually in half to search for the solution to a problem, as the bisection method does, is known as a *binary search* procedure.

The first technique, based on the Intermediate Value Theorem, is called the **Bisection**, or **Binary-search, method**.

Suppose f is a continuous function defined on the interval $[a, b]$, with $f(a)$ and $f(b)$ of opposite sign. The Intermediate Value Theorem implies that a number p exists in (a, b) with $f(p) = 0$. Although the procedure will work when there is more than one root in the interval (a, b), we assume for simplicity that the root in this interval is unique. The method calls for a repeated halving (or bisecting) of subintervals of $[a, b]$ and, at each step, locating the half containing p.

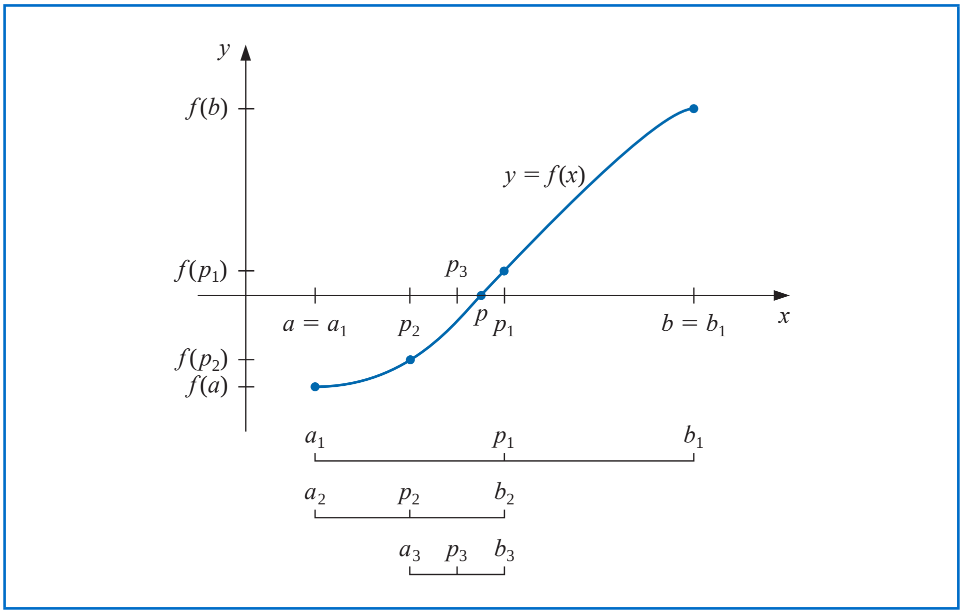
To begin, set $a_1 = a$ and $b_1 = b$, and let p_1 be the midpoint of $[a, b]$; that is,

$$p_1 = a_1 + \frac{b_1 - a_1}{2} = \frac{a_1 + b_1}{2}.$$

- If $f(p_1) = 0$, then $p = p_1$, and we are done.
- If $f(p_1) \neq 0$, then $f(p_1)$ has the same sign as either $f(a_1)$ or $f(b_1)$.
    - If $f(p_1)$ and $f(a_1)$ have the same sign, $p \in (p_1, b_1)$. Set $a_2 = p_1$ and $b_2 = b_1$.
    - If $f(p_1)$ and $f(a_1)$ have opposite signs, $p \in (a_1, p_1)$. Set $a_2 = a_1$ and $b_2 = p_1$.

Then reapply the process to the interval $[a_2, b_2]$. This produces the method described in Algorithm 2.1. (See Figure 2.1.)

Figure 2.1



ALGORITHM 2.1

Bisection

To find a solution to $f(x) = 0$ given the continuous function f on the interval $[a, b]$, where $f(a)$ and $f(b)$ have opposite signs:

```
INPUT   endpoints $a, b$; tolerance $TOL$; maximum number of iterations $N_0$.

OUTPUT   approximate solution $p$ or message of failure.

Step 1   Set $i = 1$;
             $FA = f(a)$.

Step 2   While $i \leq N_0$ do Steps 3–6.

    Step 3   Set $p = a + (b - a)/2$;   (Compute $p_i$.)
                 $FP = f(p)$.
    Step 4   If $FP = 0$ or $(b - a)/2 < TOL$ then
                 OUTPUT $(p)$;   (Procedure completed successfully.)
                 STOP.

    Step 5   Set $i = i + 1$.

    Step 6   If $FA \cdot FP > 0$ then set $a = p$;   (Compute $a_i, b_i$.)
                                           $FA = FP$
                                  else set $b = p$.   (FA is unchanged.)

Step 7   OUTPUT ('Method failed after $N_0$ iterations, $N_0$ =', $N_0$);
         (The procedure was unsuccessful.)
         STOP.
```

■

Other stopping procedures can be applied in Step 4 of Algorithm 2.1 or in any of the iterative techniques in this chapter. For example, we can select a tolerance $\varepsilon > 0$ and generate $p_1, \ldots, p_N$ until one of the following conditions is met:

$$|p_N - p_{N-1}| < \varepsilon, \tag{2.1}$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \varepsilon, \quad p_N \neq 0, \quad \text{or} \tag{2.2}$$

$$|f(p_N)| < \varepsilon. \tag{2.3}$$

Unfortunately, difficulties can arise using any of these stopping criteria. For example, there are sequences $\{p_n\}_{n=0}^{\infty}$ with the property that the differences $p_n - p_{n-1}$ converge to zero while the sequence itself diverges. (See Exercise 17.) It is also possible for $f(p_n)$ to be close to zero while p_n differs significantly from p. (See Exercise 16.) Without additional knowledge about f or p, Inequality (2.2) is the best stopping criterion to apply because it comes closest to testing relative error.

When using a computer to generate approximations, it is good practice to set an upper bound on the number of iterations. This eliminates the possibility of entering an infinite loop, a situation that can arise when the sequence diverges (and also when the program is incorrectly coded). This was done in Step 2 of Algorithm 2.1 where the bound N_0 was set and the procedure terminated if $i > N_0$.

Note that to start the Bisection Algorithm, 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; hence it is advantageous to choose the initial interval $[a, b]$ 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 Algorithm could be used on $[-4, 4]$ or on $[0, 1]$. Starting the Bisection Algorithm on $[0, 1]$ instead of $[-4, 4]$ will reduce by 3 the number of iterations required to achieve a specified accuracy.

The following example illustrates the Bisection Algorithm. The iteration in this example is terminated when a bound for the relative error is less than 0.0001. This is ensured by having

$$\frac{|p - p_n|}{\min\{|a_n|, |b_n|\}} < 10^{-4}.$$

Example 1 Show that $f(x) = x^3 + 4x^2 - 10 = 0$ has a root in $[1, 2]$, and use the Bisection method to determine an approximation to the root that is accurate to at least within 10^{-4}.

Solution Because $f(1) = -5$ and $f(2) = 14$ the Intermediate Value Theorem 1.11 ensures that this continuous function has a root in $[1, 2]$.

For the first iteration of the Bisection method we use the fact that at the midpoint of $[1, 2]$ we have $f(1.5) = 2.375 > 0$. This indicates that we should select the interval $[1, 1.5]$ for our second iteration. Then we find that $f(1.25) = -1.796875$ so our new interval becomes $[1.25, 1.5]$, whose midpoint is 1.375. Continuing in this manner gives the values in Table 2.1. After 13 iterations, $p_{13} = 1.365112305$ approximates the root p with an error

$$|p - p_{13}| < |b_{14} - a_{14}| = |1.365234375 - 1.365112305| = 0.000122070.$$

Since $|a_{14}| < |p|$, we have

$$\frac{|p - p_{13}|}{|p|} < \frac{|b_{14} - a_{14}|}{|a_{14}|} \leq 9.0 \times 10^{-5},$$

Table 2.1

n	a_n	b_n	p_n	$f(p_n)$
1	1.0	2.0	1.5	2.375
2	1.0	1.5	1.25	−1.79687
3	1.25	1.5	1.375	0.16211
4	1.25	1.375	1.3125	−0.84839
5	1.3125	1.375	1.34375	−0.35098
6	1.34375	1.375	1.359375	−0.09641
7	1.359375	1.375	1.3671875	0.03236
8	1.359375	1.3671875	1.36328125	−0.03215
9	1.36328125	1.3671875	1.365234375	0.000072
10	1.36328125	1.365234375	1.364257813	−0.01605
11	1.364257813	1.365234375	1.364746094	−0.00799
12	1.364746094	1.365234375	1.364990235	−0.00396
13	1.364990235	1.365234375	1.365112305	−0.00194

so the approximation is correct to at least within 10^{-4}. The correct value of p to nine decimal places is $p = 1.365230013$. Note that p_9 is closer to p than is the final approximation p_{13}. You might suspect this is true because $|f(p_9)| < |f(p_{13})|$, but we cannot be sure of this unless the true answer is known. ■

The Bisection method, though conceptually clear, has significant drawbacks. It is relatively slow to converge (that is, N may become quite large before $|p - p_N|$ is sufficiently small), and a good intermediate approximation might be inadvertently discarded. However, the method has the important property that it always converges to a solution, and for that reason it is often used as a starter for the more efficient methods we will see later in this chapter.

Theorem 2.1 Suppose that $f \in C[a, b]$ and $f(a) \cdot f(b) < 0$. The Bisection method generates a sequence $\{p_n\}_{n=1}^{\infty}$ approximating a zero p of f with

$$|p_n - p| \le \frac{b-a}{2^n}, \quad \text{when} \quad n \ge 1.$$

■

Proof For each $n \ge 1$, we have

$$b_n - a_n = \frac{1}{2^{n-1}}(b-a) \quad \text{and} \quad p \in (a_n, b_n).$$

Since $p_n = \frac{1}{2}(a_n + b_n)$ for all $n \ge 1$, it follows that

$$|p_n - p| \le \frac{1}{2}(b_n - a_n) = \frac{b-a}{2^n}.$$

▪ ▪ ▪

Because

$$|p_n - p| \le (b-a)\frac{1}{2^n},$$

the sequence $\{p_n\}_{n=1}^{\infty}$ converges to p with rate of convergence $O\left(\frac{1}{2^n}\right)$; that is,

$$p_n = p + O\left(\frac{1}{2^n}\right).$$

It is important to realize that Theorem 2.1 gives only a bound for approximation error and that this bound might be quite conservative. For example, this bound applied to the problem in Example 1 ensures only that

$$|p - p_9| \leq \frac{2-1}{2^9} \approx 2 \times 10^{-3},$$

but the actual error is much smaller:

$$|p - p_9| = |1.365230013 - 1.365234375| \approx 4.4 \times 10^{-6}.$$

Example 2 Determine the number of iterations necessary to solve $f(x) = x^3 + 4x^2 - 10 = 0$ with accuracy 10^{-3} using $a_1 = 1$ and $b_1 = 2$.

Solution We we will use logarithms to find an integer N that satisfies

$$|p_N - p| \leq 2^{-N}(b-a) = 2^{-N} < 10^{-3}.$$

Logarithms to any base would suffice, but we will use base-10 logarithms because the tolerance is given as a power of 10. Since $2^{-N} < 10^{-3}$ implies that $\log_{10} 2^{-N} < \log_{10} 10^{-3} = -3$, we have

$$-N \log_{10} 2 < -3 \quad \text{and} \quad N > \frac{3}{\log_{10} 2} \approx 9.96.$$

Hence, ten iterations will ensure an approximation accurate to within 10^{-3}.

Table 2.1 shows that the value of $p_9 = 1.365234375$ is accurate to within 10^{-4}. Again, it is important to keep in mind that the error analysis gives only a bound for the number of iterations. In many cases this bound is much larger than the actual number required. ■

Maple has a *NumericalAnalysis* package that implements many of the techniques we will discuss, and the presentation and examples in the package are closely aligned with this text. The Bisection method in this package has a number of options, some of which we will now consider. In what follows, Maple code is given in *black italic* type and Maple response in cyan.

Load the *NumericalAnalysis* package with the command

with(*Student*[*NumericalAnalysis*])

which gives access to the procedures in the package. Define the function with

$f := x^3 + 4x^2 - 10$

and use

Bisection $(f, x = [1, 2], \textit{tolerance} = 0.005)$

Maple returns

1.363281250

Note that the value that is output is the same as p_8 in Table 2.1.

The sequence of bisection intervals can be output with the command

Bisection $(f, x = [1, 2], \textit{tolerance} = 0.005, \textit{output} = \textit{sequence})$

and Maple returns the intervals containing the solution together with the solution

```
[1., 2.], [1., 1.500000000], [1.250000000, 1.500000000], [1.250000000, 1.375000000],
[1.312500000, 1.375000000], [1.343750000, 1.375000000], [1.359375000, 1.375000000],
[1.359375000, 1.367187500], 1.363281250
```

The stopping criterion can also be based on relative error by choosing the option

Bisection $(f, x = [1, 2], \textit{tolerance} = 0.005, \textit{stoppingcriterion} = \textit{relative})$

Now Maple returns

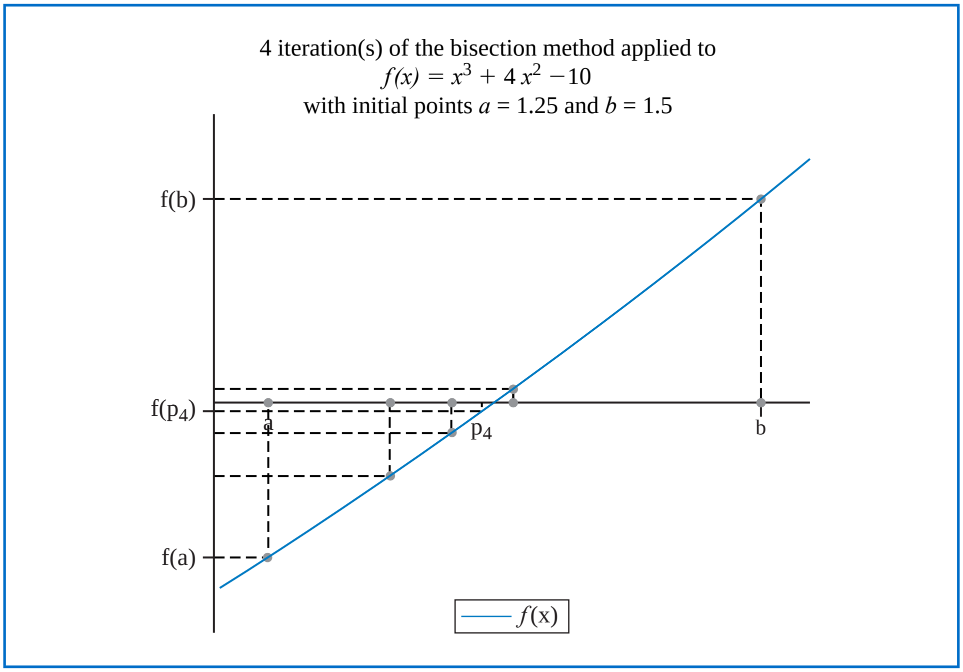
$$1.363281250$$

The option *output* = *plot* given in

$$Bisection\ (f, x = [1.25, 1.5], output = plot, tolerance = 0.02)$$

produces the plot shown in Figure 2.2.

Figure 2.2



We can also set the maximum number of iterations with the option *maxiterations* = . An error message will be displayed if the stated tolerance is not met within the specified number of iterations.

The results from Bisection method can also be obtained using the command Roots. For example,

$$Roots\left(f, x = [1.0, 2.0], method = bisection, tolerance = \frac{1}{100}, output = information\right)$$

uses the Bisection method to produce the information

n	a_n	b_n	p_n	$f\ (p_n)$	*relative error*
1	1.0	2.0	1.500000000	2.37500000	0.3333333333
2	1.0	1.500000000	1.250000000	−1.796875000	0.2000000000
3	1.250000000	1.500000000	1.375000000	0.16210938	0.09090909091
4	1.250000000	1.375000000	1.312500000	−0.848388672	0.04761904762
5	1.312500000	1.375000000	1.343750000	−0.350982668	0.02325581395
6	1.343750000	1.375000000	1.359375000	−0.096408842	0.01149425287
7	1.359375000	1.375000000	1.367187500	0.03235578	0.005714285714

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, we need to consider 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

$$p_n = a_n + \frac{b_n - a_n}{2} \quad \text{instead of} \quad p_n = \frac{a_n + b_n}{2}.$$

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, when $b_n - a_n$ is near the maximum precision of the machine, it is possible for $(a_n + b_n)/2$ to return a midpoint that is not even in the interval $[a_n, b_n]$.

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

The Latin word *signum* means "token" or "sign". So the signum function quite naturally returns the sign of a number (unless the number is 0).

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

The test

$$\operatorname{sgn}\,(f(a_n))\ \operatorname{sgn}\,(f(p_n)) < 0 \qquad \text{instead of} \qquad f(a_n)f(p_n) < 0$$

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 2.1

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. Use the Bisection method to find solutions accurate to within 10^{-5} for the following problems.

a. $x - 2^{-x} = 0 \quad \text{for } 0 \le x \le 1$

b. $e^x - x^2 + 3x - 2 = 0 \quad \text{for } 0 \le x \le 1$

c. $2x\cos(2x) - (x + 1)^2 = 0 \quad \text{for } -3 \le x \le -2 \quad \text{and} \quad -1 \le x \le 0$

d. $x\cos x - 2x^2 + 3x - 1 = 0 \quad \text{for } 0.2 \le x \le 0.3 \quad \text{and} \quad 1.2 \le x \le 1.3$

6. Use the Bisection method to find solutions, accurate to within 10^{-5} for the following problems.

a. $3x - e^x = 0$ for $1 \le x \le 2$

b. $2x + 3\cos x - e^x = 0 \quad \text{for } 0 \le x \le 1$

c. $x^2 - 4x + 4 - \ln x = 0 \quad \text{for } 1 \le x \le 2 \quad \text{and} \quad 2 \le x \le 4$

d. $x + 1 - 2\sin \pi x = 0 \quad \text{for } 0 \le x \le 0.5 \quad \text{and} \quad 0.5 \le x \le 1$

7. **a.** Sketch the graphs of $y = x$ and $y = 2\sin x$.

b. Use the Bisection method to find an approximation to within 10^{-5} to the first positive value of x with $x = 2\sin x$.

8. **a.** Sketch the graphs of $y = x$ and $y = \tan x$.

b. Use the Bisection method to find an approximation to within 10^{-5} to the first positive value of x with $x = \tan x$.

9. **a.** Sketch the graphs of $y = e^x - 2$ and $y = \cos(e^x - 2)$.

b. Use the Bisection method to find an approximation to within 10^{-5} to a value in $[0.5, 1.5]$ with $e^x - 2 = \cos(e^x - 2)$.

10. Let $f(x) = (x+2)(x+1)^2x(x-1)^3(x-2)$. To which zero of f does the Bisection method converge when applied on the following intervals?

a. $[-1.5, 2.5]$ **b.** $[-0.5, 2.4]$ **c.** $[-0.5, 3]$ **d.** $[-3, -0.5]$

11. Let $f(x) = (x+2)(x+1)x(x-1)^3(x-2)$. To which zero of f does the Bisection method converge when applied on the following intervals?

a. $[-3, 2.5]$ **b.** $[-2.5, 3]$ **c.** $[-1.75, 1.5]$ **d.** $[-1.5, 1.75]$

12. Find an approximation to $\sqrt{3}$ correct to within 10^{-4} using the Bisection Algorithm. [*Hint:* Consider $f(x) = x^2 - 3$.]

13. Find an approximation to $\sqrt[3]{25}$ correct to within 10^{-4} using the Bisection Algorithm.

14. Use Theorem 2.1 to find a bound for the number of 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.

15. Use Theorem 2.1 to find a bound for the number of 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.

16. Let $f(x) = (x-1)^{10}$, $p = 1$, and $p_n = 1 + 1/n$. Show that $|f(p_n)| < 10^{-3}$ whenever $n > 1$ but that $|p - p_n| < 10^{-3}$ requires that $n > 1000$.

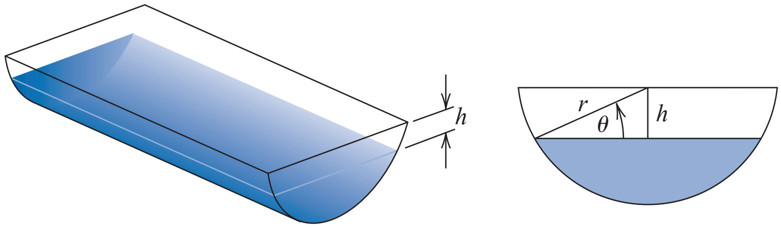
17. Let $\{p_n\}$ be the sequence defined by $p_n = \sum_{k=1}^{n} \frac{1}{k}$. Show that $\{p_n\}$ diverges even though $\lim_{n\to\infty}(p_n - p_{n-1}) = 0$.

18. 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

a. 0, if $a + b < 2$ **b.** 2, if $a + b > 2$ **c.** 1, if $a + b = 2$

19. 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

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



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.

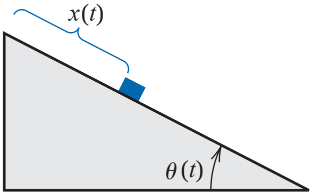
20. A particle starts at rest on a smooth inclined plane whose angle θ is changing at a constant rate

$$\frac{d\theta}{dt} = \omega < 0.$$

At the end of t seconds, the position of the object is given by

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

Suppose the particle has moved 1.7 ft in 1 s. Find, to within 10^{-5}, the rate ω at which θ changes. Assume that $g = 32.17$ ft/s^2.



2.2 Fixed-Point Iteration

A *fixed point* for a function is a number at which the value of the function does not change when the function is applied.

Definition 2.2 The number p is a **fixed point** for a given function g if $g(p) = p$. ■

In this section we consider the problem of finding solutions to fixed-point problems and the connection between the fixed-point problems and the root-finding problems we wish to solve. Root-finding problems and fixed-point problems are equivalent classes in the following sense:

Fixed-point results occur in many areas of mathematics, and are a major tool of economists for proving results concerning equilibria. Although the idea behind the technique is old, the terminology was first used by the Dutch mathematician L. E. J. Brouwer (1882–1962) in the early 1900s.

- Given a root-finding problem $f(p) = 0$, we can define functions g with a fixed point at p in a number of ways, for example, as

$$g(x) = x - f(x) \quad \text{or as} \quad g(x) = x + 3f(x).$$

- Conversely, if the function g has a fixed point at p, then the function defined by

$$f(x) = x - g(x)$$

has a zero at p.

Although the problems we wish to solve are in the root-finding form, the fixed-point form is easier to analyze, and certain fixed-point choices lead to very powerful root-finding techniques.

We first need to become comfortable with this new type of problem, and to decide when a function has a fixed point and how the fixed points can be approximated to within a specified accuracy.

Example 1 Determine any fixed points of the function $g(x) = x^2 - 2$.

Solution A fixed point p for g has the property that

$$p = g(p) = p^2 - 2 \quad \text{which implies that} \quad 0 = p^2 - p - 2 = (p+1)(p-2).$$

A fixed point for g occurs precisely when the graph of $y = g(x)$ intersects the graph of $y = x$, so g has two fixed points, one at $p = -1$ and the other at $p = 2$. These are shown in Figure 2.3. ■