

# Nonlinear Root Finding

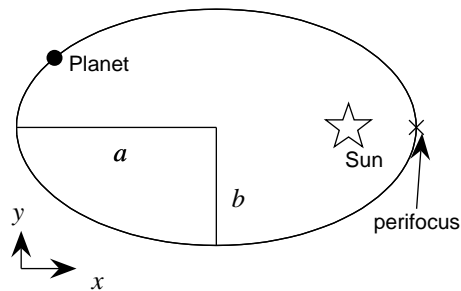
Grady Wright

November 24, 2004

## 1 Introduction

There has always been a close relationship between mathematics and the sciences (astronomy, chemistry, engineering, physics, etc.). When applying mathematics to study some physical phenomenon, it is often the case that we arrive at a mathematical model (an equation or set of equations) of the phenomenon that cannot be conveniently solved with exact formulas. The following classical example from astronomy illustrates this result:

**Example 1.1** Consider a planet in orbit around the sun:



Let  $\omega$  be the frequency of the orbit,  $t$  the time since the planet was last closest to the sun (this is called the perihelion in astronomy), and  $e = \sqrt{1 - \frac{b^2}{a^2}}$  the eccentricity of the planet's elliptical orbit. Then, according to Kepler's laws of planetary motion, the location of the planet at time  $t$  is given by

$$\begin{aligned}x &= a(\cos E - e) ; \\y &= a\sqrt{1 - e^2} \sin E ,\end{aligned}\tag{1}$$

where  $E$  is called the **eccentric anomaly** and is given by the equation

$$E = \omega t + e \sin E .\tag{2}$$

We cannot express the solution of this equation in terms of  $E$  with an exact formula. This is one of the most famous examples of a **nonlinear equation** in all of science.

While the mathematician may be interested in studying the form of the equations that arise from some mathematical model at some abstract level, this is unacceptable to the scientist. Her interest is typically in how the solution to the equations behaves as the input parameters of the problem change. For example how does the orbit change as parameters  $\omega$  and  $e$  are changed in the example above. To provide answers to these questions, it is often necessary to apply some convenient method for extracting numerical values from the equations. The development and analysis of such methods is called **numerical analysis**.

Numerical analysis has been around since the Babylonians (300 BC) began predicting the position of the sun and moon in the sky. However, it has only been considered a mathematical discipline since the advent of the digital computer. The basic mathematical operations of a computer are the same that we learn very early in our lives: addition, subtraction, multiplication, and division. The ability to do these operations extremely fast has caused the field of numerical analysis to blossom in the past half century. Today a modestly priced personal computer can do about  $10^9$  (1,000,000,000) of these basic operations in one second, while the fastest computers in the world (that we know about) can do about  $10^{12}$  (1,000,000,000,000) in one second! With this amazing speed, we can develop numerical methods for providing solutions to mathematical models for such things as blood clotting, airplane flight, the human genome, and global warming. While we will not be solving any of these problems, we will learn some basic ideas from numerical analysis by studying a common problem that arises in applications: **nonlinear root finding**.

The problem of nonlinear root finding can be stated in an abstract sense as follows:

*Given some function  $f(x)$ , determine the value(s) of  $x$  such that  $f(x) = 0$ .*

The nonlinear root finding problem that you are probably the most familiar with is determining the roots of a polynomial. For example, what are the roots of the polynomial  $f(x) = x^2 + 4x + 1$ ? This is equivalent to determining the values of  $x$  such that  $f(x) = 0$ . We all know that the solution to this problem can be obtained using the quadratic formula. This is a rare example of a nonlinear problem with a simple closed-form solution for its roots. If, for example, you were asked to compute the roots of the polynomial  $p(x) = x^5 + 4x^4 + x^3 - 5x^2 + 3x - 1$ , you would not be able to find a method involving a finite number of steps for determining a closed-form solution. Equation (2) from Example 1.1 can also be stated as a nonlinear root finding problem for  $E$  by rewriting it as follows:

$$f(E) = \omega t - E + e \sin(E) = 0 . \quad (3)$$

Similarly, a closed-form solution for this problem (for arbitrary  $e$ ,  $t$ , and  $\omega$ ) cannot be obtained in a finite number of steps. One issue that we always have to be concerned with for nonlinear root finding problems is the existence of a solution to the problem. For example, suppose we wish to solve the equation

$$\frac{3}{2} + \cos x = 0 .$$

Since  $-1 \leq \cos x \leq 1$ , for any real value of  $x$ , this equation has no real solution. (It has a complex-valued solution, but for the problems we are interested in, we will only be considering real-valued solutions.)

The central concept to all root finding methods is **iteration** or **successive approximation**. The idea is that we make some guess at the solution, and then we repeatedly improve upon that guess, using some well-defined operations, until we arrive at an approximate answer that is sufficiently close to actual answer. We refer to this process as an **iterative method**. We call the sequence of approximations the **iterates** and denote them by  $x_0, x_1, x_2, \dots, x_n, \dots$ . The following example illustrates the concept of an iterative method:

**Example 1.2** *Given some value  $a$ , make some initial guess  $x_0$  and compute  $x_0, x_1, x_2, \dots$ , using the iterative*

method

$$\begin{aligned}x_1 &= x_0(2 - ax_0) \\x_2 &= x_1(2 - ax_1) \\&\vdots \\x_{n+1} &= x_n(2 - ax_n) \\&\vdots\end{aligned}$$

Try computing a few iterations of this method for different values for  $a$  and  $x$ . Can you guess what this iterative method is converging to? (Hint: this was one of the most important iterative methods for early computers, and is still used today).

As you can see, iterative methods generally involve an infinite number of steps to obtain the exact solution. However, the beauty and power of these methods is that typically after a finite, relatively small number of steps the iteration can be terminated with the last iterate providing a very good approximation to the actual solution. One of the primary concerns of an iterative method is thus the rate at which it converges to the actual solution.

**Definition 1.3** Let  $x = \alpha$  be the actual solution to the nonlinear root finding problem  $f(x) = 0$ . A sequence of iterates is said to converge with order  $p \geq 1$  to  $\alpha$  if

$$|\alpha - x_n| \approx c|\alpha - x_{n-1}|^p \text{ as } n \rightarrow \infty, \quad (4)$$

where  $c > 0$ . If  $p = 1$  and  $c < 1$  then the convergence is called linear and we can rewrite (4) as

$$|\alpha - x_n| \approx c^n |\alpha - x_0| \text{ as } n \rightarrow \infty,$$

If  $p > 1$  then the convergence is called superlinear. The values  $p = 2$  and  $p = 3$  are given the special names quadratic and cubic convergence, respectively.

Making an analogy to cars, linear convergence is like a Honda Accord or Toyota Camery; quadratic convergence is like a Ferrari; and cubic convergence is like a rocket ship.

In addition to the order of convergence, the factors for deciding whether an iterative method for solving a nonlinear root finding problem is good are accuracy, stability, efficiency, and robustness. Each of these can be defined as follows:

- **Accuracy:** The error  $|\alpha - x_n|$  becomes small as  $n$  is increased.
- **Stability:** If the input parameters are changed by small amounts the output of the iterative method should not be wildly different, unless the underlying problem exhibits this type of behavior. For example, in the exercise above, we should be able to change  $x_0$  by a small amount and still have the iterates converge to the same number.
- **Efficiency:** The number of operations and the time required to obtain an approximate answer should be minimized.
- **Robustness:** The iterative method should be applicable to a broad range of inputs.

In the iterative methods that we study we will see how each one of these concepts applies.

## 2 Bisection Method

The bisection method is the easiest of all the iterative methods we discuss. The basic idea can be explained by the high/low game you may have played as a child. Someone chooses a number between 1 and 100 and the challenger tries to guess what it is. The challenger is told whether their guess was too high or too low. It should be rather obvious that to minimize the number of guesses, the challenger should select their successive guesses as the midpoint of the shrinking interval that brackets the solution.

In the high/low example, the function we are trying to find a root for is  $f(x) = x - \alpha$ , where  $\alpha$  is the number and  $x$  are the guesses. Since  $\alpha$  is an integer, the method above is guaranteed to converge to the correct answer in a finite number of steps. To make the problem a little more interesting, we allow an irrational value of  $\alpha$  to be chosen, but the challenger can only guess rational numbers (as is the case for a computer). Suppose the player selects  $\alpha = \sqrt{2}$ . To avoid having to know what the numerical value of  $\sqrt{2}$  is in order to determine whether a guess is too high or too low, we simply check if the square of the guess is above or below 2. Using this strategy we are actually finding the positive root of  $f(x) = x^2 - 2$ . Since  $1^2 < 2$  and  $2^2 > 2$ , the true solution is bracketed in the interval  $[1, 2]$  and we choose as our first guess the midpoint of this interval, i.e.  $x_0 = 1\frac{1}{2}$ . Since  $x_0^2 > 2$ , the next choice would be  $x_1 = 1\frac{1}{4}$ . Since  $x_1^2 < 2$ , the next choice would be  $x_2 = 1\frac{3}{8}$ . Continuing in this manner we would obtain the iterates

$$x_0 = 1\frac{1}{2}, x_1 = 1\frac{1}{4}, x_2 = 1\frac{3}{8}, x_3 = 1\frac{5}{16}, x_4 = 1\frac{13}{32}, x_5 = 1\frac{27}{64}.$$

Suppose we now want to approximate the solution to  $f(x) = 0$  for a general continuous function  $f(x)$ . As we have seen, the key to the bisection method is to keep the actual solution bracketed between the guesses. Thus, in addition to being given  $f(x)$ , we need an interval  $a \leq x \leq b$  where  $f(a)$  and  $f(b)$  differ in sign. We can write this requirement mathematically as  $f(a)f(b) < 0$ . It seems reasonable to conclude that since  $f(x)$  is continuous and has different signs at each end of the interval  $[a, b]$ , there must be at least one point  $\alpha \in [a, b]$ , such that  $f(\alpha) = 0$ . Thus,  $f(x)$  has at least one root in the interval. This result is in fact known as the **Intermediate Value Theorem**.

Now that we have an idea for how the bisection method works for a general problem  $f(x) = 0$ , it is time to write down a formal procedure for it using well defined operations. We call such a procedure an **algorithm**. We can think of an algorithm as a recipe for solving some mathematical problem. However, instead of the basic ingredients of flour, sugar, eggs, and salt, the fundamental building blocks of an algorithm are the basic mathematical operations of addition, subtraction, multiplication, and division, as well as the **for**, **if**, and **while** constructs.

The bisection algorithm is given by Algorithm 1. The line **while**  $b - a > 2\varepsilon$  in this algorithm is called the **stopping criterion**, and we call  $\varepsilon$  the error tolerance. This line says that we are going to continue bisecting the interval until the length of the interval is  $\leq 2\varepsilon$ . This guarantees that the value returned by the algorithm is at most a distance  $\varepsilon$  away from the actual solution. The value for  $\varepsilon$  is given as an input to the algorithm. Note that the smaller the value of  $\varepsilon$ , the longer it takes the bisection method to converge. Typically, we choose this value to be something small like  $\varepsilon = 10^{-6}$ . The stopping criterion that we have chosen is called an **absolute** criterion. Some other types of criterion are **relative** and **residual**. These correspond to  $b - a < 2\varepsilon|x_n|$  and  $|f(x_n)| \leq \varepsilon$ , respectively. There is no correct choice.

Let  $\alpha$  be the solution to  $f(x) = 0$  contained in the interval  $[a, b]$ . Since we are reducing the interval containing

---

**Algorithm 1** Bisection Method

---

Input: Continuous function  $f(x)$ .  
Interval  $[a, b]$ , such that  $f(a)f(b) < 0$ .  
Error tolerance  $\varepsilon$ .  
Output: Approximate solution that is within  $\varepsilon$  of a root of  $f(x)$ .

```
 $n = -1$  {initialize the counter}
while  $b - a > 2\varepsilon$  do
   $x_{n+1} = \frac{a+b}{2}$  {bisection the interval}
  if  $f(x_{n+1}) = 0$  then
    return  $x_{n+1}$  {we have found a solution, return it}
  end if
  if  $f(x_{n+1})f(a) < 0$  then
     $b = x_{n+1}$  {the guess is between  $a$  and  $x_{n+1}$ ; shrink the interval from the right}
  else
     $a = x_{n+1}$  {the guess is between  $x_{n+1}$  and  $b$ ; shrink the interval from the left}
  end if
   $n = n + 1$  {update the counter}
end while
 $x_{n+1} = \frac{b+a}{2}$  {bisection the interval one last time}
return  $x_{n+1}$  {return the solution}
```

---

$\alpha$  by  $\frac{1}{2}$  on each iteration, we have the **bound**

$$\begin{aligned} |\alpha - x_0| &\leq \frac{b-a}{2} \\ |\alpha - x_1| &\leq \frac{b-a}{2^2} \\ |\alpha - x_1| &\leq \frac{b-a}{2^3} \\ &\vdots \\ |\alpha - x_n| &\leq \frac{b-a}{2^n} \end{aligned} \tag{5}$$

(6)

Thus, the bisection method **converges linearly** to the solution at a rate of  $\frac{1}{2}$ . Note that this bound is entirely independent of the function  $f(x)$ . Equation (5) also provides a convenient bound on the total number of iterations necessary to ensure  $|\alpha - x_n| \leq \varepsilon$ . This will be satisfied when

$$\frac{1}{2^n}(b-a) \leq \varepsilon .$$

Solving this equation for  $n$  gives the bound

$$n \geq \log_2 \left( \frac{b-a}{\varepsilon} \right) . \tag{7}$$

The table below shows the numerical value and error of the first 10 iterates of the bisection algorithm for approximating  $\sqrt{2}$  using the function  $f(x) = x^2 - 2$ , the interval  $[1, 2]$ , and  $\varepsilon = 10^{-3}$ .

$n$	$x_n$	$b - a$	$ \sqrt{2} - x_n $
0	$1.5000000000 \cdot 10^0$	$1.0000000000 \cdot 10^0$	$8.5786437627 \cdot 10^{-2}$
1	$1.2500000000 \cdot 10^0$	$5.0000000000 \cdot 10^{-1}$	$1.6421356237 \cdot 10^{-1}$
2	$1.3750000000 \cdot 10^0$	$2.5000000000 \cdot 10^{-1}$	$3.9213562373 \cdot 10^{-2}$
3	$1.4375000000 \cdot 10^0$	$1.2500000000 \cdot 10^{-1}$	$2.3286437627 \cdot 10^{-2}$
4	$1.4062500000 \cdot 10^0$	$6.2500000000 \cdot 10^{-2}$	$7.9635623731 \cdot 10^{-3}$
5	$1.4218750000 \cdot 10^0$	$3.1250000000 \cdot 10^{-2}$	$7.6614376269 \cdot 10^{-3}$
6	$1.4140625000 \cdot 10^0$	$1.5625000000 \cdot 10^{-2}$	$1.5106237310 \cdot 10^{-4}$
7	$1.4179687500 \cdot 10^0$	$7.8125000000 \cdot 10^{-3}$	$3.7551876269 \cdot 10^{-3}$
8	$1.4160156250 \cdot 10^0$	$3.9062500000 \cdot 10^{-3}$	$1.8020626269 \cdot 10^{-3}$
9	$1.4150390625 \cdot 10^0$	$1.9531250000 \cdot 10^{-3}$	$8.2550012690 \cdot 10^{-4}$

The column  $b - a$  corresponds to the length of the increasingly shrinking interval around the root.

The most difficult part about using the bisection method is finding an interval  $[a, b]$  where the continuous function  $f(x)$  changes sign. Once this is found, the algorithm is guaranteed to converge. Thus, we would say that the bisection method is very robust. Also, as long as  $f(x)$  has only one root between the interval  $[a, b]$ , and it does not have another root very close to  $a$  or  $b$ , we can make small changes to  $a$  or  $b$  and the method will converge to the same solution. Thus, we would say the bisection method is stable. Additionally, (5) tells us that the error  $|\alpha - x_n|$  can be made as small as we like by increasing  $n$ . Thus, we would say the bisection method is accurate. Finally, the method converges linearly which is acceptable, but, as we will see in the next two sections, it is by no means the best we can do. Thus, we would say that the bisection method is not very efficient.

## Exercises

1. Estimate the minimum number of iterations necessary to guarantee the approximation to  $\sqrt{2}$  using the bisection method is within  $\varepsilon$  of the actual value for

- (a)  $\varepsilon = 10^{-3}$ ,  $[a, b] = [1, 2]$ ,
- (b)  $\varepsilon = 10^{-3}$ ,  $[a, b] = [0, 2]$ ,
- (c)  $\varepsilon = 10^{-3}$ ,  $[a, b] = [0, 4]$ ,
- (d)  $\varepsilon = 10^{-6}$ ,  $[a, b] = [1, 2]$ , and
- (e)  $\varepsilon = 10^{-12}$ ,  $[a, b] = [1, 2]$ .

From these results, how does the number of iterations increase as we (1) double the length of the interval containing the root? (2) double the exponent in the error tolerance  $\varepsilon$ ?

2. Use the bisection method with an error tolerance of  $\varepsilon = 10^{-3}$  to find the roots of the following equations. You may use the graph of the functions to determine a rough approximation for the interval.

- (a) The real root of  $x^3 - x^2 - x - 1 = 0$ .
- (b) The root of  $x = 1 + 0.3 \cos x$ .
- (c) The smallest positive root of  $e^{-x} = \sin x$ .
- (d) All real roots of  $x^4 - x - 1 = 0$ .

3. In this problem you will use the bisection method to compute the eccentric anomaly  $E$  in equation (2) for the earth to within an absolute error of  $10^{-3}$  and then use this value to compute the  $x$  and  $y$  coordinates of the earth. The approximate input parameters for the earth are  $e = 0.0167$ ,  $\omega = \frac{2\pi}{365.25635 \text{ days}}$ , and  $a = 149.6 \cdot 10^6$  km. Compute  $E, x$ , and  $y$  for the following values of  $t$ :

- (a)  $t = 91$  days,
- (b)  $t = 182$  days, and
- (b)  $t = 273$  days

To pick an appropriate interval for the bisection, you may consider plotting the function  $f(E)$  in (3) and then graphically determining where it crosses the  $E$ -axis.

4. Suppose you want to use your calculator to compute  $1/\alpha$  for some  $\alpha > 0$ , but that your calculator only has an add, subtract, and multiply button. Since the root of  $f(x) = 1 - \alpha x$  is  $1/\alpha$ , we can use these three operations to approximate  $1/\alpha$  by using the bisection method on this function.
  - (a) Give a method for calculating an interval  $[a, b]$  that contains  $1/\alpha$ , but does not involve any division.
  - (b) Suppose  $1 \leq \alpha \leq 2$ . Determine the number of iterations necessary to guarantee the absolute error in the approximation to  $1/\alpha$  is  $< 10^{-15}$ .
  - (c) With your strategy for choosing the interval from part (a), compute  $1/\alpha$  with the bisection method for
    - (i)  $\alpha = 7/13$ ,
    - (ii)  $\alpha = 11/81$ , and
    - (iii)  $\alpha = 13/797$ ,
 using an error tolerance of  $10^{-4}$ . Use these approximate results as the initial guess to the iterative method in Example 1.2 to compute  $1/\alpha$  to within an error of  $10^{-15}$ , thus giving a reliable and fast method for computing  $1/\alpha$ .
5. The way that the bisection method is presented in Algorithm 1, each iteration requires three evaluations of  $f(x)$ . If  $f(x)$  is some complicated function that requires a significant amount of time to compute then these function evaluations could significantly slow down the bisection method. How can we rewrite the algorithm so that only one function value is used at each iteration?

### 3 Secant method

To derive the secant method we use a geometrical approach. Suppose that we are trying to find a root of the function  $f(x)$  given in Figure 1 (the root is where  $f(x)$  crosses the  $x$ -axis) and suppose that we make two initial guesses  $x_0$  and  $x_1$  that are reasonably close to the true root of  $f(x)$ . To improve upon these guesses, we find the line  $L_1$  that passes through these points and choose as our next guess the point  $x_2$  where this line passes through  $x$ -axis. To improve upon this guess, we find the line  $L_2$  that passes through the points  $x_1$  and  $x_2$ . Our next guess becomes the point  $x_3$  where this line passes through  $x$ -axis. We repeat this process of finding where the line that passes through the previous two guesses  $x_n$  and  $x_{n-1}$  intersects the  $x$ -axis to produce the next guess  $x_{n+1}$ . Like the bisection method, this process is terminated when we reach some **stopping criterion**. We call this method the secant method because a line that passes through two points of a curve is called a secant.

To make the secant method useful, we need an algebraic formula that tells us how to generate the iterate  $x_{n+1}$  from the previous two iterates  $x_n$  and  $x_{n-1}$ . The line that passes through the previous two iterates is given in point-slope form as

$$y(x) - f(x_n) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} (x - x_n) .$$

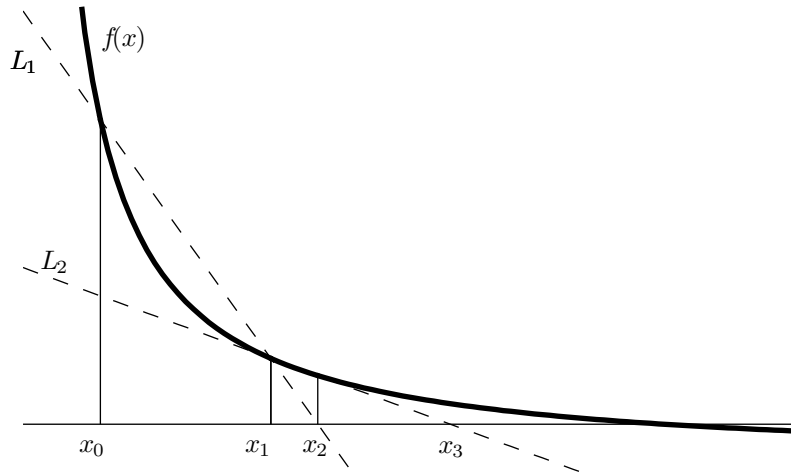


Figure 1: Geometrical illustration of the secant method.

The new iterate  $x_{n+1}$  is given by the location where this line passes through the  $x$ -axis. Setting  $y(x) = 0$  and solving for  $x = x_{n+1}$  in the above equation gives

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n), \quad n = 1, 2, \dots \quad (8)$$

Now that we know how to generate the sequence of iterates, we are ready to write down an algorithm for the secant method. This algorithm is given in Algorithm 2. The stopping criterion in this algorithm is

---

**Algorithm 2** Secant method

---

Input: Continuous function  $f(x)$ .  
Two initial guesses  $x_0$  and  $x_1$  for the root.  
Error tolerance  $\varepsilon$ .  
Output: Approximate solution  $x^*$ .

```

n = 1 {initialize the counter}
while  $|x_n - x_{n-1}| > \varepsilon$  do
   $x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n)$  {generate the next iterate}
  if  $f(x_{n+1}) = 0$  then
     $x^* = x_{n+1}$  {we have found a solution}
    return  $x^*$  {return it}
  end if
  n = n + 1 {update the counter}
end while
 $x^* = x_n$ 
return  $x^*$  {return the solution}

```

---

similar to that of the bisection algorithm. However, the bisection method is guaranteed to converge since the true solution always remains bracketed between a shrinking interval. The secant method has no such guarantee. In Exercise 3 we give an example where the method fails. Fortunately, we can typically detect this convergence failure by monitoring the difference between the successive iterates  $x_n$  and  $x_{n-1}$ . If the difference is not decreasing after several iterations, or even worse if it is increasing, then we know the method



is failing and we can attempt to restart it with some better initial guesses. Another approach is to use the bisection method until we are reasonably close to the solution and then switch to the secant method.

While we may have lost the guarantee on convergence of the secant method, we have gained an increase in the rate at which the iterates converge (when the secant method does converge.) To illustrate this increased rate, we again consider approximating  $\sqrt{2}$  by finding a root of the function  $f(x) = x^2 - 2$ . The following table lists the numerical results with the initial guess  $x_0 = 1$  and  $x_1 = 2$  and the tolerance  $\varepsilon = 10^{-6}$ :

$n$	$x_n$	$ x_n - x_{n-1} $	$ \sqrt{2} - x_n $
0	1.0000000000	-----	$4.1421356237 \cdot 10^{-1}$
1	2.0000000000	$1.0000000000 \cdot 10^0$	$5.8578643763 \cdot 10^{-1}$
2	1.3333333333	$6.6666666667 \cdot 10^{-1}$	$8.0880229040 \cdot 10^{-2}$
3	1.4000000000	$6.6666666667 \cdot 10^{-2}$	$1.4213562373 \cdot 10^{-2}$
4	1.4146341463	$1.4634146341 \cdot 10^{-2}$	$4.2058396837 \cdot 10^{-4}$
5	1.4142114385	$4.2270786659 \cdot 10^{-4}$	$2.1238982251 \cdot 10^{-6}$
6	1.4142135621	$2.1235824503 \cdot 10^{-6}$	$3.1577473969 \cdot 10^{-10}$
7	1.4142135624	$3.1577496173 \cdot 10^{-10}$	$2.2204460493 \cdot 10^{-16}$

With these two initial guesses, the secant method only requires six iterations to converge to the  $\sqrt{2}$  to the maximum precision that a standard computer or calculator possesses. Compare these results to those of the bisection method, which after 6 iterations produces an approximation with an error of  $7.66 \cdot 10^{-3}$ . In general, if  $\alpha$  is a solution to  $f(x) = 0$  and if the secant method converges to  $\alpha$  for the initial guesses  $x_0$  and  $x_1$  then it can be shown that

$$|\alpha - x_n| \approx c|\alpha - x_{n-1}|^{\frac{1+\sqrt{5}}{2}}, \text{ as } n \rightarrow \infty, \quad (9)$$

for some  $c > 0$ . The proof of this result is beyond the scope of this lesson; it involves some results from calculus and the theory of linear difference equations. You may recognize the power  $p = \frac{1+\sqrt{5}}{2} \approx 1.618$  as the **golden ratio**. Once we are close to the actual solution  $\alpha$  this convergence result tells us that the correct number of digits in  $x_n$  is roughly multiplied by 1.618 with each iteration. Unlike the bisection method, we cannot bound the maximum number of iterations it will take for the secant method to converge to within the given tolerance for a general function  $f(x)$ .

Like the bisection method, the most difficult part of the secant method is determining two good initial starting guesses for the root of the continuous function  $f(x)$ . If these are made sufficiently close to the actual answer then we are guaranteed the secant method will converge. However, since the method is not guaranteed to converge for any initial starting guesses, we would say that the secant method is only mildly robust. Also, provided the initial guesses are sufficiently close to a root  $\alpha$ , and there are not any other roots very close to  $\alpha$  then we can make small changes to the initial guesses and the method will still converge to the same root. Thus, we would say the secant method is stable. Provided that the secant method converges, (9) tells us that the error can be made as small as we like by increasing  $n$  and that the rate is **superlinear**. Thus, we would say that the secant method is accurate and fairly efficient.

## Exercises

1. Repeat exercise 2 from the previous section, but use an error tolerance of  $10^{-8}$ . You might consider using two successive approximations from the bisection method as initial guesses to the secant method.
2. Repeat Exercise 3 from the previous section, but use an error tolerance of  $10^{-8}$ .
3. Illustrate how the secant method would proceed for the three functions in the plots on page 12 by drawing the secant lines and successive iterates on the graph. Use the indicated points as the starting guesses. For which function will the secant method fail to converge to indicated root?

4. You are contracted by a power company to solve a problem related to a suspended power cable (at points of equal height) from two towers 100 yards apart. The power company requires that the cable “dip” down 10 yards at the midpoint and they have asked you to determine what the length of the cable should be. After doing some research, you find that the curve describing a suspended cable is known as a **catenary** and is given by the equation

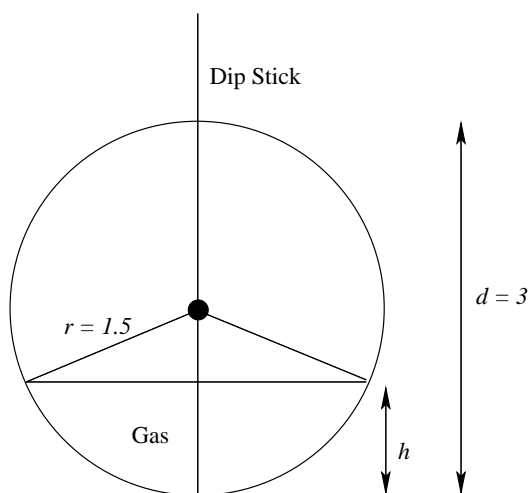
$$f(x) = \frac{c}{2} \left( e^{(x-a)/c} + e^{-(x-a)/c} \right),$$

where  $c$  is a parameter that determines how quickly the catenary “opens up”, and  $a$  is the location of the “dip”. You also find that the length of a catenary is given by the equation

$$s(x) = \frac{c}{2} \left( e^{(x-a)/c} - e^{-(x-a)/c} \right),$$

Given that the dip in the cable is 10 yards at the midpoint, use  $f(x)$  to determine the nonlinear equation for computing  $c$ . Solve this nonlinear equation using the secant method and then use the solution to approximate the required length of the power cable. Finally, send your rather large bill to the power company!

5. Suppose you and your family are on a house boat trip to Lake Powell. Before you leave the dock one of the mechanics tells you that the engines on your house boat will unfortunately stop working exactly when the gas tank is  $3/8$  full or less. After some arguing about getting another house boat, you decide to stick with this boat and to ensure that the tanks never get below  $3/8$  full. Once out on the lake, you realize the gas gauge is not working so you cannot tell how much gas is in the tank. You call the mechanic on your cell phone and she tells you to simply check the level of gas manually by using a dip stick. After locating the gas tank and a dipstick, you realize there is another problem: The gas tank is perfectly cylindrical (with a diameter of 3 feet), thus when you measure the amount of gas the length of the gas mark on the dipstick does not directly correspond to the amount of gas in the tank. For example, if the length of the mark is  $3 \cdot \frac{3}{8}$ , this does not mean the tank is  $3/8$  full. Remembering that you brought your calculator with you on your trip, determine the length of the gas mark on the dipstick that corresponds to the tank being  $3/8$  full. The following picture may help you get started:



6. Let  $h_n = x_{n-1} - x_n$  in the secant method. Then we can rewrite the secant iteration (8) as

$$x_{n+1} = x_n - \frac{f(x_n)}{\left( \frac{f(x_n+h) - f(x_n)}{h_n} \right)}, \quad n = 1, 2, \dots$$

Now consider what happens as we take the limit as  $h_n \rightarrow 0$ , (i.e. as  $x_n$  and  $x_{n-1}$  one another). For this case, we write the denominator in the above equation as

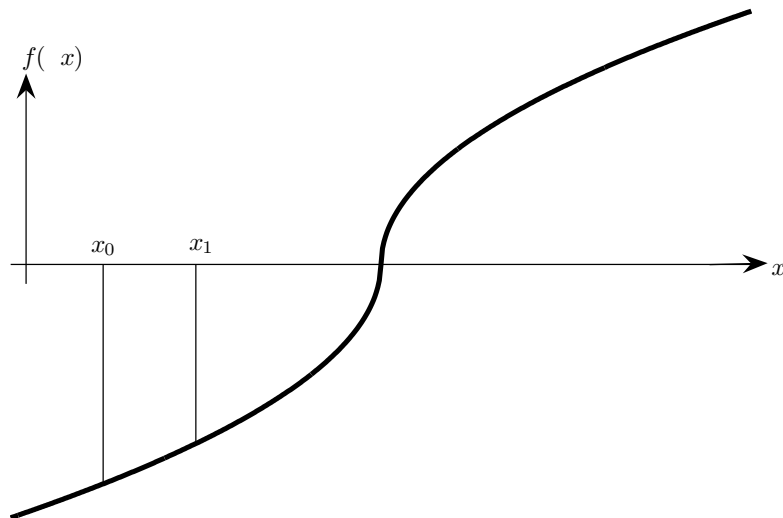
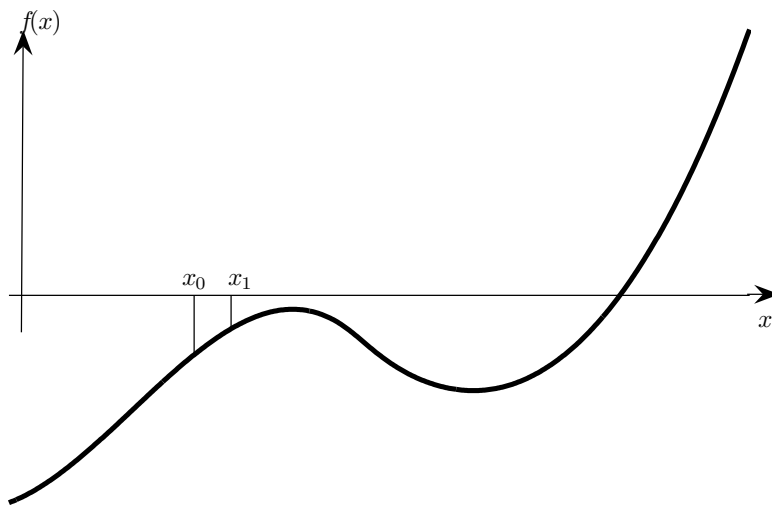
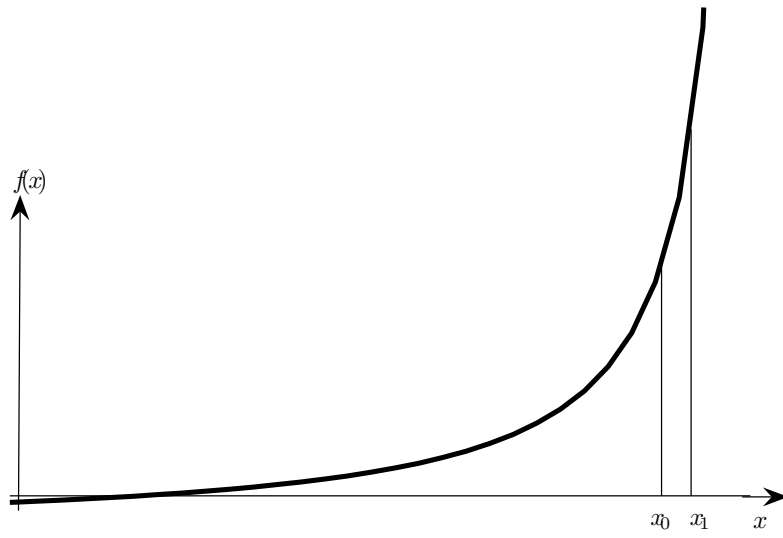
$$\lim_{h_n \rightarrow 0} \frac{f(x_n + h) - f(x_n)}{h_n}.$$

This quantity is called the derivative of  $f(x)$  at  $x_n$  and is given the name  $f'(x_n)$ . Thus, in the limit as  $h \rightarrow 0$ , the secant method is given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 1, 2, \dots \quad (10)$$

This is known as **Newton's method** and is the most widely used root finding method. Use this method for solving the problems from Exercise 1 of the previous section.

7. As the secant method is presented in Algorithm 2, it is not very efficient. The main problem is that it involves four evaluations of  $f(x)$  for each iteration which could be very time intensive. Rewrite the function so that  $f(x)$  is only evaluated once per iteration. Note that this will involve introducing two new variables to temporarily store the previous evaluations of  $f(x)$ .



## 4 Fixed point iteration

We turn our attention now towards a method for determining a fixed point of a function. This problem can be stated in an abstract sense as follows:

*Given some function  $g(x)$ , determine the value(s) of  $x$  such that  $x = g(x)$ .*

The fixed point problem is mathematically equivalent to the nonlinear root finding problem. To see this, let  $f(x) = c(g(x) - x)$  for  $c \neq 0$ , then finding the values where  $f(x) = 0$  is the same as finding a fixed point of  $g(x)$ . Conversely, let  $g(x) = cf(x) + x$ , then finding the values where  $x = g(x)$  is the same as finding a root of  $f(x)$ . It is often possible to also exploit the form of  $f(x)$  in the nonlinear root finding problem to make it a fixed point problem as the following example illustrates.

**Example 4.1** *Consider the nonlinear root finding problem*

$$f(x) = x^2 - a = 0 \quad (a > 0).$$

*The following are four examples for converting this to a fixed point problem:*

$$\begin{aligned} (a) \quad & x = x - a + x^2 \\ (b) \quad & x = \frac{a}{x} \\ (c) \quad & x = x + 1 - \frac{1}{a}x^2 \\ (d) \quad & x = \frac{1}{2} \left( x + \frac{a}{x} \right) \end{aligned}$$

*The solution to each of these problems is  $x = \pm\sqrt{a}$ .*

The simplest technique for solving the fixed point problem is to make some initial guess  $x_0$  to the solution and then refine that guess by repeatedly plugging it back into the function  $g(x)$ . This gives the following sequence of iterates

$$\begin{aligned} x_1 &= g(x_0) \\ x_2 &= g(x_1) \\ &\vdots \\ x_{n+1} &= g(x_n) \\ &\vdots \end{aligned}$$

We fittingly call this technique the **fixed point iteration method**. The following exercise illustrates how to apply this technique and how the results of the iteration depend on the formulation of the fixed point problem.

**Exercise 4.2** *Convert the fixed point problems (a)-(d) in Example 4.1 to fixed point iterations. For the initial guess  $x_0 = 3$ , write down the first 5 iterations for each of these problems with  $a = 2$ .*

From this exercise, we clearly see that the convergence of the fixed point iteration method depends on the form of fixed point problem.

Graphically, the solution to the fixed point problem  $x = g(x)$  is given by the points where the function  $y = g(x)$  intersects the line  $y = x$ . The fixed point iteration method can also be described graphically as follows:

1. Plot the line  $y = x$  and  $y = g(x)$ .
2. Make an initial guess  $x_0$ .
3. Draw a vertical line from the point  $(x_0, 0)$  to the point  $(x_0, g(x_0))$ .
4. Draw a horizontal line from the point  $(x_0, g(x_0))$  until it intersects the line  $y = x$ . The point of intersection is the next iterate  $x_1$ .
5. Repeat steps 3 and 4 with the new iterate.

Figure 2 illustrates this graphical procedure for finding the fixed point  $\alpha$  of a function  $g(x)$ . We can clearly

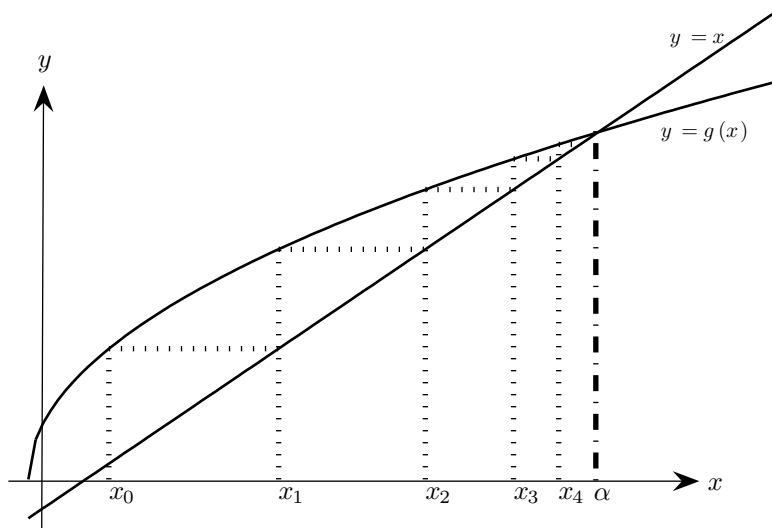


Figure 2: Graphical illustration of the fixed point iteration method.

see from the figure that the iterates are converging to  $\alpha$ .

## Exercises

1. Convert the following nonlinear root finding problems to fixed point problems  $x = g(x)$ .
  - (a)  $x^3 - x^2 - x - 1 = 0$ ;  $-\infty < x < \infty$
  - (b)  $x^3 - 2x - 5 = 0$ ;  $-\infty < x < \infty$
  - (c)  $x^3 - a = 0$ ;  $-\infty < x < \infty$  ( $a > 0$ )
  - (d)  $x - 91(2\pi/365.25635) - 0.0167 \sin x = 0$ ;  $-\infty < x < \infty$  (Kepler's problem (3) for the earth)
  - (e)  $x + \ln x = 0$ ;  $0 < x < \infty$
  - (f)  $e^x - \sin^2(x) - 2 = 0$ ;  $-\infty < x < \infty$
  - (g)  $x \cos x - \ln x^2 = 0$ ;  $0 < x < 2$
2. Each of the problems (a)–(g) in the previous exercise have one real solution in the interval specified. Write a fixed point iteration method for each problem and use it to find an approximate solution. Verify each answer by plugging it back into the original equation.

- Use the graphical fixed point iteration method on each of the plots on pages 15–16 to determine if the method converges to the fixed point  $\alpha$ . Use several different starting guesses for each plot.
- The saying “a picture says a thousand words” is also very applicable to mathematics. Using the results from the previous exercise and those from Figure 2, try to determine mathematical conditions necessary to guarantee the fixed point iteration method will converge to a fixed point  $\alpha$ .

