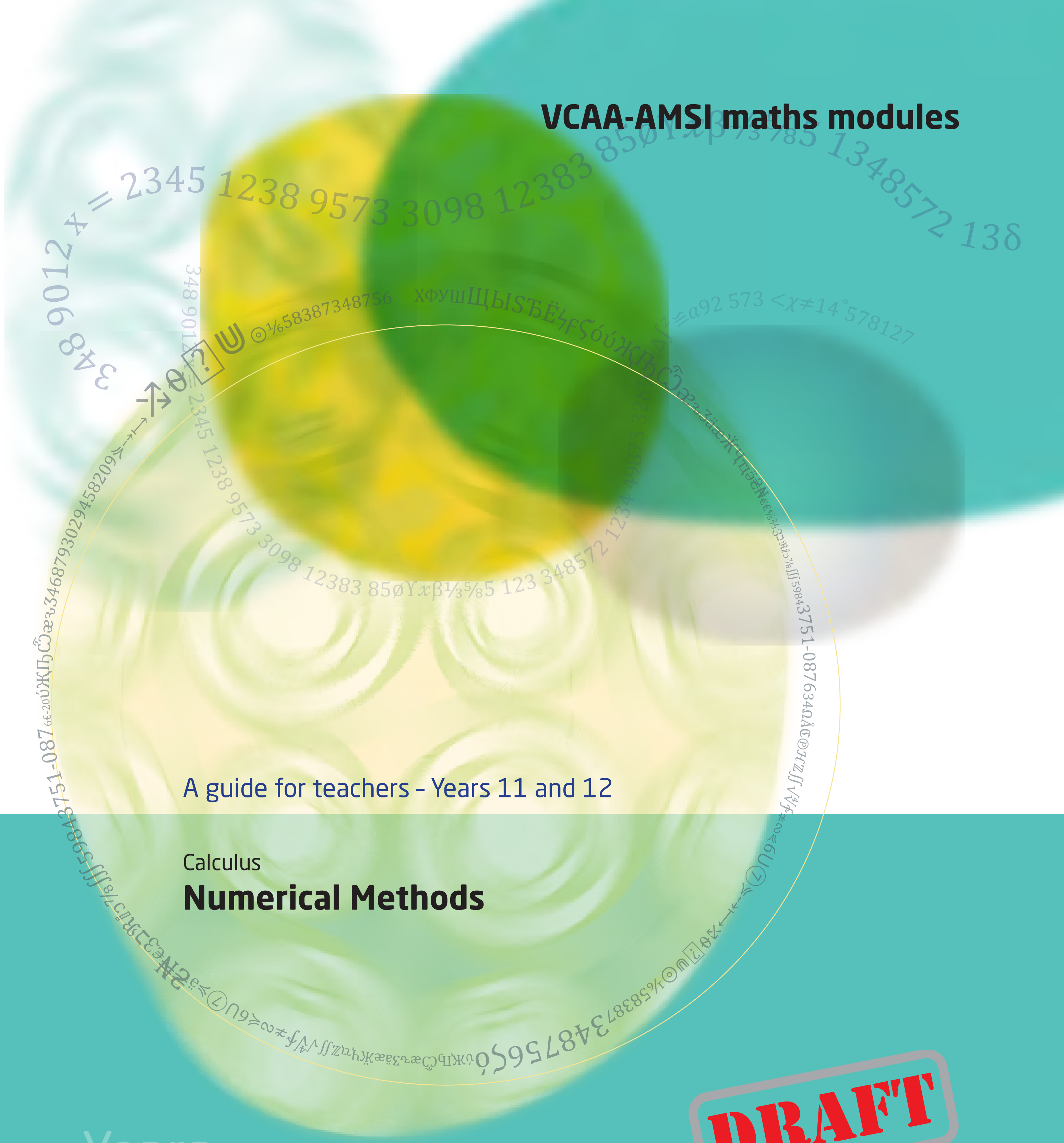
VCAA-AMSI maths modules



A guide for teachers – Years 11 and 12

Calculus

Numerical Methods

DRAFT

Years 11 & 12

AMSI
AUSTRALIAN MATHEMATICAL
SCIENCES INSTITUTE

DRAFT

Numerical Methods – A guide for teachers (Years 11–12)

Dr Daniel V. Mathews, Monash University

Illustrations and web design: Catherine Tan

Australian Mathematical Sciences Institute
Building 161
The University of Melbourne
VIC 3010
Email: enquiries@amsi.org.au
Website: www.amsi.org.au

Numerical Methods
AMSI Module

Assumed knowledge

- Familiarity with functions and their graphs.
- Familiarity with solving equations.
- Familiarity with differentiation, and the geometric meaning of the derivative as the gradient of tangent to a curve.

Motivation

> Although this may seem a paradox, all exact science is dominated by the idea of approximation. When a man tells you that he knows the exact truth about anything, you are safe in inferring that he is an inexact man.
>
> – Bertrand Russell

Sometimes we can find nice, or simple, or exact solutions to equations. Linear equations and quadratic equations, for example, can be solved exactly. But other types of equations can be much more difficult to solve exactly. In fact, sometimes it can be impossible to write down an exact expression for a solution.

Exercises in school mathematics textbooks are often deliberately designed to give nice answers. But in solving many mathematical equations deriving from real-world situations, there is no reason to expect the answer to be particularly nice. Often, the best that we can hope for is an approximate solution, to a desired degree of accuracy.

When equations are difficult to solve, we can resort to approximate numerical methods to find a solution. It is sometimes more efficient to find an approximate answer.

In this module we will examine two of the most common and useful numerical methods for finding approximate solutions to equations: the *bisection method*, and *Newton's method*. These methods are quite interesting in their own right, and lead to some beautiful pictures.

Content

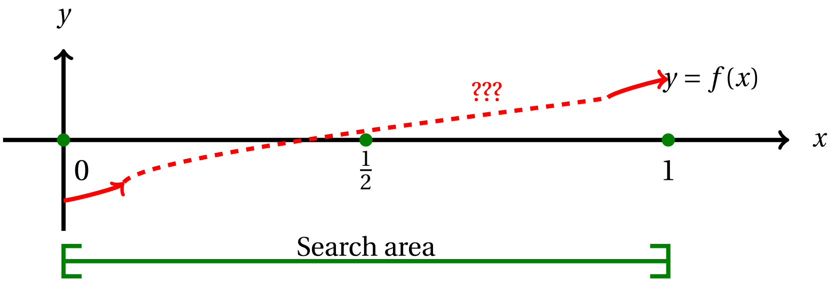
The bisection method

> Be approximately right rather than exactly wrong.
>
> – John W. Tukey

To “bisect" something means to cut it in half. The bisection method searches for a solution by bisecting: narrowing down the search area by half at each step.

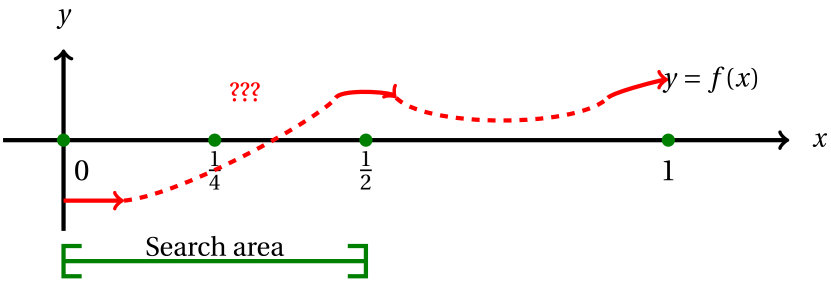
The idea is as follows. Suppose you want to solve an equation $f(x) = 0$, and you know there’s a solution somewhere between 0 and 1. For instance, suppose you know that $f(0)$ is negative, while $f(1)$ is positive. Then there must be a solution to the equation somewhere between 0 and 1! Your search area is the interval $[0,1]$.

We’ve drawn the situation below. The graph of $y = f(x)$ is drawn in red, but the dotted part is unknown; all we know is that the graph is a curve connecting the parts drawn in solid red. The search area is marked in green.

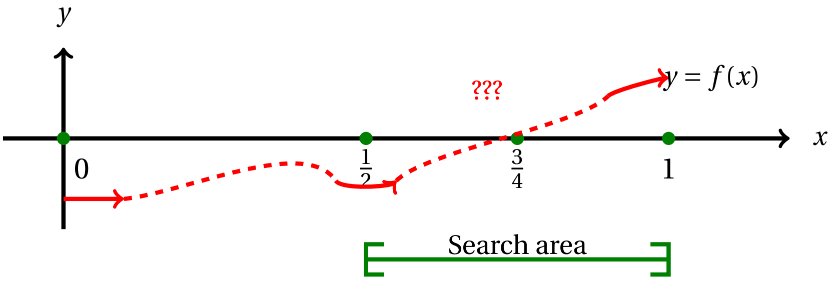


You examine the centre of the search area, evaluating $f(\frac{1}{2})$. If $f(\frac{1}{2}) = 0$, you have a solution, and you’re done! Otherwise, $f(\frac{1}{2}) \neq 0$. In this case $f(\frac{1}{2})$ could be positive or negative.

First, suppose $f(\frac{1}{2})$ is positive. Then $f(x)$ must change sign between $f(0)$, which is negative, and $f(\frac{1}{2})$, which is positive. So there must be a solution between 0 and $\frac{1}{2}$. (There could be other solutions as well, but we only need one!) You’ve narrowed down your search area to $[0, \frac{1}{2}]$, as shown.



Alternatively, $f(\frac{1}{2})$ could be negative. In this case, $f(x)$ changes sign between $f(\frac{1}{2})$ (which is negative) and $f(1)$ (which is positive). So there must be a solution between $\frac{1}{2}$ and 1. There could be other solutions as well, but there's certainly one between $\frac{1}{2}$ and 1. You've narrowed down your search area to $[\frac{1}{2}, 1]$, as shown.

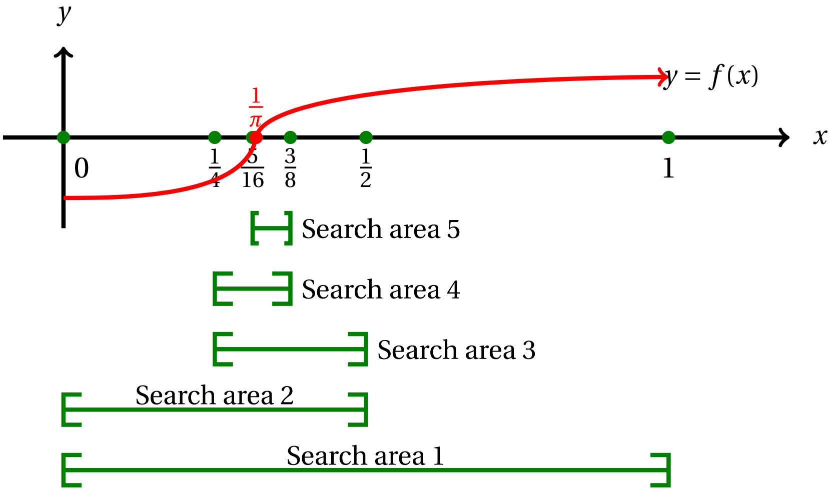


Either way, your search area has been narrowed down from $[0, 1]$, an interval of length 1, to a smaller interval, of length $\frac{1}{2}$.

You now continue searching. You take x to be in the middle of your new search area. If your search area is now $[0, \frac{1}{2}]$, you try $x = \frac{1}{4}$. If your search area is now $[\frac{1}{2}, 1]$, you try $x = \frac{3}{4}$. If you find $f = 0$ at this point, you have a solution! If you don't, you narrow down your search area by half again.

Proceeding in this way, you'll either find a solution, or get very close to one. How close? As close as you like. However closely you want to approximate a solution, you'll be able to do it with the bisection algorithm.

For instance, suppose that your equation $f(x) = 0$ had a solution precisely at $x = \frac{1}{\pi} \sim 0.3183$. (You don't know that of course; you're trying to find the solution!) You'd first narrow down the interval from $[0, 1]$ to $[0, \frac{1}{2}]$; then to $[\frac{1}{4}, \frac{1}{2}] = [0.25, 0.5]$; then to $[\frac{1}{4}, \frac{3}{8}] = [0.25, 0.375]$; then to $[\frac{5}{16}, \frac{3}{8}] = [0.3125, 0.375]$; and so on. See the figure below.

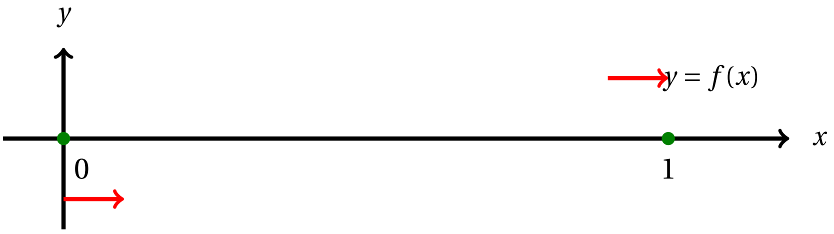


Having described the idea of the bisection method, we'll next discuss the theory behind it more rigorously.

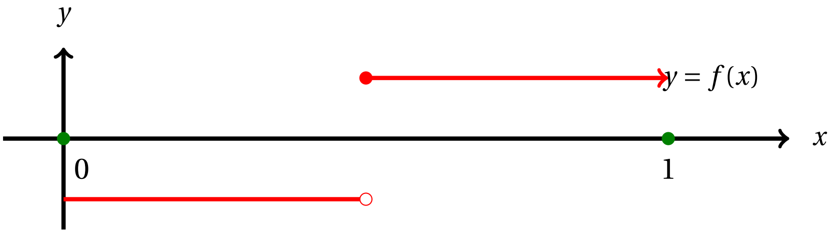
Intermediate value theorem

The bisection method relies upon an important theorem: the *intermediate value theorem*. This theorem is a very intuitive one. If you're on one side of a river, and later you're on the other side of the river, then you must have crossed the river!

Consider a function $f : [0,1] \to \mathbb{R}$ and its graph $y = f(x)$. If $f(0) < 0$, then the graph lies below the x-axis at $x = 0$. If $f(1) > 0$, then the graph lies above the x-axis at $x = 1$. (Portions of the graph must appear as shown below.) So between $x = 0$ and $x = 1$, the graph must cross the x-axis.



But beware! If f is *discontinuous*, the graph $y = f(x)$ could jump over the x-axis!



Exercise 1

Find an example, with an explicit formula, of a function $f : [0,1] \to \mathbb{R}$ such that $f(0) < 0$, $f(1) > 0$, and for all $x \in [0,1]$, $f(x) \neq 0$.

Nonetheless, provided we stick with *continuous* functions, the graph must "cross the river" of the x-axis. That is, there must be an $x \in [0,1]$ such that $f(x) = 0$.

Now, although we described the left endpoint being below the river (i.e. $f(0) < 0$) and the right endpoint being above the river (i.e. $f(1) > 0$), it could be the other way around, and the same conclusion would hold. If $f(0) > 0$ and $f(1) < 0$, there still must be a solution $x \in [0,1]$ to $f(x) = 0$.

There's nothing special about the interval $[0,1]$ either. We could replace $[0,1]$ with any interval $[a,b]$. Provided f is positive at one end and negative at the other end, then $f(x) = 0$ must have a solution in $[a,b]$.

Finally, there's also nothing special about the value 0. For any real number value c, if $f < c$ at one end of the interval $[a,b]$, and $f > c$ at the other end, then there must exist an $x \in [a,b]$ such that $f(x) = c$.

We summarise this discussion with a statement of the intermediate value theorem.

Theorem (Intermediate Value Theorem)

Let $f : [a,b] \to \mathbb{R}$ be a continuous function, and c be a real number.

a *If $f(a) < c$ and $f(b) > c$, then there exists an $x \in [a,b]$ such that $f(x) = c$.*

b *If $f(a) > c$ and $f(b) < c$, then there exists an $x \in [a,b]$ such that $f(x) = c$.*

Note that the intermediate value theorem doesn't say anything about *how many* times $f(x)$ takes the value c. There might be *many* values of x in the interval $[a,b]$ such that $f(x) = c$. All the theorem says is that there is at least one.

If you're on one side of the river, and later you're on the other side of the river, then you must have crossed the river — you might have crossed it many times, but you certainly crossed it at least once!

Bisection algorithm

We now describe the bisection algorithm in detail. Suppose we have a continuous function $f : [a,b] \to \mathbb{R}$, and we want to solve $f(x) = 0$. (To solve for some other value of f, i.e. $f(x) = c$, you can rearrange the equation to $f(x) - c = 0$.)

We assume $f(a)$ and $f(b)$ are both nonzero — otherwise we already have a solution! We also assume $f(a)$ and $f(b)$ have *opposite signs*. The intermediate value theorem then ensures that $f(x) = 0$ has a solution for some $x \in [a,b]$.

As we have seen, the idea is to look at the value of $f(x)$ at the *midpoint* $\frac{a+b}{2}$ of the interval $[a,b]$, i.e. the average (or mean) of a and b.

If $f(\frac{a+b}{2}) = 0$, then we have a solution, and we can stop. Otherwise, $f(\frac{a+b}{2})$ is either positive or negative. Depending on the signs of $f(a)$, $f(\frac{a+b}{2})$ and $f(b)$, the intermediate value theorem will tell us that one or the other of the intervals

$$\left[a, \frac{a+b}{2}\right] \quad \text{or} \quad \left[\frac{a+b}{2}, b\right]$$

contains a solution. We call this interval $[a_1, b_1]$.

We then repeat the process, evaluating f at the midpoint $\frac{a_1+b_1}{2}$ of $[a_1, b_1]$. If $f\left(\frac{a_1+b_1}{2}\right) = 0$, we have a solution, and stop. Otherwise, $f\left(\frac{a_1+b_1}{2}\right) \neq 0$ and the intermediate value theorem tells us that one or the other of the two sub-intervals obtained by bisecting $[a_1, b_1]$ contains a solution. We call this interval $[a_2, b_2]$. We continue in this fashion, obtaining intervals $[a_3, b_3]$, $[a_4, b_4]$, and so on.

We now state the bisection method rigorously.

[Bisection method] Let $f : [a,b] \to \mathbb{R}$ be a continuous function with $f(a)$ and $f(b)$ both nonzero, and of opposite sign. We seek a solution of $f(x) = 0$.

Let $[a_0, b_0] = [a, b]$. At the n'th step of the method, we have an interval $[a_{n-1}, b_{n-1}]$, where $f(a_{n-1})$ and $f(b_{n-1})$ are both nonzero and of opposite signs. Then:

1. Calculate $f\left(\frac{a_{n-1}+b_{n-1}}{2}\right)$. If $f\left(\frac{a_{n-1}+b_{n-1}}{2}\right) = 0$, $x = \frac{a_{n-1}+b_{n-1}}{2}$ is a solution, and we stop.
2. Otherwise, $f\left(\frac{a_{n-1}+b_{n-1}}{2}\right) \neq 0$. Precisely one of the following two possibilities occurs:
   - $f\left(\frac{a_{n-1}+b_{n-1}}{2}\right)$ and $f(a_{n-1})$ have opposite signs. Then let $[a_n, b_n] = \left[a_{n-1}, \frac{a_{n-1}+b_{n-1}}{2}\right]$.
   - $f\left(\frac{a_{n-1}+b_{n-1}}{2}\right)$ and $f(b_{n-1})$ have opposite signs. Then let $[a_n, b_n] = \left[\frac{a_{n-1}+b_{n-1}}{2}, b_{n-1}\right]$.
3. If the interval $[a_n, b_n]$ has the desired degree of accuracy for a solution, stop. Otherwise, return to step 1 and continue.

Note that the intervals $[a_n, b_n]$ produced by the bisection method are all *nested*:

$$[a_0, b_0] \supset [a_1, b_1] \supset [a_2, b_2] \supset \cdots.$$

Moreover, each of these intervals is half as long as the previous one: $[a_n, b_n]$ is half the length of $[a_{n-1}, b_{n-1}]$.

Exercise 2

How many times smaller is the length of the interval $[a_n, b_n]$ compared to the original $[a, b] = [a_0, b_0]$? In other words, what is

$$\frac{b_n - a_n}{b - a}?$$

Using the bisection algorithm

Let's now turn to some examples. We'll obtain some approximations to some well-known irrational numbers using the bisection method.

Example

Find π to within an accuracy of 0.05 by approximating the solution to $\sin x = 0$ in the interval $[3, 4]$ using the bisection method.

Solution

Let $f(x) = \sin x$, so we apply the bisection method to f, starting from $[a_0, b_0] = [3,4]$. As f is continuous, $f(3) = \sin 3 \sim 0.141$ is positive, and $f(4) = \sin 4 \sim -0.757$ is negative, so there is a solution in $[a_0, b_0]$.

We know from trigonometry that $\sin\pi = 0$, so $x = \pi$ is a solution. In fact, π is the *only* solution for $x \in [3,4]$. So by approximately solving $\sin x = 0$ in $[3,4]$, we approximate π.

At the first step, we consider the midpoint $\frac{7}{2} = 3.5$ of $[a_0, b_0] = [3,4]$ and compute $f\left(\frac{7}{2}\right) = \sin 3.5 \sim -0.351$, which is negative. So the sign of $f(3)$, which is positive, is opposite to the sign of $f\left(\frac{7}{2}\right)$, which is negative. Hence by the intermediate value theorem, there must be a solution for $x \in [3, \frac{7}{2}]$, and we set $[a_1, b_1] = \left[3, \frac{7}{2}\right] = [3, 3.5]$.

At the second step, consider the midpoint $\frac{13}{4} = 3.25$ of $[a_1, b_1] = [3, 3.5]$. We compute $f\left(\frac{13}{4}\right) = \sin 3.25 \sim -0.108$, which has opposite sign to $f(3)$, which is positive, and so we set $[a_2, b_2] = [3, \frac{13}{4}] = [3, 3.25]$.

At step 3, consider the midpoint $\frac{25}{8} = 3.125$ of $[a_2, b_2] = [3, 3.25]$ and compute $f\left(\frac{25}{8}\right) = \sin 3.125 \sim 0.017$, which has opposite sign to $f(3.25)$. So we take $[a_3, b_3] = \left[\frac{25}{8}, \frac{13}{4}\right] = [3.125, 3.25]$.

At step 4, consider $\frac{51}{16} = 3.1875$. We compute $f(3.1875) = \sin 3.1875 \sim -0.046$, which has opposite sign to $f(3.125)$, so $[a_4, b_4] = \left[\frac{25}{8}, \frac{51}{16}\right] = [3.125, 3.1875]$.

At step 5, we compute $f\left(\frac{101}{32}\right) = f(3.15625) = \sin 3.15625 \sim -0.015$, which has opposite sign to $f(\frac{25}{8}) = f(3.125)$, so $[a_5, b_5] = \left[\frac{25}{8}, \frac{101}{32}\right] = [3.125, 3.15625]$.

We now know that there is a solution to $\sin x = 0$ in the interval $\left[\frac{25}{8}, \frac{101}{32}\right] = [3.125, 3.15625]$, which has length $\frac{101}{32} - \frac{25}{8} = \frac{1}{32} = 0.03125$. So we know the solution to within 0.03125, which is better than the desired accuracy of 0.05.

As you can see, the bisection method can be a rather repetitive and time-consuming process! Because it is an algorithm requiring repeated evaluation of a function, it is well suited to implementation on a computer.

We can save ourselves some effort by tabulating the computations. At each stage, we can note down a_n and b_n and the signs of $f(a_n)$ and $f(b_n)$, then the midpoint $\frac{a_n+b_n}{2}$ and the sign of $f(\frac{a_n+b_n}{2})$. From this data we can calculate the next interval $[a_{n+1}, b_{n+1}]$ straightforwardly. This is quicker than writing out each step in English! The previous example could be tabulated as shown below.

Step n	a_n	Sign of $f(a_n)$	b_n	Sign of $f(b_n)$	Midpoint $\frac{a_n+b_n}{2}$	Sign of $f\left(\frac{a_n+b_n}{2}\right)$
0	3	+	4	−	$\frac{7}{2}=3.5$	−
1	3	+	$\frac{7}{2}=3.5$	−	$\frac{13}{4}=3.25$	−
2	3	+	$\frac{13}{4}=3.25$	−	$\frac{25}{8}=3.125$	+
3	$\frac{25}{8}=3.125$	+	$\frac{13}{4}=3.25$	−	$\frac{51}{16}=3.1875$	−
4	$\frac{25}{8}=3.125$	+	$\frac{51}{16}=3.1875$	−	$\frac{101}{32}=3.15625$	−
5	$\frac{25}{8}=3.125$	+	$\frac{101}{32}=3.15625$	−		

You might notice that in the column “Sign of $f(a_n)$", all the entries are the same, i.e. +; and in the column “Sign of $f(b_n)$", all the entries are also the same, i.e. −. In fact, the signs in these columns will never change.

Exercise 3

Prove that the signs of $f(a_0), f(a_1), f(a_2), \dots, f(a_n)$ are all the same. Similarly, show that the signs of $f(b_0), f(b_1), f(b_2), \dots, f(b_n)$ are all the same.

You might notice in our discussion that we continually write numbers both as fractions and decimals. Decimals are useful because they quickly convey how big each number is relative to the others. Fractions are useful because they express a number exactly and compactly.[1] Of course you do not have to do the same, but it is useful to be aware of the advantages of writing numbers both ways.

Example

Find $\sqrt{2}$ to within an accuracy of 0.01, by approximating the solution to $x^2 = 2$ in the interval $[1,2]$ using the bisection method.

Solution

Let $f(x) = x^2 - 2$, so we solve $f(x) = 0$ for $x \in [1,2]$ using the bisection method. The solutions to $x^2 - 2 = 0$ are $x = \pm\sqrt{2}$. Let $[a_0, b_0] = [1,2]$. As f is continuous, $f(1) = -1 < 0$ and $f(2) = 2 > 0$, there is a solution in this interval, i.e. $\sqrt{2}$. So we are approximating $\sqrt{2}$.

We tabulate the computations of the bisection method as follows.

[1] Also, the author is a pure mathematician and cannot bear to see only decimals!

Step		Sign		Sign	Midpoint	Sign
n	a_n	$f(a_n)$	b_n	$f(b_n)$	$\frac{a_n+b_n}{2}$	$f\left(\frac{a_n+b_n}{2}\right)$
0	1	$-$	2	$+$	$\frac{3}{2} = 1.5$	$+$
1	1	$-$	$\frac{3}{2} = 1.5$	$+$	$\frac{5}{4} = 1.25$	$-$
2	$\frac{5}{4} = 1.25$	$-$	$\frac{3}{2} = 1.5$	$+$	$\frac{11}{8} = 1.375$	$-$
3	$\frac{11}{8} = 1.375$	$-$	$\frac{3}{2} = 1.5$	$+$	$\frac{23}{16} = 1.4375$	$+$
4	$\frac{11}{8} = 1.375$	$-$	$\frac{23}{16} = 1.4375$	$+$	$\frac{45}{32} = 1.40625$	$-$
5	$\frac{45}{32} = 1.40625$	$-$	$\frac{23}{16} = 1.4375$	$+$	$\frac{91}{64} = 1.421875$	$+$
6	$\frac{45}{32} = 1.40625$	$-$	$\frac{91}{64} = 1.421875$	$+$	$\frac{181}{128} = 1.4140625$	$-$
7	$\frac{181}{128} = 1.4140625$	$-$	$\frac{91}{64} = 1.421875$	$+$		

After 7 iterations, we know there is a solution to $x^2 - 2 = 0$ in the interval $[\frac{181}{128}, \frac{91}{64}]$, which has length $\frac{91}{64} - \frac{181}{128} = \frac{1}{128} = 0.0078125 < 0.01$. So we have determined the solution $\sqrt{2}$ to within an accuracy of 0.01: it is between 1.4140625 and 1.421875.

These computations may appear quite exhausting! It took 7 iterations to obtain $\sqrt{2}$ to within 0.01, and we still do not know the second decimal digit of $\sqrt{2}$!

Exercise 4

Find e to within an accuracy of 0.02 by solving $\ln x = 1$ for x in the interval $[2,3]$.

Exercise 5

How many solutions x are there to the equation $\cos x = x$? Find all of them to within an accuracy of 0.01, using the bisection method.

Exercise 6

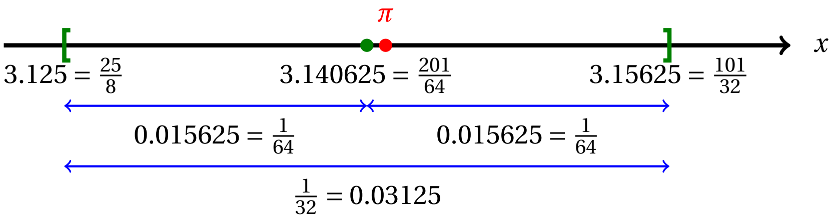
Approximate $\frac{1}{3}$ to within an accuracy of 0.001 by solving $3x - 1 = 0$ for $x \in [0,1]$ using the bisection method. What do you note about the signs of $f\left(\frac{a_n+b_n}{2}\right)$?

(This last exercise might seem silly: you know how to express 1/3 as a decimal! But an interesting pattern appears, which we discuss in the *Links Forward* section.)

How accurate is accurate?

In our first example, we found π to an accuracy of $\frac{1}{32} = 0.03125$, after 5 iterations of the bisection method: π is somewhere between 3.125 and 3.15625.

However, based on that example, what would be our best guess for π? We could guess the *midpoint* of the interval $[3.125, 3.15625]$, which is 3.140625. How close is our best guess? The situation is illustrated below.


π
x
$3.125 = \frac{25}{8}$
$3.140625 = \frac{201}{64}$
$3.15625 = \frac{101}{32}$
$0.015625 = \frac{1}{64}$
$0.015625 = \frac{1}{64}$
$\frac{1}{32} = 0.03125$

(We've illustrated the actual value of π in red, but we're not supposed to know that!)

Well, if our best guess is too high, then as the solution can't be less than $\frac{25}{8} = 3.125$, the guess can't be off by more than $\frac{201}{64} - \frac{25}{8} = \frac{1}{64} = 0.015625$.

And if our best guess is too low, then as the solution can't be more than $\frac{101}{32} = 3.15625$, the guess can't be off by more than $\frac{101}{32} - \frac{201}{64} = \frac{1}{64} = 0.015625$.

Either way, we know that our best guess of $\frac{201}{64} = 3.140625$ is within $\frac{1}{64} = 0.015625$ of a solution. In this sense, the bisection method actually gave us an answer that was accurate to within $\frac{1}{64} = 0.015625$, *twice as good* as we originally suggested.

Exercise 7

Previously we approximated $\sqrt{2}$ by performing seven iterations of the bisection method. What is our best guess for $\sqrt{2}$? How close is it to the solution?

While we have been measuring accuracy of our solution by the length of $[a_n, b_n]$, often an approximation is required to be accurate to a number of *decimal places*.

To check that you have a solution to a desired number k of decimal places, you should check that a_n and b_n agree to k decimal places.

For example, our approximation of π above does not even determine π to one decimal place! Based on the interval $[3.125, 3.15625]$, to one decimal place, π could be 3.1 or 3.2. To obtain π to one decimal place, in fact, takes 7 iterations of the bisection method; but to obtain the second decimal place takes only one further iteration.

Exercise 8

Continue the example above computing an approximation to π, up to the eighth iteration of the bisection method, and verify that you can approximate π to two decimal places.

For another example, when we approximated $\sqrt{2}$ above, it took 4 iterations to find $\sqrt{2}$ to 1 decimal place. It takes a further 7 iterations to find $\sqrt{2}$ to 2 decimal places!

Exercise 9

Continue the example above approximating to $\sqrt{2}$, up to $[a_{11}, b_{11}]$. Verify that it takes 11 iterations to approximate $\sqrt{2}$ to two decimal places.

Although the bisection method doubles its accuracy at each stage, its accuracy in terms of decimal places is much more haphazard!

We'll see later, in the *Links Forward* section, that there is a nice way to predict, based on the desired accuracy of our solution, just how many iterations of the bisection algorithm we need to calculate. On the other hand, we'll also see that, if we want a desired number of decimal places of accuracy, then it's harder to predict how many iterations of the bisection algorithm we will need.

Convergence of the bisection method

The best thing about the bisection method is that it is *guaranteed to work*. Provided you're using the method appropriately, you have a failsafe guarantee that the method works. That guarantee is the best possible type of guarantee: a mathematical theorem.

That is, if you're trying to solve $f(x) = 0$ in $[a, b]$, for a continuous function f, where $f(a)$ and $f(b)$ have opposite signs, then the bisection method is guaranteed to give you an *arbitrarily good* approximation to a solution.

"Arbitrarily good" means "as good as you want". You can say how good you want your approximation to be, and then, by applying the bisection method enough times, you can get the approximation you want.

More formally, suppose you say that you want your approximation to be accurate to within a number ϵ (the Greek letter epsilon). That is, you want to arrive at an interval $[a_n, b_n]$ where $b_n - a_n < \epsilon$. Then, by applying the bisection method enough times, you can obtain such an interval $[a_n, b_n]$ with $b_n - a_n < \epsilon$.

We can summarise the above formally as a theorem.

Theorem

Let $f : [a, b] \to \mathbb{R}$ be a continuous function and suppose $f(a)$, $f(b)$ are both nonzero and have opposite signs. Let ϵ be a positive number. Then the bisection method, after a finite number N of iterations, will produce an interval $[a_N, b_N]$ such that $b_N - a_N < \epsilon$.

That is, after some finite number N of iterations, we have a solution to $f(x) = 0$ to within *an accuracy of ϵ.*

Why is theorem true? The idea is simply that the lengths of the intervals $[a_n, b_n]$ halve at each stage. When you take a number and continually halve it, the number gets very small: as small as you like, smaller than any ϵ you prefer!

We'll see more details in the *Links Forward* section. There we will also see that there is a formula for N, the number of iterations of the bisection method required to obtain the desired accuracy.

The *best* thing about the bisection method may be that it is guaranteed to work, but the *worst* thing about the bisection method is that it is *slow*. We'll now see a method that is often much quicker in finding a solution — but not guaranteed to work every time.

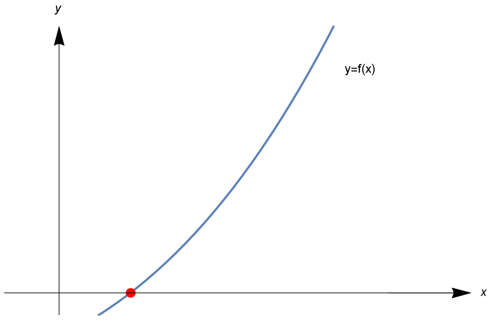
Newton's method

> Truth is much too complicated to allow anything but approximations.
>
> – John Von Neumann

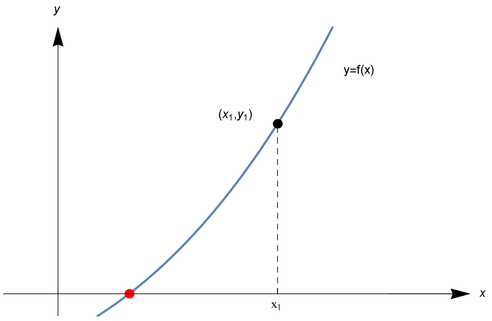
Finding a solution with geometry

Newton's method for solving equations is another numerical method for solving an equation $f(x) = 0$. It is based on the *geometry* of a curve, using the *tangent lines* to a curve. As such, it requires *calculus*, in particular *differentiation*.

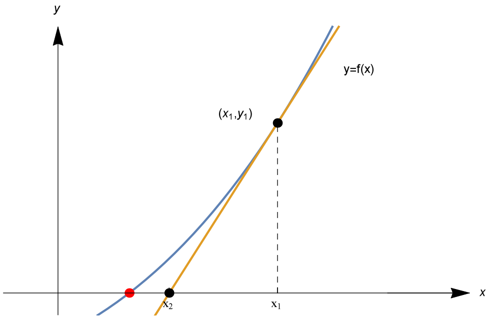
Roughly, the idea of Newton's method is as follows. We seek a solution to $f(x) = 0$. That is, we want to find the red dotted point in the picture below.



We start with an initial guess x_1. We calculate $f(x_1)$. If $f(x_1) = 0$, we are very lucky, and have a solution. But most likely $f(x_1)$ is not zero. Let $f(x_1) = y_1$, as shown.

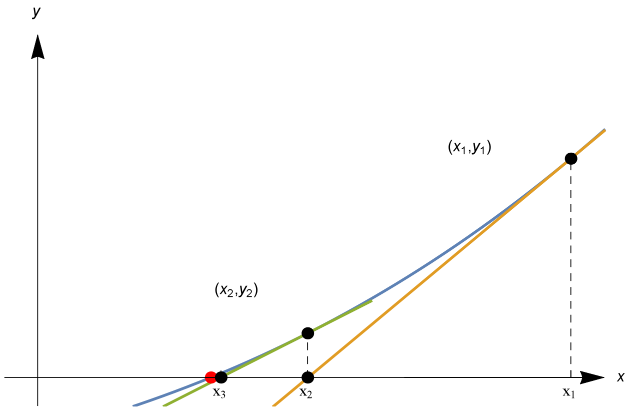


We now try for a better guess. How to find that better guess? The trick of Newton's method is to draw a *tangent line* to the graph $y = f(x)$ at the point (x_1, y_1). See below.



This tangent line is a good *linear approximation* to $f(x)$ near x_1, so our next guess x_2 is the point where the tangent line intersects the x-axis, as shown above.

We then proceed using the same method. We calculate $y_2 = f(x_2)$; if it is zero, we're finished. If not, then we draw the tangent line to $y = f(x)$ at (x_2, y_2), and our next guess x_3 is the point where this tangent line intersects the x-axis. See below.



In the figure shown, x_1, x_2, x_3 rapidly approach the red solution point!

Continuing in this way, we find points $x_1, x_2, x_3, x_4, \dots$ approximating a solution. This method for finding a solution is Newton's method.

As we'll see, Newton's method can be a very efficient method to approximate a solution to an equation — when it works.

The key calculation

As our introduction above just showed, the key calculation in each step of Newton's method is to find where the tangent line to $y = f(x)$ at the point (x_1, y_1) intersects the x-axis.

Let's find this x-intercept. The tangent line we are looking for passes through the point (x_1, y_1) and has gradient $f'(x_1)$.

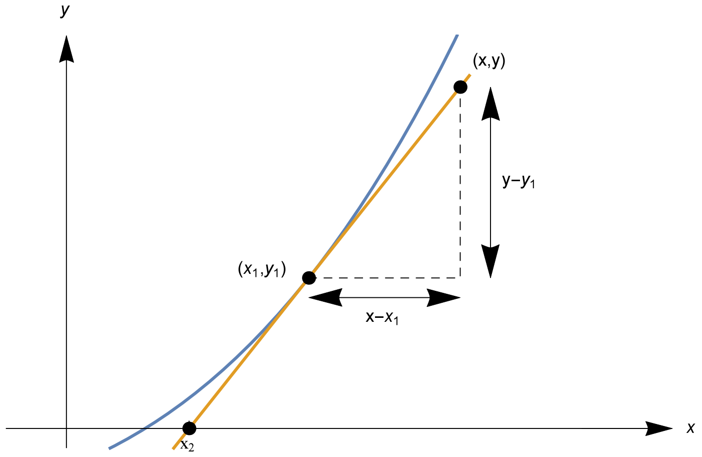
Recall that, given the gradient m of a line, and a point (x_0, y_0) on it, the line has equation

$$\frac{y - y_0}{x - x_0} = m, \quad \text{or equivalently,} \quad y = m(x - x_0) + y_0.$$

In our situation, the line has gradient $f'(x_1)$, and passes through (x_1, y_1), so has equation

$$\frac{y - y_1}{x - x_1} = f'(x_1), \quad \text{or equivalently,} \quad y = f'(x_1)(x - x_1) + y_1.$$

See the diagram below.



Setting $y = 0$, we find the x-intercept as

$$x = x_1 - \frac{y_1}{f'(x_1)} = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

The Algorithm

We can now describe Newton's method algebraically. Starting from x_1, the above calculation shows that if we construct the tangent line to the graph $y = f(x)$ at $x = x_1$, this tangent line has x-intercept given by

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

Then, starting from x_2 we perform the same calculation, and obtain the next approximation x_3 as

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}.$$

The same calculation applies at each stage: so from the n'th approximation x_n, the next approximation is given by

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

Formally, Newton's algorithm is as follows.

[Newton's method] Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function. We seek a solution of $f(x) = 0$, starting from an initial estimate $x = x_1$.

At the n'th step, given x_n, compute the next approximation x_{n+1} by

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

Some comments about this algorithm:

1 Often, Newton's method works extremely well, and the x_n converge rapidly to a solution. However, it's important to note that *Newton's method does not always work*. Several things can go wrong, as we will see shortly.

2 Note that if $f(x_n) = 0$, so that x_n is an *exact* solution of $f(x) = 0$, then the algorithm gives $x_{n+1} = x_n$, and in fact all of $x_n, x_{n+1}, x_{n+2}, x_{n+3}, \ldots$ will be equal. If you have an exact solution, Newton's method will stay on that solution!

3 While the bisection method only requires f to be continuous, Newton's method requires the function f to be *differentiable*. This is necessary for f to have a tangent line.

Using Newton's method

As we did with the bisection method, let's approximate some well-known constants.

Example

Use Newton's method to find an approximate solution to $x^2 - 2 = 0$, starting from an initial estimate $x_1 = 2$. After 4 iterations, how close is the approximation to $\sqrt{2}$?

Solution

Let $f(x) = x^2 - 2$, so $f'(x) = 2x$. At step 1, we have $x_1 = 2$ and we calculate $f(x_1) = 2^2 - 2 = 2$ and $f'(x_1) = 4$, so

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 2 - \frac{2}{4} = \frac{3}{2} = 1.5.$$

At step 2, from $x_2 = \frac{3}{2}$ we calculate $f(x_2) = \frac{1}{4}$ and $f'(x_2) = 3$, so

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = \frac{3}{2} - \frac{\frac{1}{4}}{3} = \frac{17}{12} \sim 1.41666667 \quad \text{(to 8 decimal places).}$$

At step 3, from $x_3 = \frac{17}{12}$ we have $f(x_3) = \frac{1}{144}$ and $f'(x_3) = \frac{17}{6}$, so

$$x_4 = x_3 - \frac{f(x_3)}{f'(x_3)} = \frac{17}{12} - \frac{\frac{1}{144}}{\frac{17}{6}} = \frac{577}{408} \sim 1.41421569 \quad \text{(to 8 decimal places).}$$

At step 4, from $x_4 = \frac{577}{408}$ we calculate $f(x_4) = \frac{1}{166{,}464}$ and $f'(x_4) = \frac{577}{204}$, so

$$x_5 = x_4 - \frac{f(x_4)}{f'(x_4)} = \frac{665{,}857}{470{,}832} \sim 1.414213562375 \quad \text{(to 12 decimal places).}$$

In fact, to 12 decimal places $\sqrt{2} \sim 1.414213562373$. So after 4 iterations, the approximate solution x_5 agrees with $\sqrt{2}$ to 11 decimal places. The difference between x_5 and $\sqrt{2}$ is

$$\frac{665{,}857}{470{,}832} - \sqrt{2} \sim 1.59 \times 10^{-12} \quad \text{to 3 significant figures.}$$

Newton's method in the above example is *much* faster than the bisection algorithm! In only 4 iterations we have 11 decimal places of accuracy! The following table illustrates how many decimal places of accuracy we have in each x_n.

n	1	2	3	4	5	6	7
# decimal places accuracy in x_n	0	0	2	5	11	23	47

The number of decimal places of accuracy approximately *doubles* with each iteration!

Exercise 10

(*Harder.*) This problem investigates this doubling of accuracy:

a First calculate that $x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}$ and that, if $y_n = x_n - \sqrt{2}$, then $y_{n+1} = \frac{y_n^2}{2(\sqrt{2}+y_n)}$.

b Show that, if $x_1 > 0$ (resp. $x_1 < 0$), then $x_n > \sqrt{2}$ (resp. $x_n < -\sqrt{2}$) for all $n \geq 2$.

c Show that if $0 < y_n < 10^{-k}$ for some positive integer k, then $0 < y_{n+1} < 10^{-2k}$.

d Show that, for $n \geq 2$, if x_n differs from $\sqrt{2}$ by less than 10^{-k}, then x_{n+1} differs from $\sqrt{2}$ by less than 10^{-2k}.

Example

Find an approximation to π by using Newton's method to solve $\sin x = 0$ for 3 iterations, starting from $x_1 = 3$. To how many decimal places is the approximate solution accurate?

Solution

Let $f(x) = \sin x$, so $f'(x) = \cos x$. We compute x_2, x_3, x_4 directly using the formula

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

We compute directly each x_n to 25 decimal places, for 3 iterations:

$$\begin{aligned}
x_2 &= x_1 - \tan x_1 = 3 - \tan 3 \sim 3.1425465430742778052956354 \\
x_3 &= x_2 - \tan x_2 \sim 3.1415926533004768154498858 \\
x_4 &= x_3 - \tan x_3 \sim 3.1415926535897932384626434
\end{aligned}$$

To 25 decimal places, $\pi = 3.1415926535897932384626433$; x_4 agrees to 24 places.

In the above example, convergence is even more rapid than for our $\sqrt{2}$ example. The number of decimal places accuracy roughly *triples* with each iteration!

n	1	2	3	4	5
# decimal places accuracy in x_n	0	2	9	24	87

Exercise 11

Find all solutions to the equation $\cos x = x$ to 9 decimal places using Newton's method. Compare the convergence to what you obtained with the bisection method in exercise 5.

What can go wrong

While Newton's method can give fantastically good approximations to a solution, several things can go wrong. We now examine some of this less fortunate behaviour.

The first problem is that the x_n may not even be defined!

Example

Find an approximate solution to $\sin x = 0$ using Newton's method starting from $x_1 = \frac{\pi}{2}$.

Solution

We set $f(x) = \sin x$, so $f'(x) = \cos x$. Now $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$, but $f'(x_1) = \cos\frac{\pi}{2} = 0$, so x_2 is not defined, hence nor are any subsequent x_n.

This problem occurs whenever when $f'(x_n) = 0$. If x_n is a stationary point of f, then Newton's method attempts to divide by zero — and fails.

The next example illustrates that even when the "approximate solutions" x_n exist, they may not "approximate" a solution very well.

Example

Approximate a solution to $-x^3 + 4x^2 - 2x + 2 = 0$ using Newton's method from $x_1 = 0$.

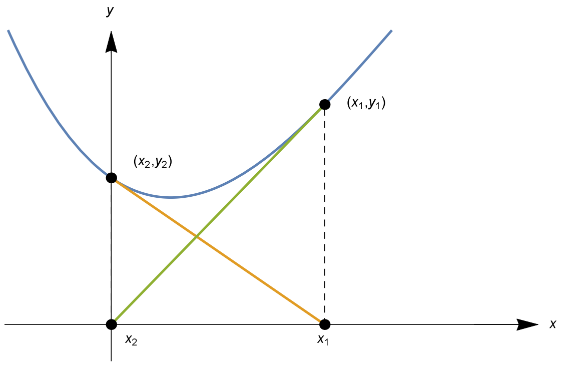
Solution

Let $f(x) = -x^3 + 4x^2 - 2x + 2$, so $f'(x) = -3x^2 + 8x - 2$. Starting from $x_1 = 0$, we have

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = -\frac{f(0)}{f'(0)} = -\frac{2}{-2} = 1, \quad \text{then} \quad x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = 1 - \frac{f(1)}{f'(1)} = 1 - \frac{3}{3} = 0.$$

We have now found that $x_3 = x_1 = 0$. The computation for x_4 is then exactly the same as for x_2, so $x_4 = x_2 = 1$. Thus $x_n = 0$ for all odd n, and $x_n = 1$ for all even n. The x_n are locked into a cycle, alternating between two values.

In the above example, there is actually a solution, $x = \frac{1}{3}\left(4 + \sqrt{10} + \sqrt[3]{100}\right) \sim 3.59867$. Our "approximate solutions" x_n fail to approximate it, instead cycling as illustrated below.



The third problem shows that all x_n can get further and further *away* from a solution!

Example

Find an approximate solution to $\arctan x = 0$, using Newton's method starting at $x_1 = 3$.

Solution

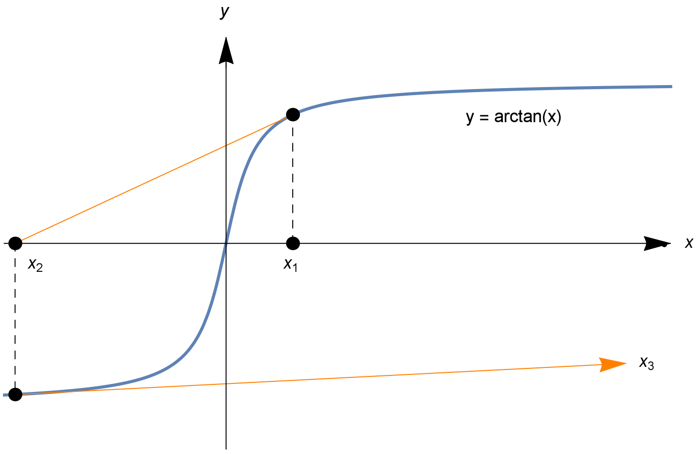
Let $f(x) = \arctan x$, so $f'(x) = \frac{1}{1+x^2}$. From $x_1 = 3$ we can compute successively $x_2, x_3, \ldots$. We write them to 2 decimal places.

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 3 - \frac{\arctan 3}{\frac{1}{1+3^2}} = 3 - 10\arctan 3 \sim -9.49.$$

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} \sim 124.00$$

$$x_4 = x_3 - \frac{f(x_3)}{f'(x_3)} \sim -23{,}905.94$$

The “approximations" x_n rapidly diverge to infinity.

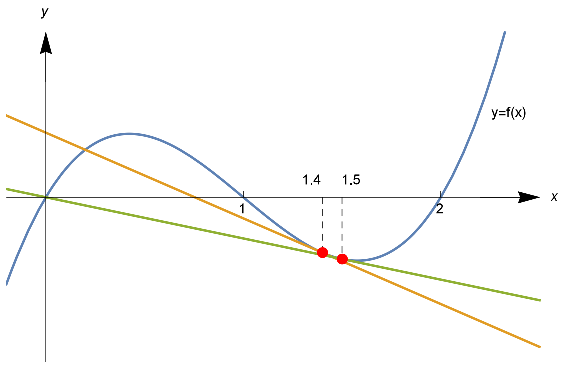


The above example may seem a little silly, since finding an exact solution to $\arctan x = 0$ is not difficult: the solution is $x = 0$! But similar problems will happen applying Newton's method to any curve with a similar shape. The fact that $f'(x_n)$ is close to zero means the tangent line is close to horizontal, and so may travel far away to arrive at its x-intercept of x_{n+1}.

Sensitive dependence on initial conditions

Sometimes the choice of initial conditions can be ever so slight, yet lead to a radically different outcome in Newton's method.

Consider solving $f(x) = 2x - 3x^2 + x^3 = 0$. It's possible to factorise the cubic, $f(x) = x(x-1)(x-2)$, so the solutions are just $x = 0, 1$ and 2.



If we apply Newton's method with different starting estimates x_1, we might end up at any of these three solutions... or somewhere else.

As illustrated above, we apply Newton's method starting from $x_1 = 1.4$ and $x_1 = 1.5$. Starting from $x_1 = 1.4 = \frac{7}{5}$ we compute, to 6 decimal places,

$$x_1 = \frac{7}{5} = 1.4,\; x_2 \sim 0.753846,\; x_3 \sim 1.036456,\; x_4 \sim 0.999903,\; x_5 \sim 1.000000.$$

and the x_n converge to the solution $x = 1$. And starting from $x_1 = 1.5$ we compute $x_2 = 0$, an exact solution, so all $x_3 = x_4 = \cdots = 0$.

We might then ask: where, between the initial estimates $x_1 = 1.4$ and $x_1 = 1.5$, does the sequence switch from approaching $x = 1$, to approaching $x = 0$?

We find that the the initial estimate $x_1 = 1.45$ behaves similarly to 1.5: the sequence x_n approaches the solution $x = 0$. On the other hand, taking $x_1 = 1.44$ behaves similarly to

1.4: x_n approaches the solution $x = 1$. The outcome seems to switch somewhere between the initial estimates $x_1 = 1.44$ and $x_1 = 1.45$. With some more computations, we can narrow down further and find that the switch happens somewhere between $x_1 = 1.447$ and $x_1 = 1.448$. The computations (to 3 decimal places) from these initial values are shown in the table below.

So, if we try $x_1 = 1.4475$, will the x_n converge to the solution $x = 1$ or $x = 0$? The answer is *neither*: as shown in the table, it converges to the solution $x = 2$!

x_1	1.4	1.44	1.447	1.4475	1.448	1.45	1.5
x_2	0.754	0.594	0.554	0.551	0.548	0.536	0
x_3	1.036	1.266	1.440	1.458	1.477	1.567	0
x_4	1.000	0.952	0.596	0.484	0.317	-9.156	0
x_5	1.000	1.000	1.260	2.370	-0.598	-5.793	0
x_6	1.000	1.000	0.956	2.111	-0.225	-3.562	0
x_7	1.000	1.000	1.000	2.015	-0.050	-2.091	0
x_8	1.000	1.000	1.000	2.000	-0.003	-1.135	0
x_9	1.000	1.000	1.000	2.000	0.000	-0.536	0
x_{10}	1.000	1.000	1.000	2.000	0.000	-0.192	0
x_{11}	1.000	1.000	1.000	2.000	0.000	-0.038	0
x_{12}	1.000	1.000	1.000	2.000	0.000	-0.002	0

Furthermore, if we try $x_1 = 1 + \frac{1}{\sqrt{5}} \sim 1.447214$, we find different behaviour again! As it turns out,

$$x_2 = 1 - \frac{1}{\sqrt{5}} \quad \text{and} \quad x_3 = 1 + \frac{1}{\sqrt{5}}$$

so $x_1 = x_3$ and the x_n alternate between these two values.

Exercise 12

Verify that if $f(x) = 2x - 3x^2 + x^3 = x(x-1)(x-2)$ and $x_1 = 1 + \frac{1}{\sqrt{5}}$, then $x_2 = 1 - \frac{1}{\sqrt{5}}$ and $x_3 = 1 + \frac{1}{\sqrt{5}} = x_1$.

So there is no simple transition from those values of x_1 which lead to the solution $x = 1$ and $x = 0$. The behaviour of Newton's method depends extremely sensitively on the initial x_1. Indeed, we have only tried only a handful of values for x_1, and have only seen the tip of the iceberg of this behaviour.

In the *Links Forward* section we examine this behaviour further, showing some pictures of this sensitive dependence on initial conditions, which is indicative of mathematical *chaos*. Pictures illustrating the behaviour of Newton's method show the intricate detail of a *fractal*.

Getting Newton's method to work

In general, it is difficult to state precisely when Newton's method will provide good approximations to a solution of $f(x) = 0$. But we can make the following notes, summarising our previous observations.

- If $f'(x_n) = 0$ then Newton's method immediately fails, as it attempts to divide by zero.
- The "approximations" $x_1, x_2, x_3, \ldots$ can cycle, rather than converging to a solution.
- The "approximate solutions" $x_1, x_2, x_3, \ldots$ can even diverge to infinity. This problem happens when $f'(x_n)$ is small, but not zero.
- The behaviour of Newton's method can depend extremely sensitively on the choice of x_1, and lead to chaotic dynamics.

However, a useful rule of thumb is as follows:

- Suppose you can sketch a graph of $y = f(x)$, and you can see the approximate location of a solution, and $f'(x)$ is not too small near that solution. Then taking x_1 near that solution, and away from other solutions or critical points, Newton's method will tend to produce x_n which converge rapidly to the solution.

Links Forward

> Everything we know is only some kind of approximation, because we know that we do not know all the laws yet. Therefore, things must be learned only to be unlearned again or, more likely, to be corrected.
>
> – Richard Feynman

Speed of the bisection algorithm

As we saw previously, one nice aspect of the bisection algorithm is that it always works. We now consider its accuracy more precisely, and *how long* it takes to work.

In order to examine the accuracy of the method, we find the length of each $[a_n, b_n]$. Exercise 2 asked you to find the length of each $[a_n, b_n]$ in terms of the original interval

$[a_0, b_0] = [a, b]$. Now $[a_0, b_0] = [a, b]$ has length $b - a$, and at each step we bisect this interval. So the length $b_n - a_n$ of $[a_n, b_n]$ is obtained by dividing $b - a$ by 2, n times:

$$b_n - a_n = \frac{b-a}{2^n}.$$

Hence, $[a_n, b_n]$ provides us an accuracy of $\frac{b-a}{2^n}$ for a solution.

Exercise 13

Suppose you apply the bisection method to solve $f(x) = 0$ in the interval $[a, b]$. After n iterations,you make a best guess for the solution. How close is your guess to a solution?

In fact, we can calculate just how many steps are required in the bisection algorithm to obtain a desired degree of accuracy.

Example

Using the bisection algorithm to find a solution to the equation $f(x) = 0$, starting from the interval $[a, b] = [3, 5]$, to an accuracy of 0.001, how many steps are required?

Solution

From above, the n'th interval $[a_n, b_n]$ has length

$$b_n - a_n = \frac{b-a}{2^n} = \frac{5-3}{2^n} = 2 \cdot 2^{-n}.$$

To obtain a solution to within an accuracy of 0.001, we require $b_n - a_n < 0.001$, so

$$2 \cdot 2^{-n} < 0.001, \quad \text{hence} \quad \frac{2}{0.001} < 2^n$$

$$n > \log_2\left(\frac{2}{0.001}\right) = \log_2 2000 \sim 10.966.$$

So if we take $n = 11$ steps, we will have the desired accuracy.

Note that to find how many steps are required, we do not need *any* information about the function f! All we need to know is the length of $[a, b]$, and the desired accuracy.

Exercise 14

Starting from the interval $[a, b] = [-5, 5]$, how many steps are required in the bisection algorithm to obtain a solution to an accuracy of 0.05?

In general, to obtain an accuracy of ϵ, starting from the interval $[a, b]$, the bisection method will require N iterations, where

$$N > \log_2\left(\frac{b-a}{\epsilon}\right) \quad \text{steps.}$$

Exercise 15

Prove this.

Bisection and binary numbers

The binary number system is in many ways well suited to the bisection method.

To see why, suppose we are searching for a solution to $f(x) = 0$ in the interval $[0,32]$, and suppose there is a unique solution $x = 19$ (although we don't know that yet!).

Using the bisection method, we will obtain several intervals $[a_n, b_n]$, and when we arrive at $x = 19$, we will discover that we have found the solution exactly. We can write out these intervals; we do so in both decimal and binary notation.

Interval	$[a_0, b_0]$	$[a_1, b_1]$	$[a_2, b_2]$	$[a_3, b_3]$	$[a_4, b_4]$
Decimal	$[0,32]$	$[16,32]$	$[16,24]$	$[16,20]$	$[18,20]$
Binary	[0, 100000]	[10000, 100000]	[10000, 11000]	[10000, 10100]	[10010, 10100]

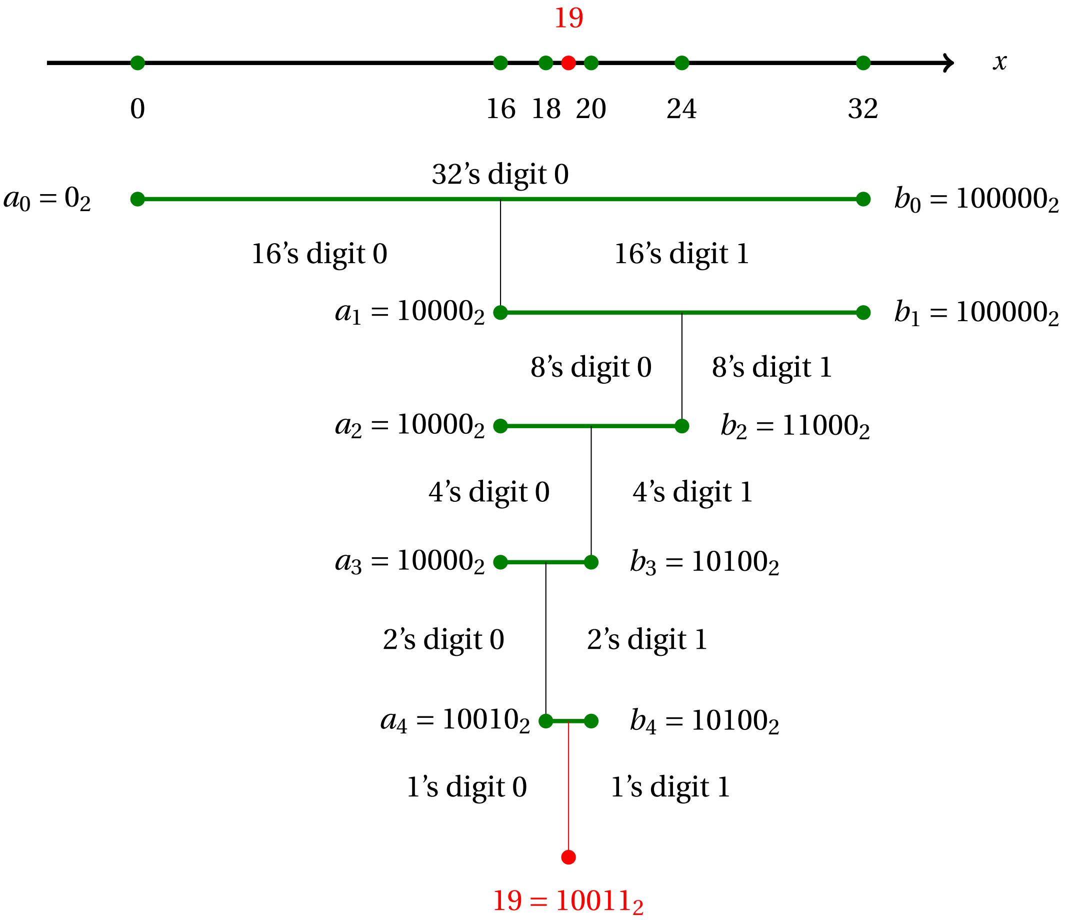
We illustrate these intervals below, denoting binary numbers with a subscript 2. As in previous illustrations, the intervals $[a_n, b_n]$ are shown in green and the solution in red.

At the first step, we split the interval from 0 to 32 at the midpoint 16. Now the numbers from 0 to 16 are precisely those which have a 0 in the 16's digit, and the numbers from 16 to 32 are those which have a 1 in the 16's digit. Therefore, at the first step of the bisection algorithm, we determine the 16's digit of the solution. Since in our example $[a_1, b_1] = [16,32]$ on the right, the solution has a 1 in the 16's digit.

At the second step, we split the interval from 16 to 32 at the midpoint 24; the lower half (from 16 to 24) consists of numbers with a 0 in the 8's digit, and the upper half (from 16 to 32) consists of numbers with a 1 in the 8's digit. So at the second step of the bisection method, we determine the 8's digit of the solution. Since $[a_2, b_2]$ is the lower half $[16,24]$, the solution has a 0 in the 8's digit.

Continuing in this way, each step of the bisection method provides another binary digit of the solution. The next interval $[a_3, b_3]$ is the lower half of $[a_2, b_2]$, so the solution has a 0 in the 4's digit. Then $[a_4, b_4]$ is the upper half of $[a_3, b_3]$, so the solution has a 1 in the 2's digit. Finally we arrive at the solution, which is 10011 in binary.

Provided that we start from an interval which fits nicely with the binary number system, each step of the bisection algorithm determines a further binary digit of the solution.



This continues even when we consider fractions: we determine binary digits which appear after the “decimal" point.

The bisection method is well suited to the *binary*, or *base 2* number system, because it splits the interval at each stage into *two* pieces, which can align with binary digits.

The bisection method is less well suited to the *decimal* number system. The intervals $[a_n, b_n]$ do not correspond nicely to decimal digits. As we have found previously, the number of iterations required to find a solution to a desired number of decimal places can vary rather haphazardly.

Exercise 16

Can you describe a “decimation method" for solving an equation $f(x) = 0$, analogous to the bisection method, but which at each stage produces an interval *one tenth* the size of the previous interval, and at each stage determines one *decimal* digit of the solution?

Complex numbers and Newton fractals

Newton's method works just as well for complex numbers as for real numbers: sometimes finding a solution at blistering speed, and sometimes failing to work at all.

For instance, suppose we want to solve the equation $z^3 = 1$. There is just one *real* solution $z = 1$, but over the complex numbers there are *three* solutions:

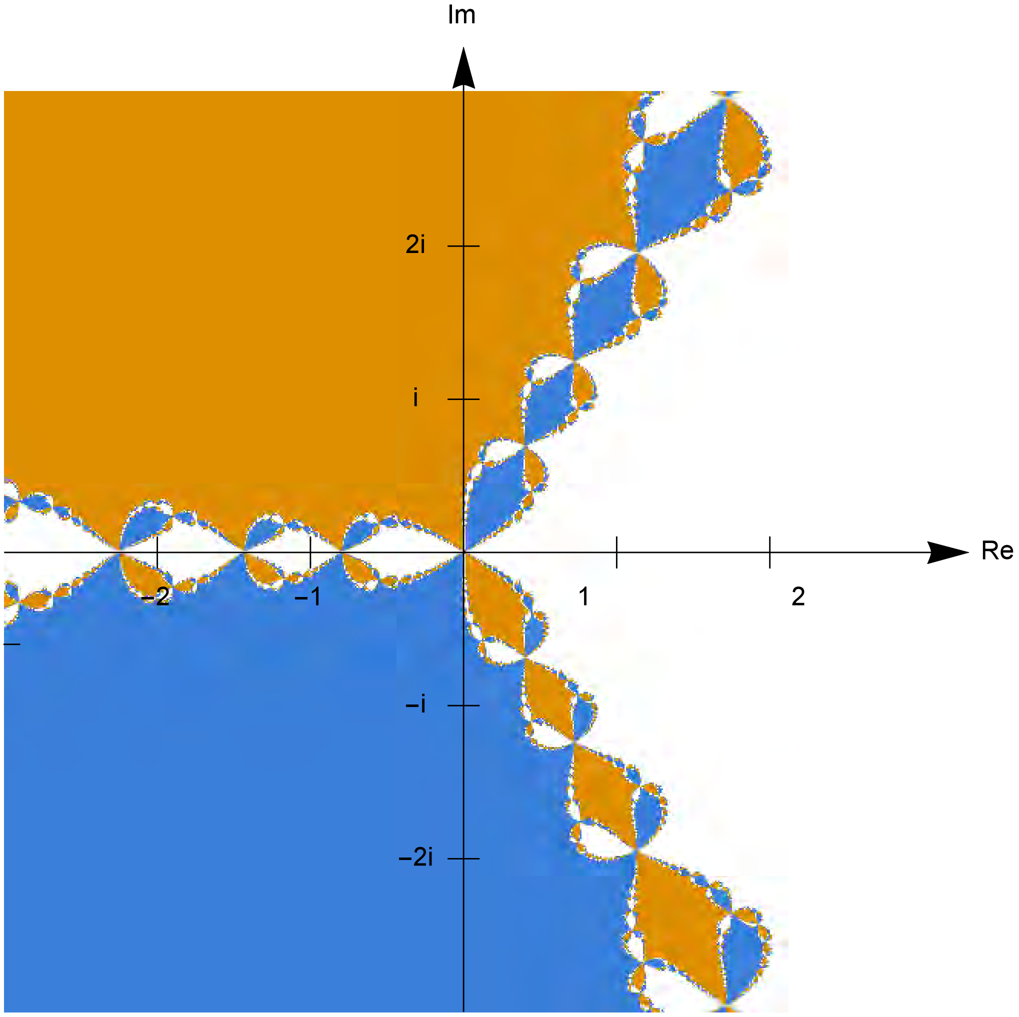
$$z = 1,\ -\frac{1}{2} + \frac{\sqrt{3}}{2}i,\ -\frac{1}{2} - \frac{\sqrt{3}}{2}i.$$

Starting from an initial point z_1, Newton's method works just as over the reals. We let $f(z) = z^3 - 1$, so $f'(z) = 3z^2$, and then calculate

$$z_2 = z_1 - \frac{f(z_1)}{f'(z_1)} = z_1 - \frac{z^3 - 1}{3z^2}.$$

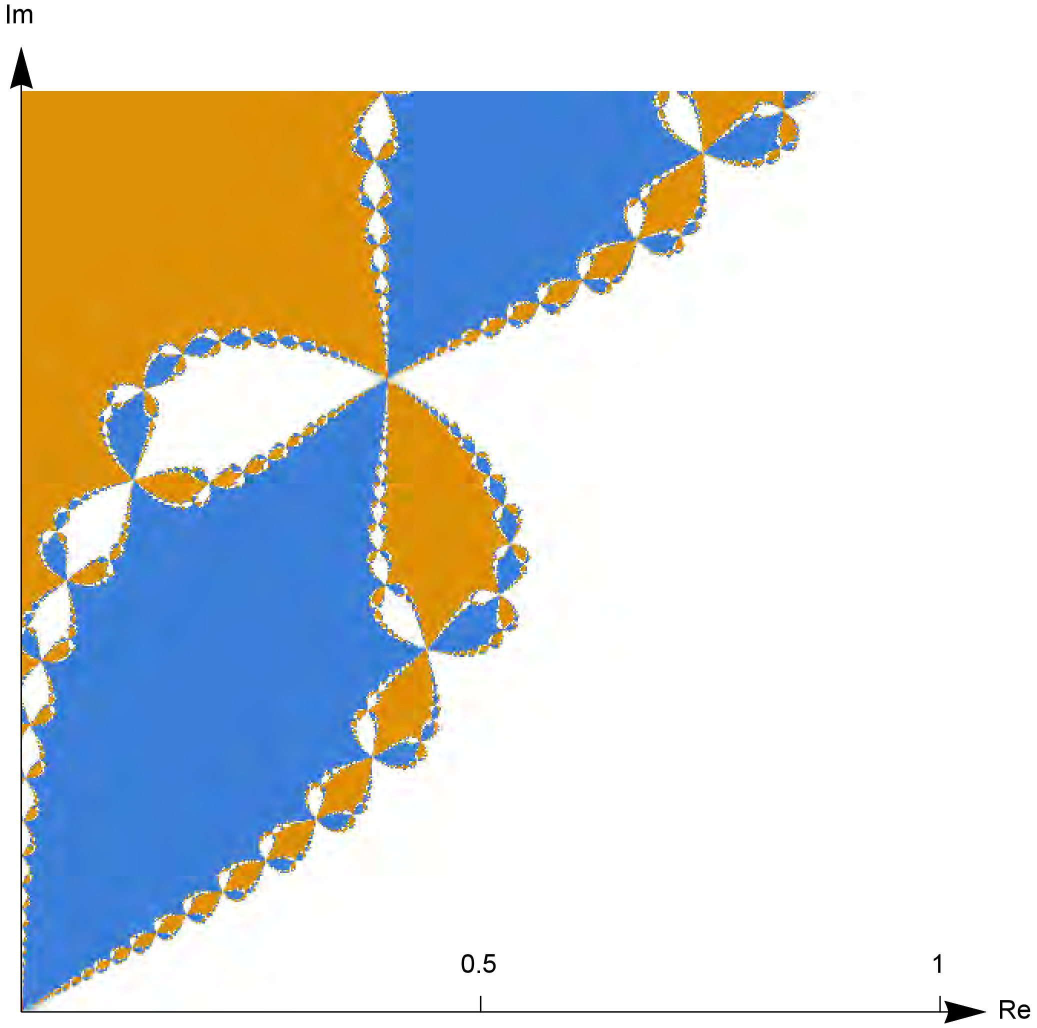
From z_2, we calculate z_3, and then z_4, and so on.

We can ask: which choices of z_1 lead to which solutions? This leads to a very interesting picture. The picture below shows the complex plane, with real and imaginary axes labelled. For each point complex number z, we run Newton's method with $z_1 = z$, and see where the z_n go. If they approach $-\frac{1}{2} + \frac{\sqrt{3}}{2}i$, we colour the point brown. If they approach $-\frac{1}{2} - \frac{\sqrt{3}}{2}i$, we colour the point blue. And if they approach 1, we colour the point white.



We can see that, although there are large regions which converge to the three roots, there is an intricately complicated structure between these regions. We find complex numbers very close together, converging to different solutions, arranged in an intricate pattern.

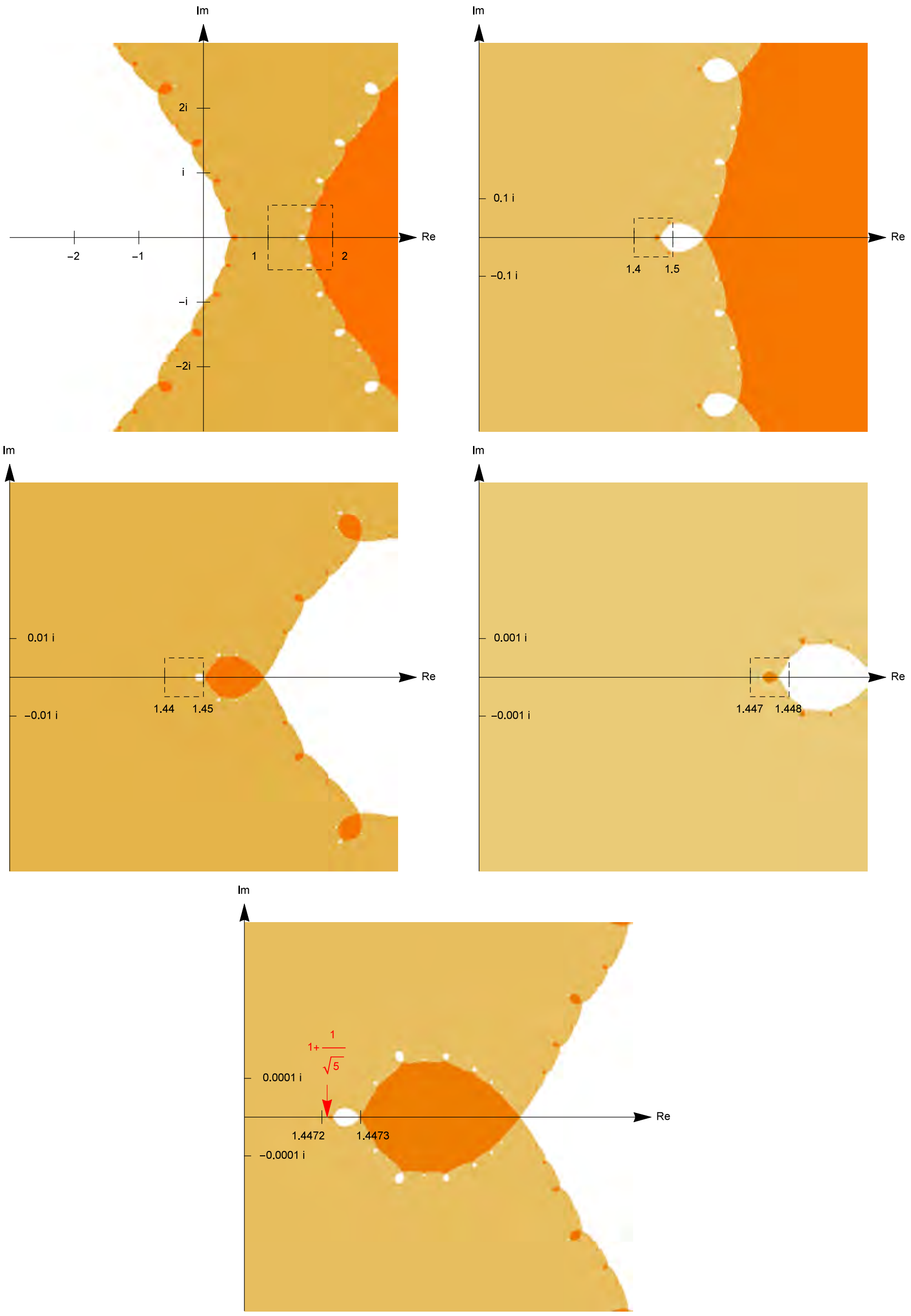
In fact, if you zoom in on the details, you see even more detail. The more you zoom in, the more detail you see — and the details as you zoom in are similar to the details seen before you zoomed in. The pattern is *self-similar*, and is an object known as a *fractal*. For instance, if we zoom in on the square with corners $0, 1, i, 1+i$, we see an equally intricate picture.



The intricate arrangement of colours show just how sensitive Newton's method can be to initial conditions. The tiniest change in the initial estimate x_1 can move to a region of different colour, so that Newton's method leads to a completely different outcome.

Previously, we made a table showing outcomes when trying to solve the equation $2x - 3x^2 + x^3 = 0$. The solutions are $x = 0, 1, 2$, but we found very sensitive dependence on initial conditions for x_1 between 1.4 and 1.5. In fact, the structure is much more intricate.

In the pictures below, points are coloured white, light brown, or orange, respectively as Newton's method approaches 0, 1 or 2. Each subsequent picture is obtained by repeatedly zooming in on the dotted square in the previous picture. The final picture shows the value of $1 + \frac{1}{\sqrt{5}}$, which gives no solution but rather cycles between 2 values.



Answers to exercises

Exercise 1

For example, we could define

$$f(x) = \begin{cases} -1 & 0 \le x < \frac{1}{2} \\ 1 & \frac{1}{2} \le x \le 1 \end{cases}$$

Exercise 2

As each successive $[a_n, b_n]$ has half the length of $[a_{n-1}, b_{n-1}]$, the interval $[a_n, b_n]$ is $\frac{1}{2^n}$ times the size of $[a_0, b_0]$. Thus $\frac{b_n - a_n}{b-a} = \frac{1}{2^n}$.

Exercise 3

Suppose that $f(a_0)$ is positive and $f(b_0)$ is negative. We will prove by induction that for all non-negative integers n, $f(a_n)$ is positive and $f(b_n)$ is negative. The claim is clearly for $n = 0$. Now suppose the claim holds for $n = k$, so $f(a_k) > 0$ and $f(b_k) < 0$; we will prove the claim for $n = k+1$, showing that $f(a_{k+1}) > 0$ and $f(b_{k+1}) < 0$.

Bisecting $[a_k, b_k]$, we consider $f\left(\frac{a_k+b_k}{2}\right)$. If $f\left(\frac{a_k+b_k}{2}\right)$ is positive, then $[a_{k+1}, b_{k+1}] = \left[\frac{a_k+b_k}{2}, b_{k+1}\right]$, so $f(a_{k+1}) > 0$ and $f(b_{k+1}) < 0$ as claimed. Alternatively, if $f\left(\frac{a_k+b_k}{2}\right) < 0$, then $[a_{k+1}, b_{k+1}] = \left[a_k, \frac{a_k+b_k}{2}\right]$, so $f(a_{k+1}) > 0$ and $f(b_{k+1}) < 0$ again. Either way, $f(a_{k+1}) > 0$ and $f(b_{k+1}) < 0$. By induction, all $f(a_n) > 0$ and all $f(b_n) < 0$.

One can prove, by a similar method, that if $f(a_0)$ is negative and $f(b_0)$ is positive, then for all non-negative integers n, $f(a_n)$ is negative and $f(b_n)$ is positive.

Exercise 4

Let $f(x) = \ln x - 1$, so we solve $f(x) = 0$ using the bisection method with $[a_0, b_0] = [2, 3]$. We tabulate the calculation as follows.

Step		Sign		Sign	Midpoint	Sign
n	a_n	$f(a_n)$	b_n	$f(b_n)$	$\frac{a_n+b_n}{2}$	$f\left(\frac{a_n+b_n}{2}\right)$
0	2	–	3	+	$\frac{5}{2} = 2.5$	–
1	$\frac{5}{2} = 2.5$	–	3	+	$\frac{11}{4} = 2.75$	+
2	$\frac{5}{2} = 2.5$	–	$\frac{11}{4} = 2.75$	+	$\frac{21}{8} = 2.625$	–
3	$\frac{21}{8} = 2.625$	–	$\frac{11}{4} = 2.75$	+	$\frac{43}{16} = 2.6875$	–
4	$\frac{43}{16} = 2.6875$	–	$\frac{11}{4} = 2.75$	+	$\frac{87}{32}$	+
5	$\frac{43}{16} = 2.6875$	–	$\frac{87}{32} = 2.71875$	+	$\frac{173}{64}$	–
6	$\frac{173}{64} = 2.703125$	–	$\frac{87}{32} = 2.71875$	+		

Thus $e \in \left[\frac{173}{64}, \frac{87}{32}\right] = [2.703125, .71875]$, and e has been calculated to within an accuracy of $\frac{87}{32} - \frac{173}{64} = \frac{1}{64} = 0.015625 < 0.02$.

Exercise 5

There is just one solution. Let $f(x) = \cos x - x$. We note $f(0) = 1$ and $f(1) < 0$, so there is a solution in $[0,1]$. (One can check that for negative x, $f(x) > 0$, and for $x > 1$, $f(x) < 0$.) Using the bisection method starting from $[a_0, b_0] = [0,1]$, after 7 iterations we arrive at $[a_7, b_7] = \left[\frac{47}{64}, \frac{95}{128}\right] = [0.734375, 0.7421875]$, which has length $1/128 < 0.01$. To 6 decimal places the solution is 0.739085.

Exercise 6

Letting $f(x) = 3x - 1$, we apply the bisection method to solve $f(x) = 0$ starting from the interval $[0,1]$.

Step		Sign of		Sign of	Midpoint	Sign of
n	a_n	$f(a_n)$	b_n	$f(b_n)$	$\frac{a_n+b_n}{2}$	$f\left(\frac{a_n+b_n}{2}\right)$
0	0	$-$	1	$+$	$\frac{1}{2} = 0.5$	$+$
1	0	$-$	$\frac{1}{2} = 0.5$	$+$	$\frac{1}{4} = 0.25$	$-$
2	$\frac{1}{4} = 0.25$	$-$	$\frac{1}{2} = 0.5$	$+$	$\frac{3}{8} = 0.375$	$+$
3	$\frac{1}{4} = 0.25$	$-$	$\frac{3}{8} = 0.375$	$+$	$\frac{5}{16} = 0.3125$	$-$
4	$\frac{5}{16} = 0.3125$	$-$	$\frac{3}{8} = 0.375$	$+$	$\frac{11}{32} = 0.34375$	$+$
5	$\frac{5}{16} = 0.3125$	$-$	$\frac{11}{32} = 0.34375$	$+$	$\frac{21}{64} = 0.328125$	$-$
6	$\frac{21}{64} = 0.328125$	$-$	$\frac{11}{32} = 0.34375$	$+$	$\frac{43}{128} = 0.3359375$	$+$
7	$\frac{21}{64} = 0.328125$	$-$	$\frac{43}{128} = 0.3359375$	$+$	$\frac{85}{256} = 0.33203125$	$-$
8	$\frac{85}{256} = 0.33203125$	$-$	$\frac{43}{128} = 0.3359375$	$+$	$\frac{171}{512} = 0.333984375$	$+$
9	$\frac{85}{256} = 0.33203125$	$-$	$\frac{171}{512} = 0.333984375$	$+$	$\frac{341}{1024} = 0.3330078125$	$-$
10	$\frac{341}{1024} = 0.3330078125$	$-$	$\frac{171}{512} = 0.333984375$	$+$		

So to within accuracy of $\frac{1}{1024} < 0.001$, $1/3$ is approximated to lie in the interval $\left[\frac{341}{1024}, \frac{171}{512}\right] = [0.3330078125, 0.333984374]$.

We note that the signs of $f\left(\frac{a_n+b_n}{2}\right)$ alternate $+$, $-$, $+$, $-$, etc. This corresponds to the fact that, in binary, $1/3$ is written as $0.0101010101\cdots = 0.\overline{01}$. The alternating zeroes and ones correspond to the alternating between positive and negative signs. See the section “Bisection and binary numbers" above for details.

Exercise 7

In the example, we obtained an interval of $\left[\frac{181}{128}, \frac{91}{64}\right] = [1.4140625, 1.421875]$ of length $\frac{1}{128}$ approximating $\sqrt{2}$. The best guess for $\sqrt{2}$ is then the midpoint of this interval, $\frac{363}{256} = 1.41796875$. It is within $\frac{1}{256} = 0.00390625$ of a solution.

Exercise 8

The next intervals obtained after $[a_5, b_5] = \left[\frac{25}{8}, \frac{101}{32}\right] = [3.125, 3.15625]$ are $[a_6, b_6] = [3.140625, 3.15625]$, $[a_7, b_7] = [3.140625, 3.1484375]$, $[a_8, b_8] = [3.140625, 3.14453125]$. As both a_8 and b_8 round to 3.14 to two decimal places, we have approximated $\pi \sim 3.14$, to two decimal places.

Exercise 9

After $[a_7, b_7] = 1.4140625, 1.421875]$ the bisection method yields $[a_8, b_8] = [1.4140625, 1.41796875]$, $[a_9, b_9] = [1.4140625, 1.416015625]$, $[a_{10}, b_{10}] = [1.4140625, 4.4150390625]$, $[a_{11}, b_{11}] = [1.4140625, 1.41455078$ Both endpoints round to 1.41 to 2 decimal places, so $\sqrt{2} \sim 1.41$ to 2 decimal places.

Exercise 10

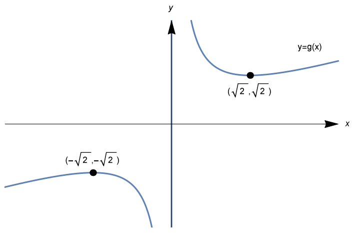
For part (a), since $f(x) = x^2 - 2$ and $f'(x) = 2x$, we have

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - 2}{2x_n} = x_n - \frac{1}{2}x_n + \frac{1}{x_n} = \frac{x_n}{2} + \frac{1}{x_n}.$$

Now substituting $x_n = y_n + \sqrt{2}$ gives $y_{n+1} + \sqrt{2} = \frac{y_n + \sqrt{2}}{2} + \frac{1}{y_n + \sqrt{2}}$ so

$$y_{n+1} = \frac{y_n + \sqrt{2}}{2} + \frac{1}{y_n + \sqrt{2}} - \sqrt{2} = \frac{(y_n + \sqrt{2})^2 + 2 - 2\sqrt{2}(y_n + \sqrt{2})}{2(\sqrt{2} + y_n)} = \frac{y_n^2}{2(\sqrt{2} + y_n)}.$$

For part (b), let $g(x) = \frac{x}{2} + \frac{1}{x}$, so $x_{n+1} = g(x_n)$. The function g is graphed below. When $x > 0$, $g(x) > 0$; and when $x < 0$, $g(x) < 0$. For $x > 0$, the minimum value of g is given by $g(\sqrt{2}) = \sqrt{2}$; when $x < 0$, the maximum value of g is given by $g(-\sqrt{2}) = -\sqrt{2}$. Hence if $x > 0$ then $g(x) \geq \sqrt{2}$; and if $x < 0$ then $g(x) \leq -\sqrt{2}$. So if $x_1 > 0$ then $x_2 = g(x_1) \geq \sqrt{2}$, and all subsequent $x_n \geq \sqrt{2}$. Similarly, if $x_1 < 0$ then all subsequent $x_n \leq -\sqrt{2}$.



For part (c), suppose $0 < y_n < 10^{-k}$. Then $y_{n+1} = \frac{y_n^2}{2(\sqrt{2}+y_n)}$ is positive, and

$$y_{n+1} = \frac{y_n^2}{2(\sqrt{2} + y_n)} < \frac{y_n^2}{2\sqrt{2}} < \frac{10^{-2k}}{2\sqrt{2}} < 10^{-2k} \quad \text{as desired.}$$

For part (d), if x_n differs from $\sqrt{2}$ by less than 10^{-k}, for an integer $k \geq 1$, then x_n is positive. By (b) above x_1 is positive; and as $n \geq 2$ then $x_n > \sqrt{2}$. So $y_n = x_n - \sqrt{2} \in (0, 10^{-k})$. By (c) above then $y_{n+1} \in (0, 10^{-2k})$, so x_{n+1} differs from $\sqrt{2}$ by less than 10^{-2k}.

Exercise 11

We apply Newton's method to $f(x) = \cos x - x$. We know, by the intermediate value theorem, that there is a solution in $[0,1]$, so we can start at $x_1 = 1$. We obtain (to 9 decimal places) $x_1 = 0.750363868$, $x_2 = 0.739112891$, $x_3 = 0.739085133$, $x_4 = 0.739085133$. After only 3 iterations we have the solution to 9 decaimal places. This is much better than the bisection method, which took 7 iterations to obtain 6 decimal places.

Exercise 12

We have $f(x) = x(x-1)(x-2) = 2x - 3x^2 + x^3$ and $f'(x) = 2 - 6x + 3x^2 = 3\left[(x-1)^2 - \frac{1}{3}\right]$. Substituting $x = 1 \pm \frac{1}{\sqrt{5}}$, we obtain

$$f(x) = x(x-1)(x-2) = \left(1 \pm \frac{1}{\sqrt{5}}\right)\left(\pm\frac{1}{\sqrt{5}}\right)\left(-1 \pm \frac{1}{\sqrt{5}}\right) = \frac{\pm 1}{5\sqrt{5}}\left(1^2 - \left(\sqrt{5}\right)^2\right) = \frac{\mp 4}{5\sqrt{5}}$$

$$f'(x) = 3\left[\left(\pm\frac{1}{\sqrt{5}}\right)^2 - \frac{1}{3}\right] = 3\left[\frac{1}{5} - \frac{1}{3}\right] = \frac{-2}{5}$$

$$x - \frac{f(x)}{f'(x)} = 1 \pm \frac{1}{\sqrt{5}} - \frac{\frac{\mp 4}{5\sqrt{5}}}{\frac{-2}{5}} = 1 \pm \frac{1}{\sqrt{5}} \mp \frac{2}{\sqrt{5}} = 1 \mp \frac{1}{\sqrt{5}}$$

Thus if $x_1 = 1 + \frac{1}{\sqrt{5}}$ then $x_2 = 1 - \frac{1}{\sqrt{5}}$ and $x_3 = 1 + \frac{1}{\sqrt{5}}$.

Exercise 13

The original interval has length $b - a$. Afer n iterations the interval has length $\frac{b-a}{2^n}$. The best guess is the midpoint of this interval; its distance from the solution can be no more than half the length of this interval. So this best guess is within $\frac{b-a}{2^{n+1}}$ of a solution.

Exercise 14

The first interval $[a_0, b_0] = [-5, 5]$ has length 10, so after n iterations, the interval $[a_n, b_n]$ has length $\frac{10}{2^n}$. To obtain an accuracy of 0.05 we require $\frac{10}{2^n} < 0.05$, so $2^n > \frac{10}{0.05} = 200$, hence $n > \log_2(200) \sim 7.64$, so 8 steps are required.

Exercise 15

The first interval has length $b - a$, so after n iterations, the interval $[a_n, b_n]$ has length $\frac{b-a}{2^n}$. To obtain an accuracy of ϵ, we require $\frac{b-a}{2^n} < \epsilon$. Rearranging this gives $2^n > \frac{b-a}{\epsilon}$, or $n > \log_2\left(\frac{b-a}{\epsilon}\right)$. So the bisection method will require at least $\log_2\left(\frac{b-a}{\epsilon}\right)$ steps.

Exercise 16

Given an interval $[a, b] = [a_0, b_0]$, you could divide it into *ten* subintervals at the 9 points $\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}$ from a to b, i.e. at the points $a + \frac{1}{10}(b-a)$, $a + \frac{2}{10}(b-a)$, ..., $a + \frac{9}{10}(b-a)$. The

j'th point is $a+\frac{j}{10}(b-a) = \frac{10-j}{10}a+\frac{j}{10}b$. You could then evaluate f at these 9 points. If $f = 0$ at any of these points, we have an exact solution. Otherwise, as f changes sign between a and b, it must change sign on at least one of the sub-intervals; one such sub-interval is chosen as $[a_1, b_1]$. Repeating this method produces a sequence of nested intervals $[a_0, b_0] \supset [a_1, b_1] \supset [a_2, b_2] \supset \cdots$, each of which contains a solution, and each of which is one tenth the size of the previous interval. If $[a_0, b_0] = [0, 1]$, then each successive $[a_n, b_n]$ determines a successive decimal digit of the solution.

Years

11 & 12