Iterated Functions

Tom Davis
tomrdavis@earthlink.net
http://www.geometer.org/mathcircles
October 19, 2011

Abstract

In this article we will examine various properties of iterated functions. If $f(x)$ is a function, then the iterates of f are: $f(x), f(f(x)), f(f(f(x))), \ldots$.

1 Introduction

I first became fascinated by iterated functions when I had a scientific calculator for the first time and repeatedly pressed the "cosine" button.

The calculator was in "radians" mode, so the angles were interpreted in radians rather than degrees[1], but I found that no matter what number I started with, after enough button presses of the "cosine" button, the numbers seemed to approach $.739085133215$.

What I was doing, of course, was iterating the cosine function. If my starting number was 1, then pressing the cosine button repeatedly was generating the following sequence of numbers:

$$\begin{aligned}
\cos(1) &= .540302305868\\
\cos(\cos(1)) &= .857553215846\\
\cos(\cos(\cos(1))) &= .654289790498\\
\cos(\cos(\cos(\cos(1)))) &= .793480358742\\
\cos(\cos(\cos(\cos(\cos(1))))) &= .701368773623
\end{aligned}$$

As I continued pressing the cosine button, the numerical results kept oscillating to values above and below, but each time closer to, the final limiting value of approximately $.739085133215$.

Of course there's nothing special about the cosine function; any function can be iterated, but not all iterated functions have the same nice convergence properties that the cosine function has. In this paper, we'll look at various forms of iteration.

2 A Simple Practical Example

Suppose you put some money (say x dollars) in a bank at a fixed interest rate. For example, suppose the bank offers simple interest at 10% per year. At the end of one year, you will have your original x dollars plus $(.10)x$ dollars of interest, or $x + (.10)x = (1.10)x$ dollars. In other words, if you begin with any amount of money, one year later you will have that amount multiplied by 1.10.

[1] One radian is equal to about 57.2957795131 degrees. If you've never seen it, this seems like a strange way to measure angles but it makes very good sense. In a unit circle (a circle with radius 1), the circumference is 2π, and if we measure in terms of that length instead of in terms of 360 degrees, we find that 2π radians is 360 degrees, from which the conversion above can be derived.

Suppose you'd like to know how much money you have after 5 or 10 years. If you consider the increase in value over one year to be a function named f, then we will have:

$$f(x) = (1.10)x.$$

The function f will take any input value and tell you the resulting output value if that input value is left in the bank for one year. Thus if you start with x dollars, then after one year, you will have $f(x)$ dollars. At the beginning of the second year you have $f(x)$ dollars, so at the end of the second year, you'll have $f(f(x))$ dollars. Similar reasoning yields $f(f(f(x)))$ dollars after the third year, $f(f(f(f(x))))$ dollars at the end of the fourth year, and so on.

It would be nice to have an easier notation for function iteration, especially if we iterate 100 or 1000 times. Some people use an exponent, like this:

$$f(f(f(f(x)))) = f^4(x),$$

but there's a chance that this could be confused with regular exponentiation (and it could, especially with functions like the cosine function, mentioned above). So we will use the following notation with parentheses around the iteration number:

$$f(f(f(f(x)))) = f^{(4)}(x).$$

As with standard exponentiation, we'll find it is sometimes useful to define:

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

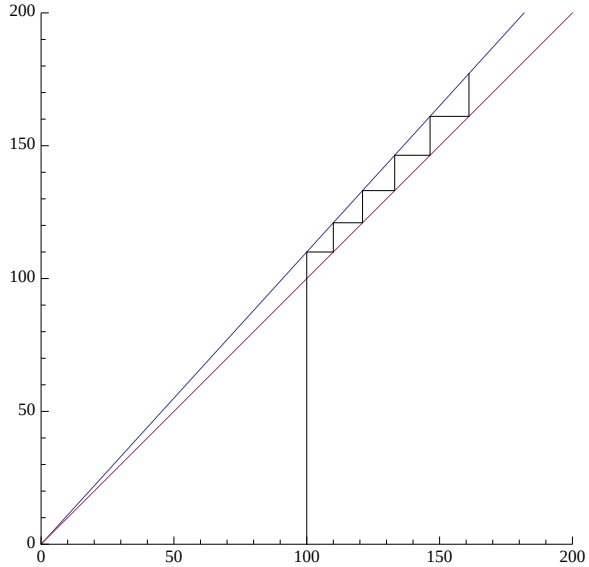
Returning to our example where $f(x)$ represents the amount of money in your account a year after investing x dollars, then the amount you'll have after 10 years would be $f^{(10)}(x)$.

It is fairly easy to derive exactly the form of $f^{(n)}(x)$ for this example. Since each year, the amount of money is multiplied by 1.10, we can see that:

$$f^{(n)}(x) = (1.10)^n x,$$

and this even works for the case where $n = 0$.

Although this example is so simple that it is easy to give a closed form for $f^{(n)}(x)$, we will use that example to show how iteration can be analyzed graphically. First, imagine the graph of $f(x) = (1.10)x$ versus x.



In the graph above, two lines are plotted that pass through the origin. The upper line corresponds to the function $y = f(x) = (1.10)x$ and the lower one, to the function $y = x$ which forms a $45°$ angle with the x and y axes. Now suppose we begin with an investment of $\$100$. To find out how much money we have after one year, find the 100 on the x-axis and go up to the line $f(x)$. The height of that line (which will be 110) represents how much we'll have after one year. But now we would like to put that 110 back into the function f, so we need to find 110 on the x-axis and look above that.

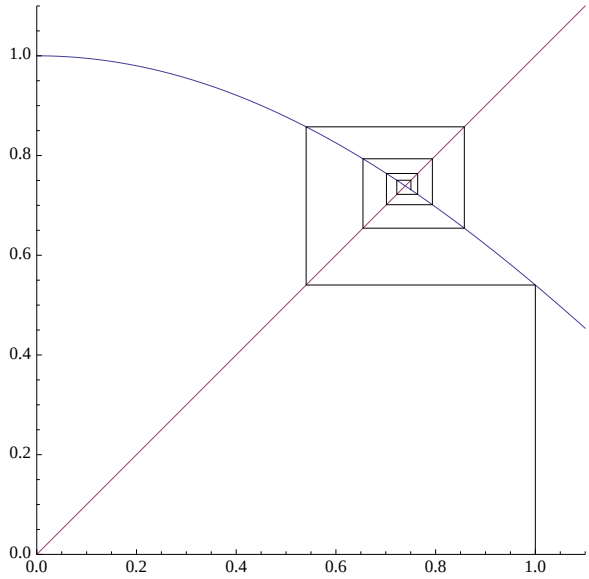
But what this amounts to is copying the y-value (the height of the line from the x-axis to the line $f(x)$ to the x-axis. Here is where the line $y = x$ suddenly becomes useful: If we begin at the point $(100, 110)$ and move horizontally to the line $y = x$, we will be situated exactly over the value 110 on the x-axis (since on the line $y = x$, the y-coordinate is the same as the x-coordinate). Since we're already exactly over 110, we can just go up from there to the line $f(x)$ to find the value of our investment after two years: $\$121$.

To find the value after three years, we can do the same thing: the height of the point $(110, 121)$ needs to be copied to the x-axis, so move right from that point to $(121, 121)$, putting us over 121 on the x-axis, and from there, we go up to $f(121)$ to find the amount after three years: $\$133.10$.

The same process can be used for each year, and the final height of the zig-zagging line at the upper-rightmost point represents the value of the original $\$100$ investment after six years: about $\$177.16$.

3 General Iteration

A little thought shows us that there is nothing special about the $f(x)$ we used in the previous section in a graphical interpretation. We can see what happens with the cosine function mentioned in the introduction, by using exactly the same sort of graph, except that $f(x) = \cos(x)$ and we will begin our iteration with $x = 1$:

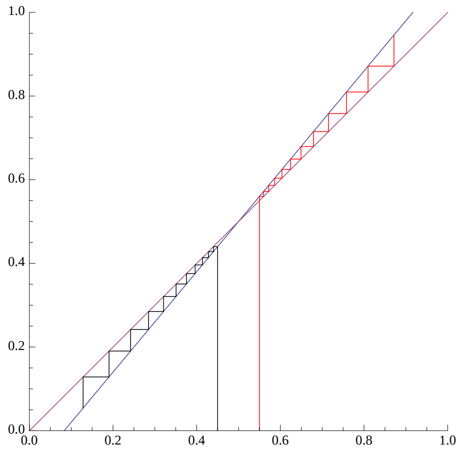


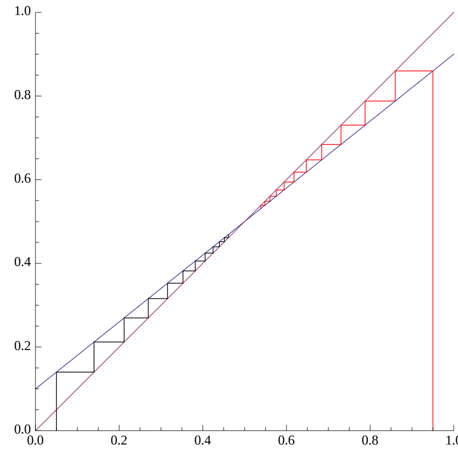
Although the picture looks different (instead of a zig-zagging line, we've got a spiral), exactly the same thing is going on. We begin with an input value of $x = 1$, we go up to the line $y = \cos(x)$ to find the result of one press of the cosine button. The height of that point has to be used as an input, so we move horizontally from it to the line $y = x$, at which point we're above a point x which is equal to the y-coordinate of the point $(1, \cos(1))$. Move vertically from there to the line $y = \cos(x)$, then horizontally to $y = x$, and repeat as many times as desired. It should be clear from the illustration

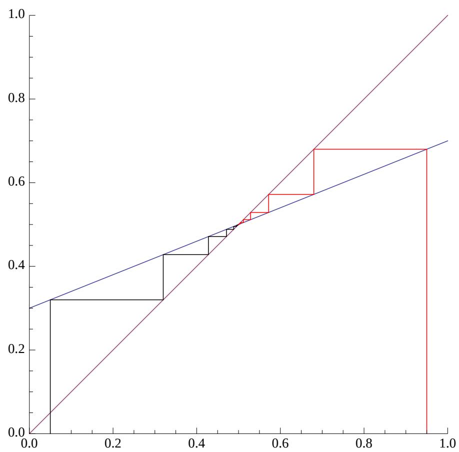
that as more and more iterations are performed, the spiral will converge to the point where the line $y = x$ meets the curve $y = \cos(x)$, and that point will have a value of approximately $(.739085133215, .739085133215)$.

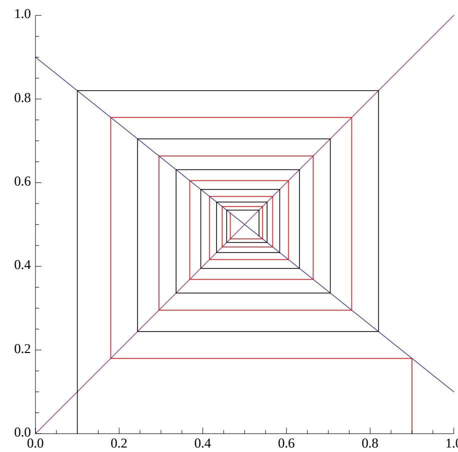
The major difference between the two examples is that it should be clear that in the case of bank interest, the zig-zag line will keep going to the right and up forever, the spiraling line for the cosine function will converge more and more closely to a limiting point. What we would like to do is examine the geometry of the curves $f(x)$ that either causes them to converge (like the cosine curve) or to diverge (like the bank interest function). We'd also like to know if there are other possibilities (other than convergence or divergence). We will examine graphically a number of examples in the next section.

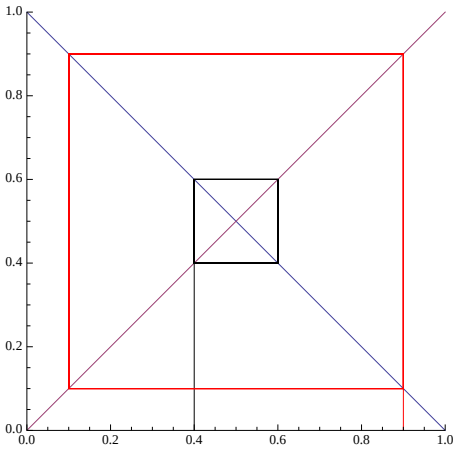
4 Graphical Examples

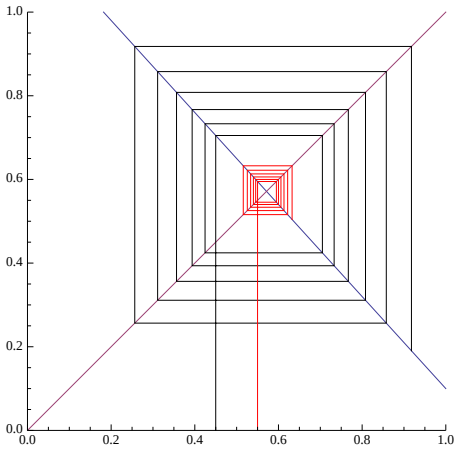












Every example on the previous page shows what happens for a particular function of x, assuming two different starting points. All of the functions on that page are linear (straight lines) and they illustrate the convergence (or divergence) properties. All the examples have the line $y = f(x)$ crossing the line $y = x$ at $(.5, .5)$. Basically, the only difference between the different functions f is the slope of that line. We call the point of intersection of $y = f(x)$ and $y = x$ a "fixed point", since if the input happened to have exactly that value, the output would be the same, or fixed.

In the example on the upper left we have basically the same sort of situation we had with the bank account balance except that in the bank balance case, the two lines met at $(0, 0)$. In this case, if we start a bit above the intersection, or fixed point, the values diverge, getting larger and larger. If we begin below the intersection, the values diverge, but to smaller and smaller values. The slope of $f(x)$ in this example is 1.2.

On the upper right, the slope of $f(x)$ is 0.8, and now the iterations converge to the fixed point from both directions.

The example on the middle left is similar, but the slope is less: only 0.4. The convergence is the same, from above or below, but notice how much faster it converges to the fixed point.

The next three examples show what happens when $f(x)$ has a negative slope. The example on the middle right, the slope is negative, but not too steep (in fact, it is -0.8). Convergence occurs from any starting point, but instead of zig-zagging to the fixed point, it spirals in.

The example on the bottom left shows a very special case where the slope of $f(x)$ is exactly -1. In this case, from any starting point there is neither convergence nor divergence; the output values fall into "orbits" of two values.

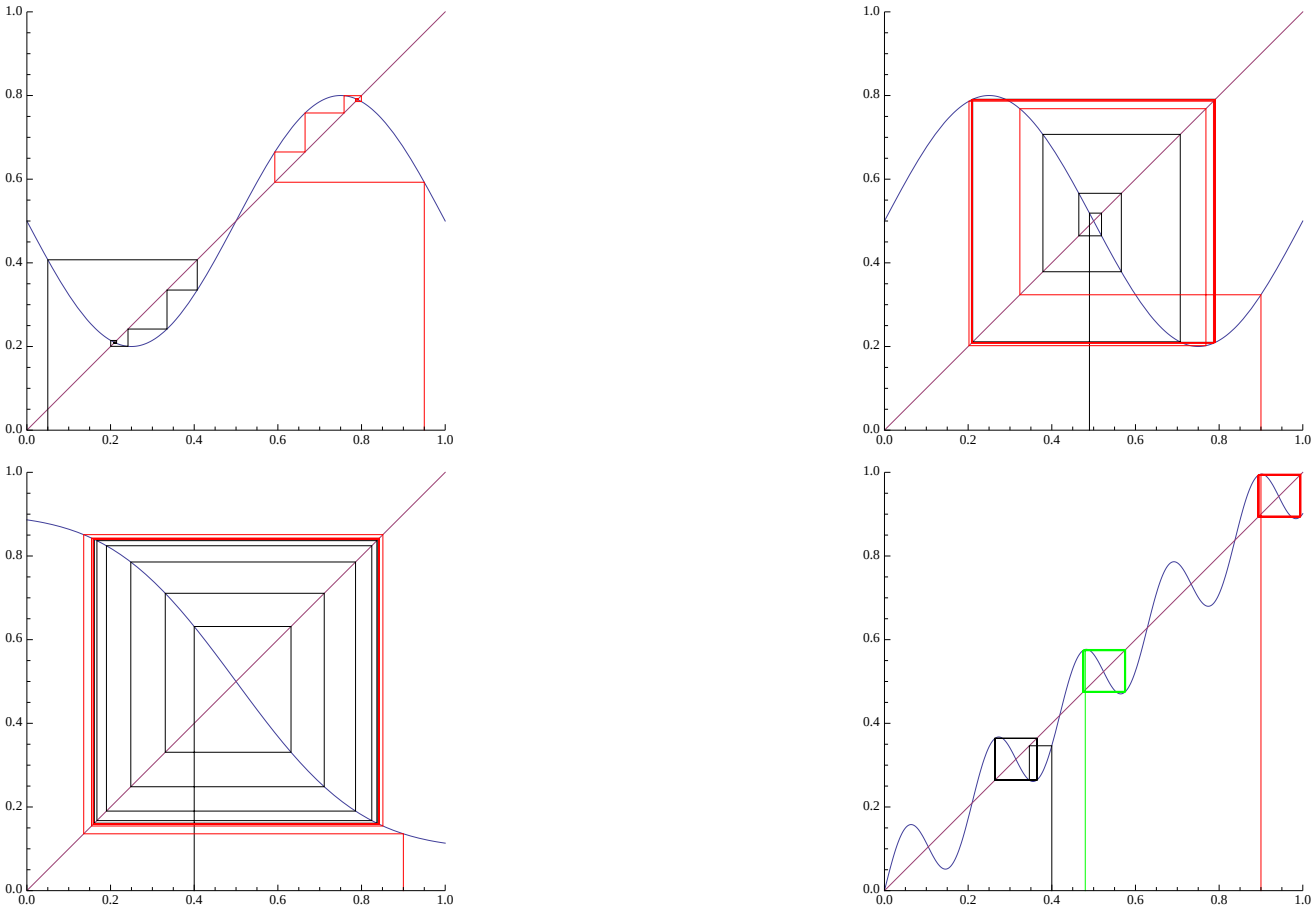
Finally, in the example on the lower right, the slope is negative and steeper than -1.0, and any input, except for the fixed point itself, will diverge in a spiral pattern, as shown.

The examples on the previous page pretty much cover all the possibilities for linear functions $f(x)$. If the slope, in absolute value, is less than 1, iterations converge to the fixed point. If the slope's absolute value is greater than 1, it diverges, and if the slope's value is equal to 1 or -1, there is a fixed orbit of one or two points. (Note that the function with slope 1 is $y = f(x) = x$ which is identical to the line $y = x$, so every point is a fixed point whose orbit contains exactly that point.)

It will turn out that if the function f is not a straight line, the resulting situation is usually not too different. Compare the example of the cosine function with the similar example on the middle right. In both cases the slopes of the (curve, line) are negative, but between -1 and 0.

In the next section we will examine more complex situations where the iteration results are not quite so simple.

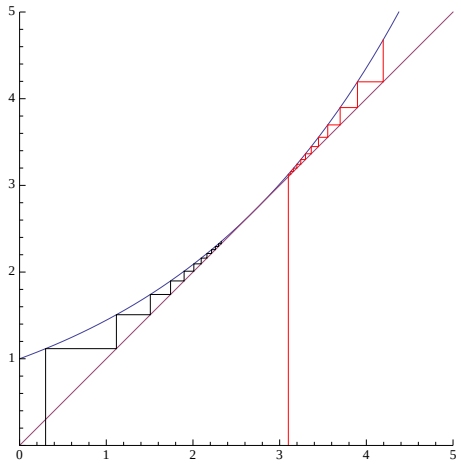
5 More Graphical Examples



Here are four more examples. The graph in the upper left shows that there may be more than one fixed point.

The other three illustrate limit cycles. In the examples in the upper right and lower left, different starting points converge not to a point, but to a cycle. On the upper right, the cycles first step inside the limiting cycle and then spiral out. In the lower right, one cycles in while the other cycles out.

Finally, the example in the lower right shows that there may be many limiting cycles, and the particular cycle into which a starting point falls depends on the position of the starting point.



The figure above shows another interesting example. The curve $f(x)$ is actually tangent to the line $x = y$, and initial values less than the point of tangency converge to that point, while initial values larger than it diverge.

You may find it interesting to try to construct examples with combinations of fixed points, cycles and sequences that diverge.

6 A Simple Practical Example

Suppose you want to approximate the square root of a number, for example $\sqrt{3}$. The following strategy works:

Make a guess, say x. If x is the correct answer, then $x^2 = 3$, and another way of saying that is: $x = 3/x$. If x is not the exact square root, then if x is too small, $3/x$ will be too large and vice-versa. Thus the two numbers x and $3/x$ will lie on both sides of the true square root, and thus their average has to be closer to the true square root than either one of them.

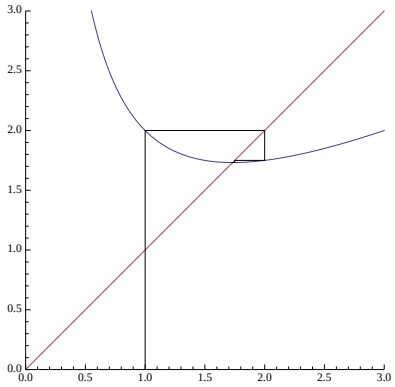
If x is not the exact square root, then the value $x_1 = (x+3/x)/2$ is closer to the square root than x was. We can then use x_1 as a new guess at the square root and repeat the procedure above as many times as desired to obtain more and more accurate approximations. The following illustrates this method to find $\sqrt{3}$, starting with an initial guess $x_0 = 1$, and approximated to 20 decimal places. The final line displays the actual result, again accurate to 20 decimal places:

$$\begin{aligned}
x_0 &= 1.00000000000000000000 \\
x_1 = (x_0 + 3/x_0)/2 &= 2.00000000000000000000 \\
x_2 = (x_1 + 3/x_1)/2 &= 1.75000000000000000000 \\
x_3 = (x_2 + 3/x_2)/2 &= 1.73214285714285714286 \\
x_4 = (x_3 + 3/x_3)/2 &= 1.73205081001472754050 \\
x_5 = (x_4 + 3/x_4)/2 &= 1.73205080756887729525 \\
\sqrt{3} &= 1.73205080756887729352
\end{aligned}$$

The fifth approximation is accurate to 17 decimal places. As you can see, this method converges rather rapidly to the desired result. In fact, although we will not prove it here, each iteration approximately doubles the number of decimal places of accuracy.

Obviously, what we are doing is iterating the function $f(x) = (x + 3/x)/2$ and approaching the fixed point which will be the square root of 3. Also, obviously, there is nothing special about 3. If we wanted to approximate the square root of n, where n is fixed, we simply need to iterate the function $f(x) = (x + n/x)/2$.

On the following page is a graphical illustration of the convergence of the iterates of $f(x)$ as described above, beginning with an initial guess of $x = 1$. After the first few steps, due to the flatness of the curve, you can see that the convergence is very rapid.



An interesting strategy that can be performed on almost any iterated function is this: we can double the speed of the convergence using a different function. In the example above, we approximated $\sqrt{3}$ by iterating $f(x) = (x^2 + 3)/(2x)$. Suppose we iterate $f(f(x))$? That should go twice as fast. In this case,

$$f(f(x)) = \frac{\left(\frac{x^2+3}{2x}\right)^2 + 3}{2\left(\frac{x^2+3}{2x}\right)}.$$

A little algebra yields:

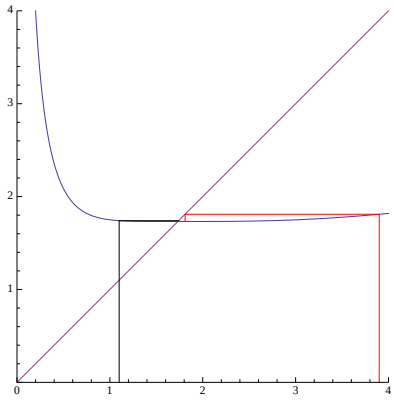
$$f(f(x)) = \frac{x^4 + 18x^2 + 9}{4(x^3 + 3x)}.$$

Iterations of that function are shown below, and it's obvious that the new function is a lot flatter, so convergence will be a lot faster.

Obviously, there is no reason we couldn't do the algebra for three, four or more iterations in the same way. The disadvantage, of course, is that the functions that we evaluate are more complicated, so the calculation time for each step may increase more rapidly than the time gained.

Here are three iterations to 30 decimal places of the function $f(f(x))$ calculated above, beginning with a (bad) initial guess of 4:

$$\begin{aligned} f(4) &= 1.819078947368421052631578947 37 \\ f(f(4)) &= 1.732052057147012130209726967 08 \\ f(f(f(4))) &= 1.732050807568877293527446400 16 \\ \sqrt{3} &= 1.732050807568877293527446341 51 \end{aligned}$$



7 A Simple Impractical Example

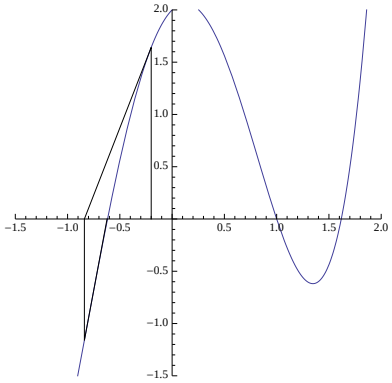
The following is not a particularly good way to do it, but it does use a simple iterated function to converge to the value of π. Unfortunatly, to do the calculations requires that we be able to calculate the sine function to many decimal places which is at least equally difficult as computing π using standard methods.

But here's the function to iterate: $p(x) = x + \sin(x)$. It's a good exercise to plot the function and see why a fixed point is at $x = \pi$. In any case, here are the results of repeatedly applying that function from $x_0 = 1$ (accurate to 50 decimal places):

$$
\begin{aligned}
x_0 &= 1.0000000000000000000000000000000000000000000000000000 \\
x_1 &= 1.8414709848078965066525023216302989996225630607984 \\
x_2 &= 2.8050617093497299136094750235092172055890412289217 \\
x_3 &= 3.1352763328997160003524035699574829475623914958979 \\
x_4 &= 3.1415926115906531496011613183596168589038475058108 \\
x_5 &= 3.1415926535897932384626310360379289800865645064638 \\
x_6 &= 3.1415926535897932384626433832795028841971693993751 \\
\pi &= 3.1415926535897932384626433832795028841971693993751
\end{aligned}
$$

8 Newton's Method

The method of approximating a square root used in the previous section is just a special case of a general method for finding the zeroes of functions.



Newton's method works as follows. (See the figure above.) Suppose that the curved line represents a function for which you would like to find the roots. In other words, you want to find where the function crosses the x-axis. First make a guess. In the figure, this initial guess is $x = -0.2$. Find the point on the curve corresponding to that guess (which will be $(x, f(x))$) and find the tangent line to the curve at that point. That tangent line will intersect the x-axis at some point, and since the tangent line is a rough approximation of the curve, it will probably intersect near where the curve does. In the example above, this will be at approximately $x = -0.84$.

We then iterate the process, going from that intersection point to the curve, finding the tangent to the curve at that point, and following it to where it intersects the x-axis again. In the figure above, the tangent line and the curve are so similar that it's almost impossible to see a difference between the curve and the line, so the intersection of that line and the x-axis is almost exactly the root of the function. But if not, the process can be iterated as many times as desired.

In the example above, we are trying to solve $f(x) = x^4 - 4x^2 + x + 2 = 0$. The initial guess is $x_0 = -0.2$, the second is $x_1 = -0.839252$, and the third is $x_2 = -0.622312$. To show that x_2 is very close to a root, we have: $f(-0.622312) = -0.0214204$.

To find the equation of the tangent line requires a bit of calculus, but assume for a moment that we can do that. In fact, for well-behaved functions, the number $f'(x)$ represents the slope of the function $f(x)$ at the point $(x, f(x))$.

If a line passes through the point (x_0, y_0) and has slope m at that point, the equation for that line is:

$$y - y_0 = m(x - x_0).$$

This is usually called the "point-slope" equation of a line.

In our case, the initial guess will be x_0, y_0 will be $f(x_0)$ and $m = f'(x_0)$. (Remember that $f'(x_0)$ is the slope of the function at $(x_0, f(x_0))$.) The equation for the tangent line will thus be:

$$y - f(x_0) = f'(x_0)(x - x_0).$$

This may look confusing, but there are only two variables above, x and y. All the others: x_0, $f(x_0)$ and $f'(x_0)$ are just numbers.

We want to find where this line crosses the x-axis, so set $y = 0$ and solve for x. We obtain:

$$x = x_0 - \frac{f(x_0)}{f'(x_0)},$$

and this equation is used to obtain successive approximations using Newton's method.

But as before, all we are doing to find better and better approximations for the root of a function is to iterate a particular function. The solution x in the equation above serves as the next guess for the root of $f(x) = 0$.

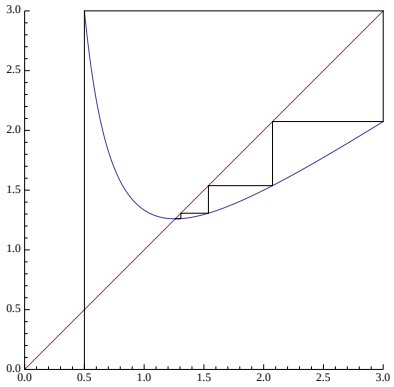
Let us show the graphical interpretation of the iterations performed to use Newton's method to find the cube root of a number; say, the cube root of 2.

In this case, we would like to solve the equation $f(x) = x^3 - 2 = 0$. If you don't know any calculus, take it on faith that the slope function (called the "derivative" in calculus) is $f'(x) = 3x^2$.

If we start with any guess x_0, the next guess will be given by:

$$x_0 - \frac{f(x_0)}{f'(x_0)} = x_0 - \frac{x_0^3 - 2}{3x_0^2} = \frac{2x_0^3 + 2}{3x_0^2}.$$

The figure at the top of the next page represents the iteration of the function above (beginning with a particularly terrible first guess of $x_0 = .5$) that approaches the cube root of 2, which is approximately 1.25992105.



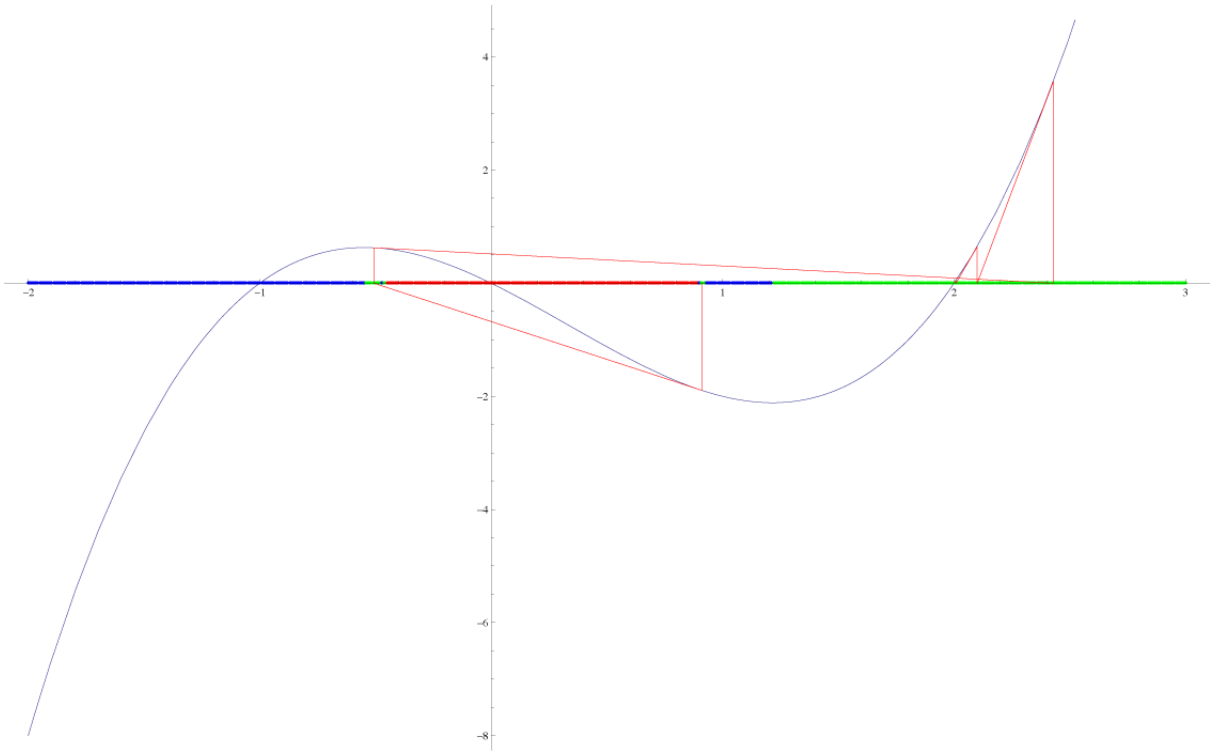
8.1 Domains of Attraction

When you use Newton's method to approximate the roots of a polynomial which has more than one root, the particular root to which the method converges depends on your choice of starting point. Generally, if you pick a first guess that is quite close to a root, Newton's method will converge to that root, but an interesting question is this: "Exactly which points on the line, if used as the initial guess, will converge to each root?"

Some points may not converge at all. For example, if your initial guess happened to lie exactly under a local maximum or minimum of the curve, the approximating tangent line would be parallel to the x-axis, and the iteration process would fail, since that line never intersects the x-axis. And there are other points where the process might fail. Suppose your initial guess produced a tangent line that intersected the x axis exactly under a local maximum or minimum. Or if the same thing happened after two iterations, et cetera?

And what happens if you pick a guess near where the function has a local maximum or minimum? The tangent line could intersect far from that initial guess, and in fact, into a region that converged to a root far away from the initial guess.

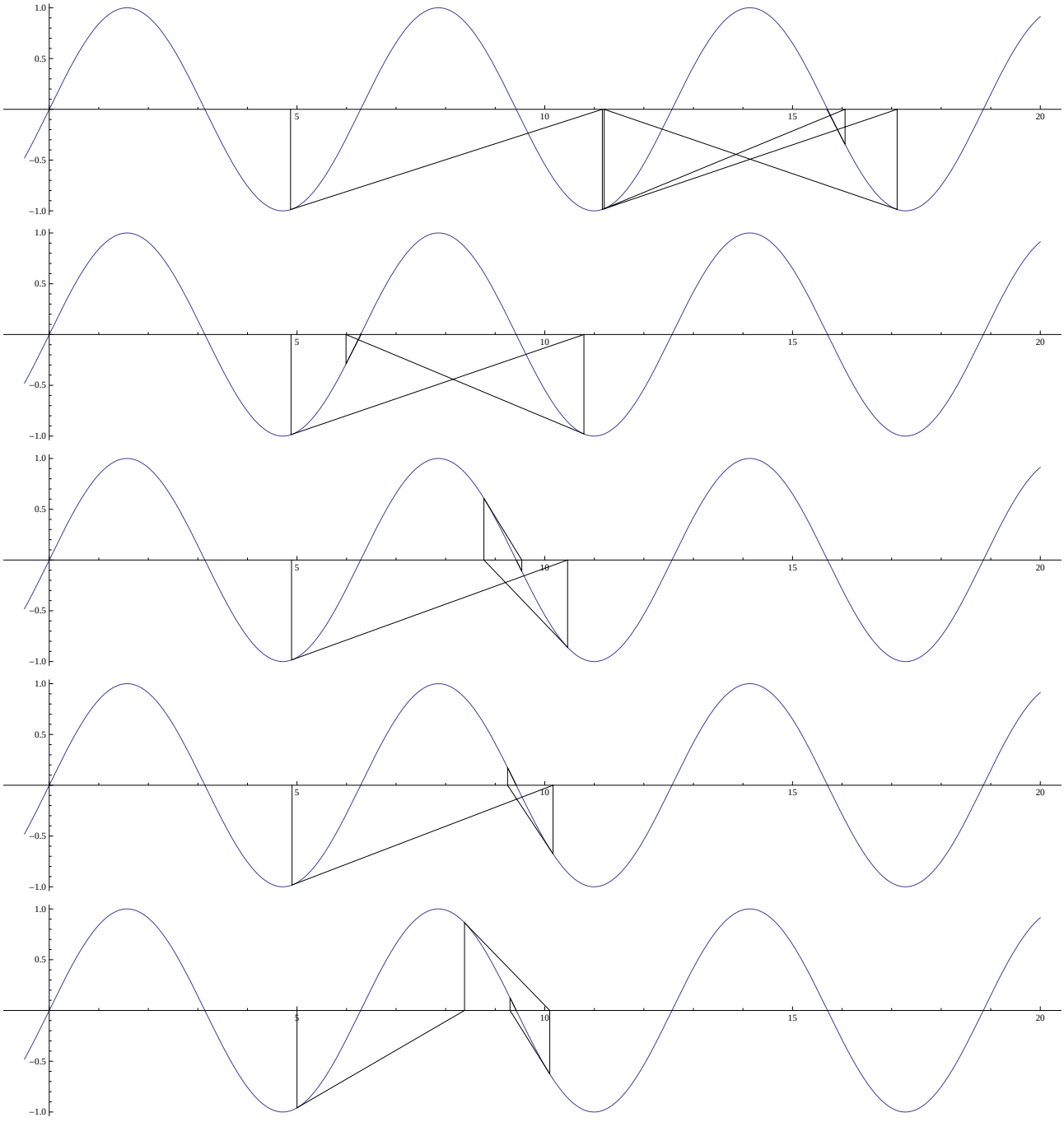
For any particular root, let us define the "domain of attraction" to be the set of initial guesses on the x-axis that will eventually converge to that root. As an illustration of some domains of attraction, consider the equation $f(x) = x(x+1)(x-2) = 0$ which obviously has the three roots $x = -1, x = 0$ and $x = 2$. We can still apply Newton's method to initial guesses, and most of them will converge to one of these three roots. In the figure below, points on the x-axis have been colored blue, red or green, depending on whether they would converge to the roots -1, 0 or 1, respectively.

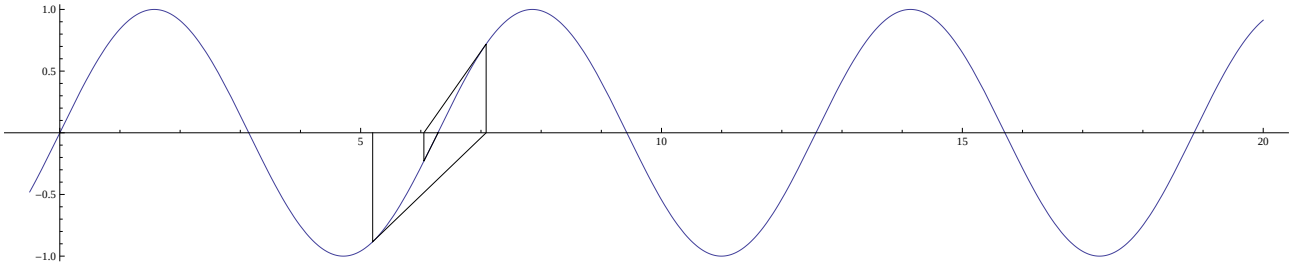


Notice that near the roots, the points are colored solidly, but as we move away from the roots to the regions near where the relative maxima an minima occur, the tangent lines will throw the next guess into regions that converge to non-nearby roots. An example of that is illustrated above. With a particular initial guess that is a bit less than 1.0, successive applications of Newton's method are illustrated by the red line which shows that although the initial guess is a bit closer to the root at zero, the first approximation takes it to a position closer to the root at -1, and then out beyond 2, after which it converges rapidly to the root at 2. The image is not infinitely detailed, and in fact, the actual shape of these domains of attraction can be quite complex. In fact, in Section 13 we will see some incredibly complex domains of attraction when Newton's method is extended to search for roots in the complex number plane.

8.2 Dependence on starting point

Here is a series of plots illustrating the use of Newton's method on the function $\sin(x)$, using starting values of 4.87, 4.88, 4.89, 4.90, 5.00 and 5.20, in that order. Notice how small changes in the input values can produce wildly-different convergence patterns. In every case below, six steps in the approximation are taken. Some converge so rapidly that only the first three steps or so are visible. Some are quite a bit slower.


1.0
0.5
−0.5
−1.0
5
10
15
20



9 Continued Fractions and Similar

Sometimes you see mathematical objects like these:

$$\begin{aligned} x &= \sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}} \\ x &= 1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cdots}}} \end{aligned}$$

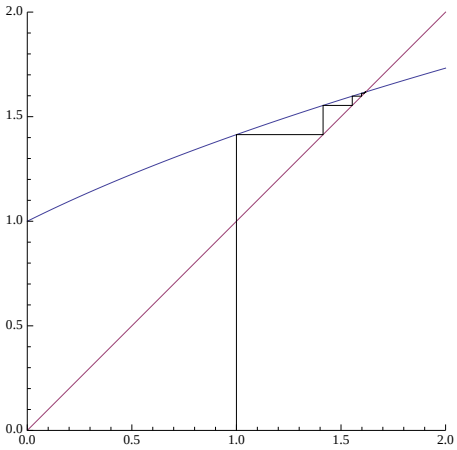
What do they mean? The usual interpretation is that each expression can be terminated after one, two, three, four, ... steps, and the terminated forms can be evaluated. If the evaluations tend to a limit, then the infinite expression is interpreted to be that limiting value.

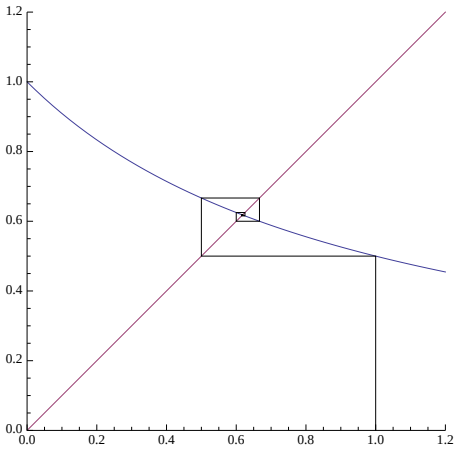
For example, in the first case above, we could evaluate:

$$\begin{aligned} \sqrt{1} &= 1.000000000 \\ \sqrt{1+\sqrt{1}} &= 1.414213562 \\ \sqrt{1+\sqrt{1+\sqrt{1}}} &= 1.553773974 \\ \sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1}}}} &= 1.598053182 \\ \cdots &= \cdots \end{aligned}$$

The second case (which is called a "continued fraction") can be evaluated similarly.

Both, however, can be evaluated using function iteration as well. In the first case, if we have the value x_0 for a certain level of iteration, the next level can be obtained by calculating $f(x_0)$, where $f(x) = \sqrt{1+x}$. For the continued fraction, the corresponding value of $f(x)$ would be $f(x) = 1/(1+x)$. Below are the graphical iterations. On the left is the nested square root and on the right, the continued fraction. Both begin with an initial approximation of 1:





Notice that there is another way to evaluate these expressions, assuming that the limits exist. In the first case, if we let:

$$x = \sqrt{1 + \sqrt{1 + \sqrt{1 + \sqrt{1 + \cdots}}}},$$

then a copy of x appears inside the expression for x and we have:

$$x = \sqrt{1 + x}.$$

Squaring both sides, we obtain $x^2 = 1 + x$, which has as a root the golden ratio:

$$x = \frac{1 + \sqrt{5}}{2} = 1.6180339887498948482045868343 7$$

Similarly, if x is set equal to the continued fraction, we can derive:

$$x = 1/(1 + x),$$

Which becomes $x^2 + x = 1$, and has the solution:

$$x = \frac{\sqrt{5} - 1}{2} = .6180339887498948482045868343 7,$$

one less than the golden ratio.

10 Optimal Stopping

Thanks to Kent Morrison, from whom I stole this idea.

Suppose you want to score as high as possible, on average, when you play the following game. The game goes for n rounds, and you know the value of n before you start. On each round, a random number uniformly distributed between 0 and 1 is selected. After you see the number, you can decide to end the game with that number as your score, or you can play another round. If play up to the n^{th} round, then you get whatever number you get on that last round. What is your optimal strategy?

The topic is called "optimal stopping" because you get to decide when to stop playing.

As with most games, it's usually a good idea to analyze simple cases first, and the simplest of all is the "game" when $n = 1$. It's not really a game, since once the number is selected, and since there are no more rounds, you will be stuck with that number. Since the numbers are uniformly distributed between 0 and 1, your average score, when $n = 1$, is $1/2$.

Let us give a name to the expected value of your score for a game with up to n rounds as E_n. From the previous paragraph, $E_1 = 1/2$.

What about if $n = 2$? What is E_2? On the first round a number is chosen, and based on that number, you can decide to use it as your score, or to discard it and play one more round. If your initial score is less than $1/2$, it's clear that you should play again, since on average playing the final round will yield, on average, a score of $1/2$. But if your initial score is larger than $1/2$, if you discard it, you'll do worse, on average.

So the optimal strategy for $n = 2$ is this: Look at the first score. If it's larger than $1/2$, stop with it. If not, discard it, and play the final round. What is your expected score in this game? Half the time you will stop immediately, and since you know your score is above $1/2$ it will be uniformly picked between $1/2$ and 1, or in other words, will average $3/4$. The other half of the time you will be reduced to playing the game with $n = 1$, which you already solved, and your expected score then will be $1/2$. So half the time you'll average $3/4$ and half the time you'll average $1/2$, yielding an expected value for $n = 2$ of:

$$E_2 = \frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{2} = \frac{5}{8}.$$

What is E_3? After the first round, you have a score. If you discard that score, you will be playing the game with only 2 rounds left and you know that your expected score will be $5/8$. Thus it seems obvious that if the first-round score is larger than $5/8$ you should stick with that, and otherwise, go ahead and play the $n = 2$ game since, on average, you'll get a score of $5/8$. Thus $5/8$ of the time you'll play the game with $n = 2$ and $3/8$ of the time you stop, with an average score midway between $5/8$ and 1, or $13/16$. Expected score will be:

$$E_3 = \frac{3}{8} \cdot \frac{13}{16} + \frac{5}{8} \cdot \frac{5}{8} = \frac{89}{128}.$$

The same sort of analysis makes sense at every stage. In the game with up to n rounds, look at the first round score, and if it's better than what you'd expect to get in a game with $n - 1$ rounds, stick with it; otherwise, play the game with $n - 1$ rounds.

Suppose we have laboriously worked out E_1, E_2, E_3, $\ldots E_{n-1}$ and we wish to calculate E_n. If the first score is larger than E_{n-1}, stick with it, but otherwise, play the game with $n - 1$ rounds. What will the average score be? Well, $1 - E_{n-1}$ of the time you'll get an average score mid-way between E_{n-1} and 1. The other E_{n-1} of the time, you'll get an average score of E_{n-1}.

The number mid-way between E_{n-1} and 1 is $(1 + E_{n-1})/2$, so the expected value of your score in the game with n rounds is:

$$E_n = (1 - E_{n-1}) \cdot \left(\frac{1 + E_{n-1}}{2}\right) + E_{n-1} \cdot E_{n-1} = \frac{1 + E_{n-1}^2}{2}.$$

Notice that the form does not really depend on n. To get the expected score for a game with one more round, the result is just $(1 + E^2)/2$, where E is the expected score for the next smaller game. We can check our work with the numbers we calculated for E_1, E_2 and E_3. We know that $E_1 = 1/2$, so

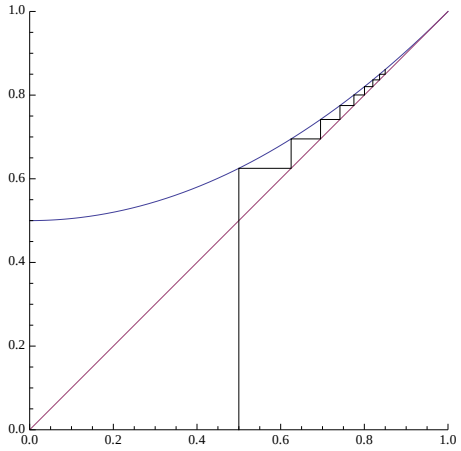
$$E_2 = \frac{1 + (1/2)^2}{2} = \frac{5}{8}, \text{ and } E_3 = \frac{1 + (5/8)^2}{2} = \frac{89}{128},$$

so we seem to have the right formula.

Notice that to obtain the next expected value, we simply take the previous one and plug it into the function $f(x) = (1 + x^2)/2$, so basically we just iterate to find successive expected values for successive games with larger n. Here are the first few values:

$$\begin{aligned} E_1 &= \frac{1}{2} = 0.50000 \\ E_2 &= \frac{5}{8} = 0.62500 \\ E_3 &= \frac{89}{128} \approx 0.695313 \\ E_4 &= \frac{24305}{32768} \approx .74173 \\ E_5 &= \frac{1664474849}{2147483648} \approx .775082 \\ E_6 &= \frac{7382162541380960705}{9223372036854775808} \approx .800376 \end{aligned}$$

As before, we can look at the graphical display of the iteration, and it's easy to see from that that the average scores increase gradually up to a limiting value of 1:



11 Biology: Heterozygote Advantage

Some knowledge of probability is required to understand this section. A little biology wouldn't hurt, either.

The canonical example of this is the presence of the gene that causes sickle-cell anemia in humans. Basically, if an individual has this disease he/she will not live long enough to have children. Normally, genes like this are eliminated from the population by natural selection, but in human populations in Africa, the gene is maintained in the population at a rate of about 11%. In this section we will see why.

The sickle-cell trait is carried on a single gene, and there are two types: A, the normal type, does not cause anemia. The sickle-cell gene is called a. Every individual has two copies of each gene (one from the mother and the other from the father), so there are three possibilities for the genotype of an individual: AA, Aa, or aa. Suppose that at some point, there is a proportion p of gene A in the population and a proportion $q = 1 - p$ of the gene a.

We will simply assume that in the next generation that the genes are thoroughly mixed, and therefore, at birth, the three types of individuals will appear with frequencies p^2 (of type AA), $2pq$ (of type Aa) and q^2 (of type aa).

But years later, when it is time for those individuals to breed, all of the ones of type aa are dead. In other words, genotype aa is a recessive lethal gene.

It may also be true that the individuals of types AA and Aa have different chances of survival. We don't know exactly what these are, but let us just say that individuals of type AA are $(1+s)$ times as likely to live as individuals of type Aa. As childbearing adults then, we will find $(1+s)p^2$ individuals of type AA for every $2pq$ individuals of type Aa.

We'd like to count the total number of a genes in the next generation. They can only come from the $2pq$ proportion having type Aa, and only half of the genes from those people will be a since the other half are of type A. Thus there will be a proportion pq of them. The total number of genes will be $2(1-s)p^2 + 2pq$.

The proportion of a genes after breeding will thus be:

$$q' = \frac{pq}{(1+s)p^2 + 2pq}.$$

But genes are either of type A or a, so $p = 1 - q$ and we have:

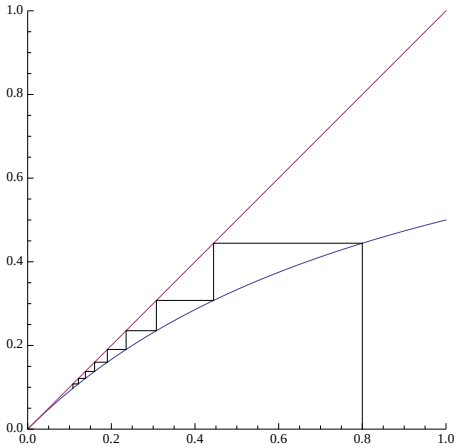
$$q' = \frac{(1-q)q}{(1+s)(1-q)^2 + 2q(1-q)} = \frac{q}{(1+s)(1-q) + 2q}.$$

To obtain the proportion of a genes in each successive generation, we simply have to put the value of q into the equation above to find the new value, q' in the next generation. To find the proportion after ten generations, just iterate 10 times. This is just another iterated function problem!

Let's look at a few situations. Suppose first that the AA individuals and the Aa individuals are equally fit. There is no disadvantage to having a single copy of the a gene. Then $s = 0$, and the function we need to iterate looks like this:

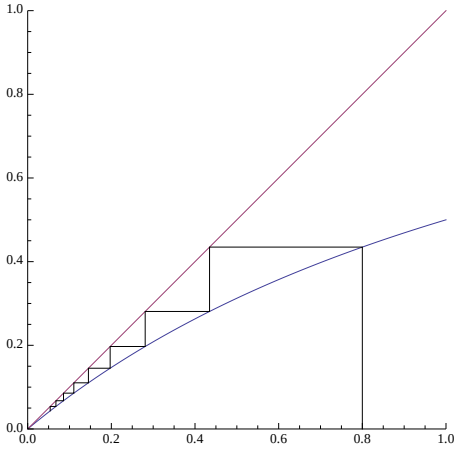
$$q' = f(q) = \frac{q}{1+q}.$$

Let's assume that for some reason there is a huge proportion of a genes, say 80%. Here is the graphical iteration in that case:



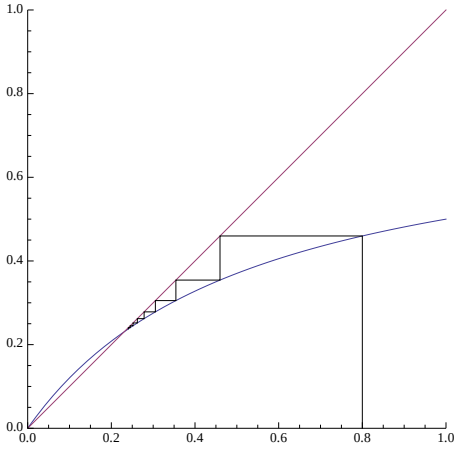
Notice that the fixed point is at zero, so with time, the sickle-cell gene should be eliminated from the population. In other words, the probability that an a gene will appear drops to zero.

Now suppose that even having a single copy of the a gene is a disadvantage. Suppose that it is 20% more likely that an AA individual will live to breeding age than an Aa individual. This makes $s = .2$, and the corresponding graphical iteration looks like this:



Not surprisingly, the sickle-cell gene is again driven out of the population, and if you compare the two graphs, you can see that it will be driven out more rapidly (fewer generations to reduce it the same amount) in this second case. With the same number of iterations, the gene is about half as frequent if the AA individuals have a 20% advantage over the aa individuals.

But in the real world, something else happens. In Africa where there is a lot of malaria, individuals with a single copy of the sickle-cell gene (individuals of type Aa) actually have an advantage over those of type AA because they have a better chance of surviving a case of malaria. We can use the same equation, but simply make s negative. Let's look at the graph with $s = -.3$:



Now notice that q tends to a non-zero limit, in this case, a bit more than 23%. In other words, the advantage to the individuals who have a single copy of the a gene is enough that the certain deaths of aa individuals is not enough to eliminate the gene. In fact, a value of $s = -.3$ is not required; any negative value for s would make this occur, although a smaller advantage of Aa would translate to a smaller limiting value of q.

As in every other case we've looked at, we could find the exact limiting value by setting $q' = q$:

$$q' = q = \frac{q}{(1 - .3)(1 - q) + 2q}.$$

If we solve that for q, we obtain: $q = .23077$: a bit more than 23%, as we noted above.

Individuals with two copies of the same gene (like aa or AA) are called homozygous, and individuals with one copy of each type are called heterozygotes. In the case of the sickle-cell gene in a malarial area, there is a heterozygote advantage; hence, the title of this section.

12 Markov Processes

To understand this section you will need to know something about matrix multiplication and probability.

Imagine a very simple game. You have a board that is nine squares long with a marker beginning on the first square. At each stage, you roll a single die, and advance the marker by the number of steps shown on the die, if that is possible. The game is over when you land on the ninth square. You cannot overshoot the last square, so, for example, if you are on square 5 and roll a 6 the marker stays in place because you cannot advance six spaces. In other words, in order to finish, you need to land on the last square exactly. With this game, we can ask questions like, "How likely is it that the game will have ended after 7 rolls of the die?"

Up to now we have been looking at one-dimensional iteration, but now we will look at a multi-dimensional situation. At every stage of the game, there are exactly 9 possible situations: you can be on square 1, square 2, ..., square 8. At the beginning of the game, before the first roll, you will be on square 1 with probability 1. After the game begins, however, all we can know is the probability of being on various squares.

For example, after one roll there is a $1/6$ chance of being on squares 2, 3, 4, 5, 6, and 7, and no chance of being on squares 1 or 8, et cetera.

We can, however, easily write down the probability of moving from square i to square j on a roll. For example, the chance of moving from square 1 to square 4 is $1/6$. The probability of moving from square 4 to itself is $4/6 = 2/3$, since only rolls of 1 or 2 will advance the marker. Rolls of 3, 4, 5 or 6 require impossible moves, beyond the end of the board. We can arrange all those probabilities in a matrix that looks like this:

$$M = \begin{pmatrix} 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 \\ 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 \\ 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{2}{3} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{5}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The number in row i and column j is the probability of moving from square i to square j in a single roll. We have put a 1 in row 8, column 8, since if the game is over, it stays over.

As we stated earlier, all we can know is the probability of getting to various squares after a certain number of rolls. At any stage there is a probability of being on square 1, 2, 3, ..., 9. We will write these in a row vector, and that vector, initially (at time zero), looks like this:

$$P_0 = (1, 0, 0, 0, 0, 0, 0, 0, 0).$$

In other words, we are certain to be on square 1.

The nice thing about the matrix formulation is that given a distribution of probabilities P of being on the 9 squares, if we multiply P by M (using matrix multiplication), the result will be a new P' that shows the odds of being on each square

after a roll of the die. To calculate the probabilities of being on the various squares after 10 rolls, just iterate the matrix multiplication 10 times.

To show how this works, let us call the probability distributions after one roll P_1, after two rolls P_2, and so on. Here are the numeric results:

$$\begin{aligned}
p_1 &= (0, 0.166667, 0.166667, 0.166667, 0.166667, 0.166667, 0.166667, 0, 0) \\
p_2 &= (0, 0, 0.027778, 0.083333, 0.138889, 0.194444, 0.25, 0.166667, 0.138889) \\
p_3 &= (0, 0, 0, 0.0185185, 0.0648148, 0.138889, 0.240741, 0.25463, 0.282407) \\
p_4 &= (0, 0, 0, 0.003086, 0.0246914, 0.0833333, 0.197531, 0.289352, 0.402006) \\
p_5 &= (0, 0, 0, 0.000514, 0.0087449, 0.0462963, 0.150206, 0.292567, 0.501672) \\
p_6 &= (0, 0, 0, 0.000086, 0.0030007, 0.0246914, 0.109396, 0.278099, 0.584727) \\
p_7 &= (0, 0, 0, 0.000014, 0.0010145, 0.0128601, 0.077560, 0.254612, 0.653939)
\end{aligned}$$

If we look at the last entry in p_7, we can conclude that after 7 rolls, there is a slightly better than 65% chance that the game will be over.

Note that this game is incredibly simple, but much more complicated situations can be modeled this way. For example, imagine a game where some of the squares are marked, "Go to jail", or "Go forward 4 steps". All that would be affected would be the numbers in the array M. For a board with 100 positions, the array would be 100×100, but the idea is basically the same.

Suppose you are designing a board game for very young children. You would like to make sure that the game is over in fewer than, say, 50 moves, so you could simply make an array corresponding to a proposed board, iterate as above, and make sure that the game is very likely to be over in 50 moves.

13 Final Beautiful Examples

13.1 Newton's Method for Complex Roots

In Section 8 we learned how Newton's method can be used to find the roots of functions as long as we are able to calculate the slopes of the functions at any point on the curve. In this final section, we are going to do the same thing, but instead of restricting ourselves to the real numbers, we are going to seek roots in the complex plane.

We will use Newton's method to find the roots of $x^3 = 1$ – in other words, we will find the cube root of 1. On the surface this seems silly, because isn't the cube root of 1 equal to 1? Well, it is, but it turns out that 1 has three different cube roots. Here's why:

$$x^3 - 1 = (x - 1)(x^2 + x + 1).$$

If we set that equal to zero, we get a root if either $x - 1 = 0$ or if $x^2 + x + 1 = 0$. The first equation yields the obvious root $x = 1$, but the second can be solved using the quadratic formula to obtain two additional roots:

$$x = \frac{-1 + \sqrt{3}i}{2} \quad \text{and} \quad \mathrm{x} = \frac{-1 - \sqrt{3}\mathrm{i}}{2},$$

where i is the imaginary $\sqrt{-1}$. If you haven't seen this before, it's a good exercise to cube both of those results above to verify that the result in both cases is 1.

If you plot the three roots, they lie on an equilateral triangle centered at the origin.

Since we will be working with complex numbers, we will change the variable names to be in terms of z instead of x. This is not required, but it is the usual convention, and it constantly reminds us that the variables are not necessarily restricted to be real numbers.

As we derived in Section 8, if z_0 is an initial guess for the cube root, the next approximation, z_1, can be obtained as follows:

$$z_1 = \frac{2z_0^3 + 1}{3z_0^2}.$$

(Actually, this is not exactly the same, since we are trying to find the cube root of 1 instead of 2 as we did in the previous section, but the only difference is that the constant 2 in that section is replaced by a 1 here.)

The other difference, of course, is that we want to allow z_0 and z_1 to take on complex values instead of just real ones. The calculations are a bit messy, but straightforward (see Appendix A for details). As an example, consider performing the iteration above when the initial value is $.9 + .4i$:

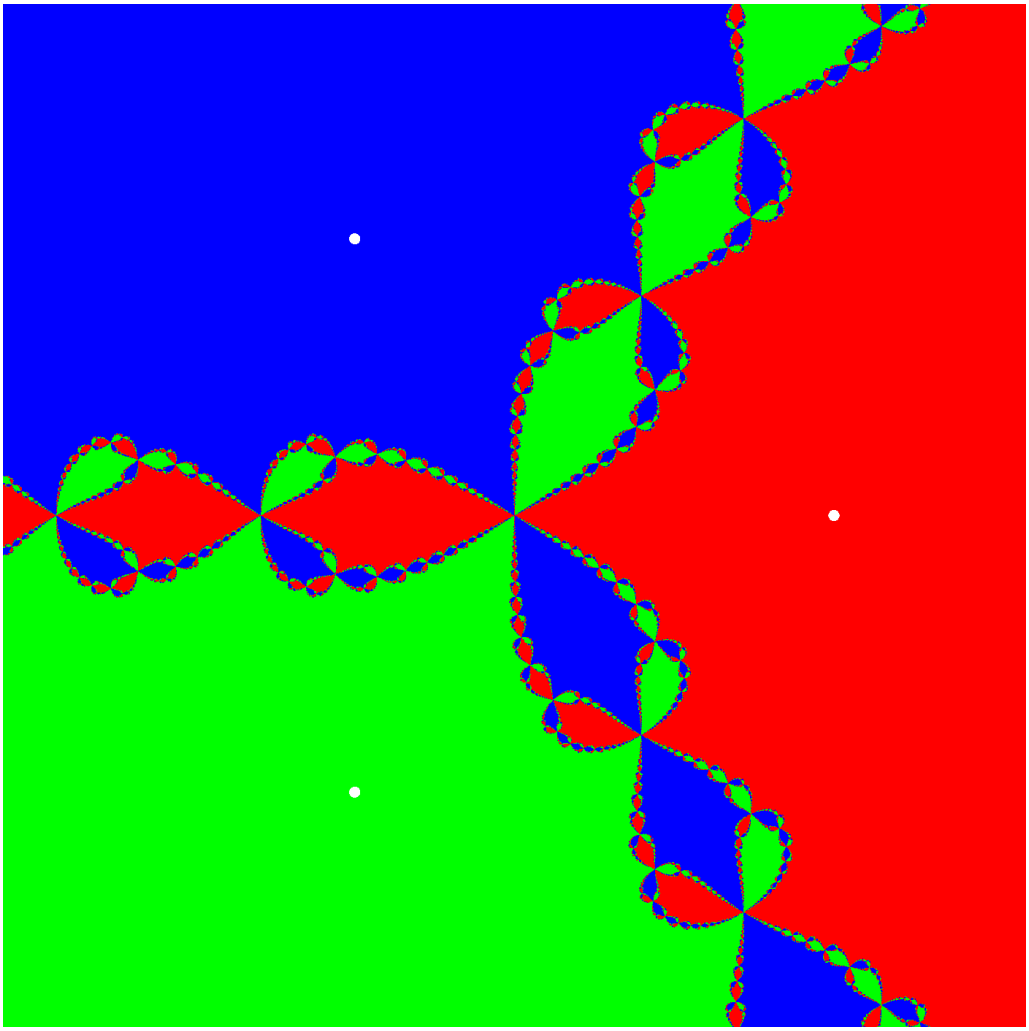
$$\begin{aligned} z_0 &= 0.9 + 0.4i \\ z_1 &= 0.830276 + 0.0115917i \\ z_2 &= 1.03678 - 0.00576866i \\ z_3 &= 1.00126 - 0.000395117i \\ z_4 &= 1.00000, -.0000009935i \end{aligned}$$

It is fairly clear that the sequence above converges to the root $z = 1$. Let's try one more, beginning at $-1 + i$:

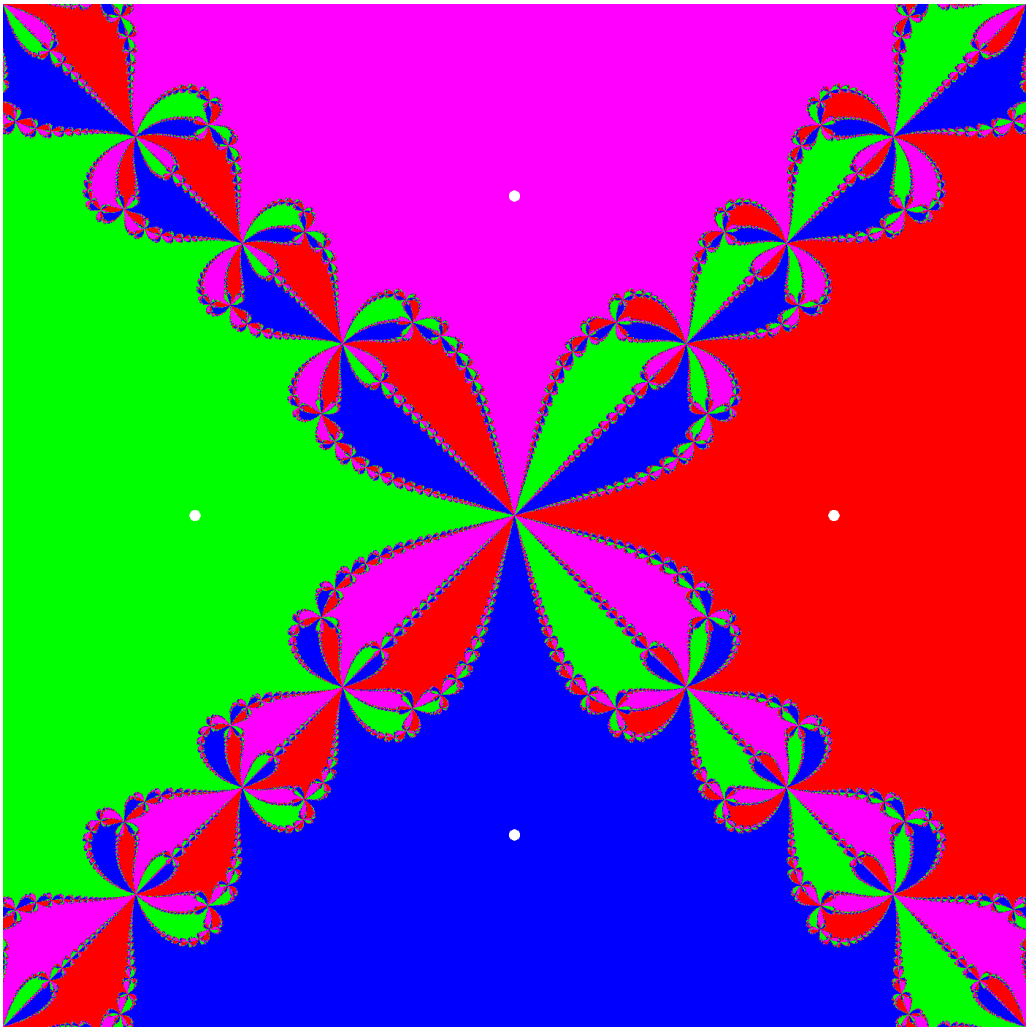
$$\begin{aligned} z_0 &= -1.0 + 1.0i \\ z_1 &= -0.666667 + 0.833333i \\ z_2 &= -0.508692 + 0.8411i \\ z_3 &= -0.49933 + 0.866269i \\ z_4 &= -0.5 + 0.866025i \end{aligned}$$

It is clear that this one converges to $z = (-1 + \sqrt{3}i)/2 = -.5 + .866025i$.

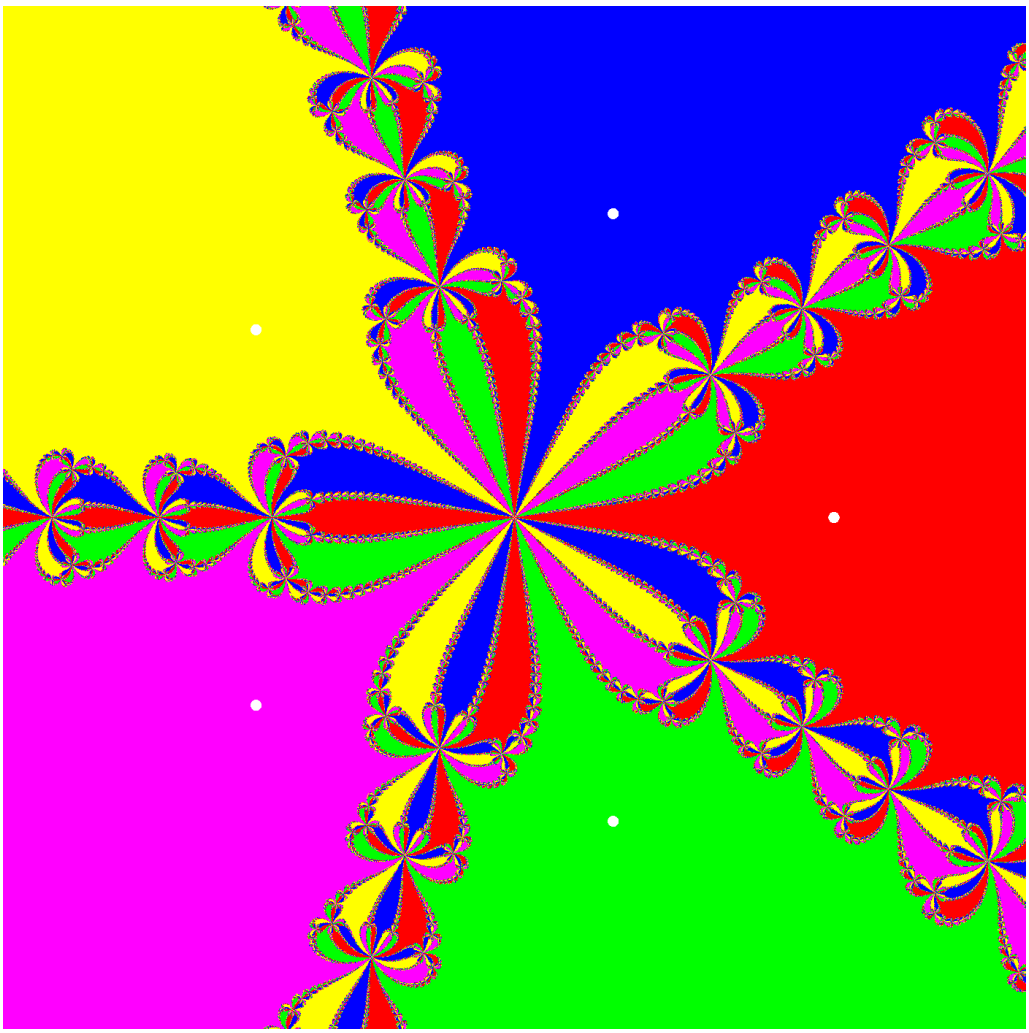
What makes Newton's method in the complex plane interesting is that many initial values lead to multiple interesting jumps around the plane before they converge to one of the roots. In fact, if we look at every point in the plane and color it red if it converges to 1, color it green if it converges to $(-1 - \sqrt{3})/2$ and blue if it converges to $(-1 + \sqrt{3})/2$, then the region around the origin ($-1.6 \leq x, y \leq 1.6$) would be colored as in the following illustration. These are the domains of attraction in the complex plane for Newton's method applied to this polynomial. The white dots in the red, green and blue regions represent the three roots of the equation $z^3 - 1 = 0$.



Just because it's another pretty picture, here is another image which corresponds in exactly the same way as the previous one to the use of Newton's method to solve the equation $z^4 - 1 = 0$. This time there are four roots: 1, -1, i and $-i$, and the colors correspond to regions of input values that will eventually converge to those roots.



Finally, here's exactly the same thing, but solutions of $z^5 - 1 = 0$:



13.2 The Mandelbrot Set

Newton's method iteratively applies a function to attempt to converge on a root, but other sorts of iteration can lead to different artistic results. One of the most famous is called the Mandelbrot set.

The definition of the Mandelbrot set is simple. Consider the function: $f(z) = z^2 + c$, where z and c are complex numbers. For any particular fixed c, we consider the iterates: $f^{(0)}(0)$, $f^{(1)}(0)$, $f^{(2)}(0)$, $f^{(3)}(0)$, ... In other words, iterated the function f beginning with the value 0. If the absolute value of z ever gets larger than 2, successive iterates of f from then on increase in size without bound. Another way to say that is that the iterates "diverge to infinity".

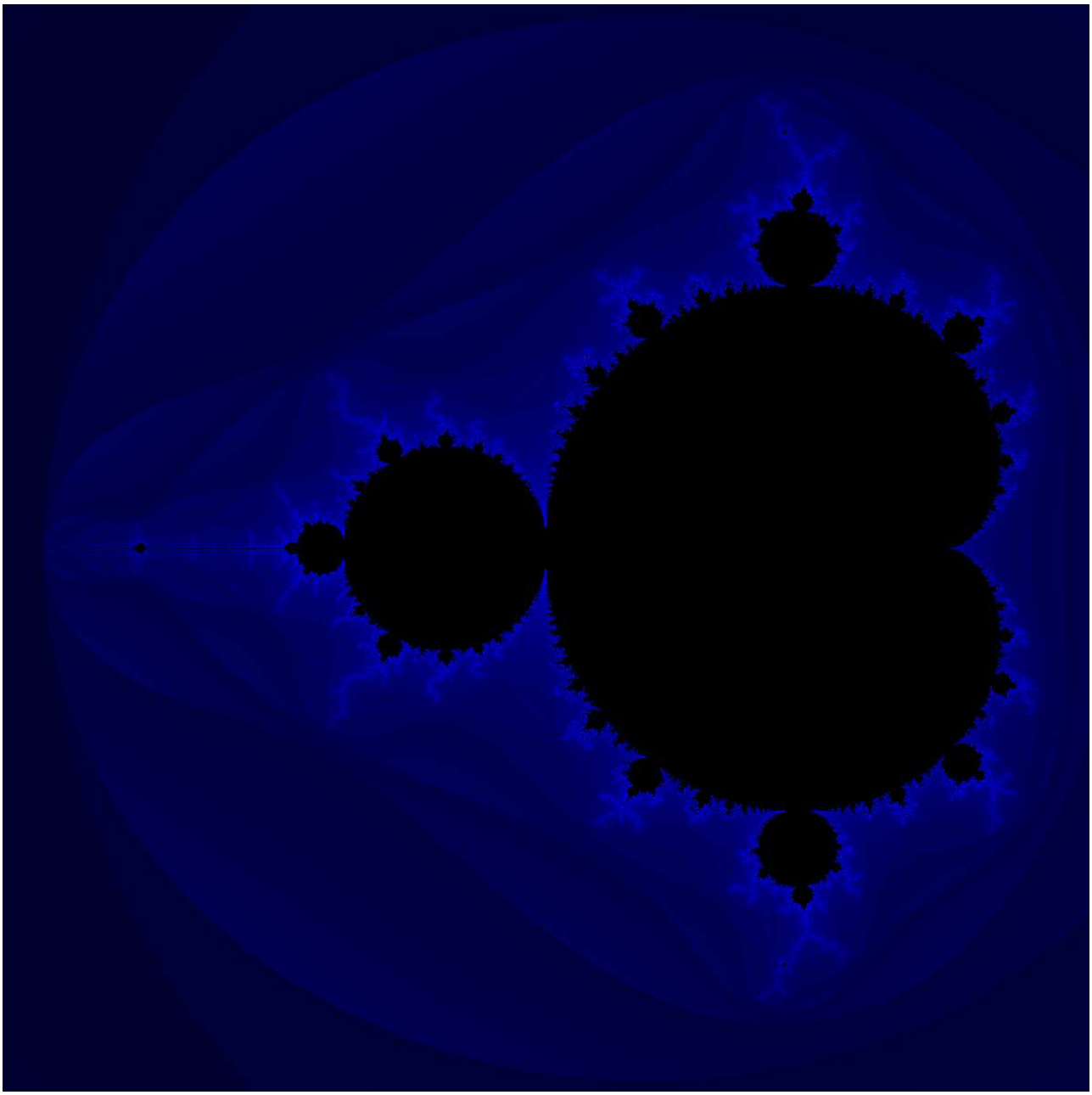
Thus the series of iterates will always diverge if $|c| > 2$, but for some values of c this divergence does not occur. An obvious example is $c = 0$, in which case $f^{(n)}(0) = 0$, but there are a lot of others. For example, if we let $c = 0.3 + 0.3i$, then the

first few iterations, beginning with zero, yield:

$$\begin{aligned} f^{(0)}(0) &= 0.00000 + 0.00000i \\ f^{(1)}(0) &= 0.30000 + 0.30000i \\ f^{(2)}(0) &= 0.30000 + 0.48000i \\ f^{(3)}(0) &= 0.15960 + 0.58800i \\ f^{(4)}(0) &= -0.02027 + 0.48769i \\ f^{(5)}(0) &= 0.06257 + 0.28023i \end{aligned}$$

The sequence above wanders around for a long time, but eventually converges to $0.143533 + 0.420796i$, which you can check to be a solution for $f(z) = z$, when $c = 0.3 + 0.3i$.

The Mandelbrot set is simply the set of complex numbers c for which the series of iterates of f, beginning at zero, do not diverge. The following is an illustration of the Mandelbrot set in the complex plane, where the members of the set itself are drawn in black. The somewhat artistic blue outline of the Mandelbrot set is an indication of how quickly those points diverge. The ones that diverge the slowest (and are, in some sense, closest to the boundary of the Mandelbrot set) are painted in the brightest color.



A Iterates for Newton's Method

To iterate the function representing Newton's method for the solution in the complex plane of the equation $z^3 - 1 = 0$, we must be able to compute:

$$z_1 = \frac{2z_0^3 + 1}{3z_0^2},$$

where z_0 is given.

We write $z_0 = x + yi$, where x is the real part of z_0 and y is the imaginary part. Then we have:

$$\begin{aligned} z_1 &= \frac{2(x+yi)^3 + 1}{3(x+yi)^2} \\ z_1 &= \frac{2(x^3 + 3x^2yi - 3xy^2 - 3y^3i) + 1}{3(x^2 + 2xyi - y^2)} \\ z_1 &= \frac{(2x^3 - 6xy^2 + 1) + (6x^2y - 2y^3)i}{(3x^2 - 3y^2) + (6xy)i} \end{aligned}$$

The fraction above has the form:

$$z_1 = \frac{a + bi}{c + di},$$

where $a = 2x^3 - 6xy^2 + 1$, $b = 6x^2y - 2y^3$, $c = 3x^2 - 3y^2$ and $d = 6xy$. We have:

$$\begin{aligned} z_1 = \frac{a+bi}{c+di} &= \frac{(a+bi)(c-di)}{(c+di)(c-di)} \\ z_1 &= \frac{(ac+bd) + (bc-ad)i}{(c^2+d^2)} \\ z_1 &= \Big(\frac{ac+bd}{c^2+d^2}\Big) + \Big(\frac{bc-ad}{c^2+d^2}\Big)i \end{aligned}$$

and the final line shows us how to calculate the real and imaginary parts of z_1 in terms of the real and imaginary parts of z_0.

Let's illustrate this with $z_0 = -1 + i$ which was an example in Section 13. We have $x = -1$ and $y = 1$. We then have:

$$z_1 = \frac{(-2 + 6 + 1) + (6 - 2)i}{(3 - 3) - 6i}.$$

So $a = 5$, $b = 4$, $c = 0$ and $d = -6$, yielding:

$$z_1 = \frac{-24}{36} + \frac{30}{36}i = -\frac{2}{3} + \frac{5}{6}i,$$

which is what we obtained earlier.