The basic idea behind the classic *generating function* is easy to explain; it is a trick to turn an infinite sequence into a function. Classic mathematics simply knows a lot more about functions than it does about infinite sequences. Sometimes sequences can be bounded from above or below and in this way proven to converge or not. A few sequences even have known limits. However, so much more can be accomplished when we know how to change arbitrary sequences into functions; specifically, recursions involving sequence elements become algebraic equations when using generating functions.

Given a sequence $s_0, s_1, s_2, \ldots$, its generating function is defined to be

$$s(x) \equiv \sum_{n=0}^{\infty} s_n x^n \tag{1}$$

basically an infinite polynomial in x. The variable x itself is entirely artificial, being introduced solely for the purpose of giving the generating function a domain. It is easily seen that the correspondence between a sequence and its generating function is one-to-one; different sequences correspond to different generating functions, and different generating functions generate different sequences. In a way, generating sequences are the opposite of Taylor expansions. A Taylor expansion takes a function $s(x)$ and creates a sequence of coefficients s_n, while the generating function does just the opposite. The Taylor coefficients give us intuition as to the behavior of the function, while the generating function gives us insight as to the behavior of the sequence.

We can demonstrate the strength of the generating function technique with a simple example, that of the *Fibonacci sequence* f_n. This famous sequence, invented by Leonardo of Pisa (nicknamed Fibonacci) in 1202, models the number of female rabbits in successive years. We assume that each mature female rabbit produces a female offspring each year and that no rabbit ever dies. We start with a single female rabbit ($f_0 = 1$); there is still only that rabbit after one year ($f_1 = 1$), since it takes a year for the rabbit to reach maturity. In the second year a new baby rabbit is born ($f_2 = 2$), and another in the third ($f_3 = 3$). Thereafter in each year we have the number of rabbits alive in the previous year *plus* those born to rabbits who were alive two years ago. We can deduce the recursive definition

$$f_0 = 1 \qquad f_1 = 1 \qquad f_n = f_{n-1} + f_{n-2} \qquad \text{for}\, n \geq 2 \tag{2}$$

that produces the values $1, 1, 2, 3, 5, 8, 13, 21, \ldots$. However, were we to need f_{137} we would have no recourse other than to recurse 137 times. Is there an explicit (nonrecursive) formula for f_n? It's hard to think of any way of finding one, but that is where the generating function can help. Generating functions convert complex recursions into simple algebraic equations that can often be solved.

The generating function for the Fibonacci sequence is

$$f(x) = \sum_{n=0}^{\infty} f_n x^n = 1 + x + 2x^2 + 3x^3 + 5x^4 + 8x^5 + \dots$$

and this is what we wish to evaluate. To proceed, take the recursion that defines the Fibonacci sequence

$$f_n = f_{n-1} + f_{n-2}$$

multiply both sides by x^n and sum from $n = 2$ to infinity (from 2 since f_{n-2} is only defined from there!).

$$\sum_{n=2}^{\infty} f_n x^n = \sum_{n=2}^{\infty} f_{n-1} x^n + \sum_{n=2}^{\infty} f_{n-2} x^n$$

Now we can manipulate the right hand side to try to change the bottom index in the sums.

$$\begin{aligned} \sum_{n=2}^{\infty} f_n x^n &= x \sum_{n=2}^{\infty} f_{n-1} x^{n-1} + x^2 \sum_{n=2}^{\infty} f_{n-2} x^{n-2} \\ &= x \sum_{n=1}^{\infty} f_n x^n + x^2 \sum_{n=0}^{\infty} f_n x^n \end{aligned}$$

Note carefully what we did with the indexes - we will do this a lot of later. In the first sum we had $f_{n-1}x^{n-1}$; define $m = n - 1$ in order to make this $f_m x^m$; fix the sum to run from $m = 1$ since when $n = 2$ we have $m = n-1 = 1$ (the top index is infinity so changing by one doesn't do anything); but m is a dummy index, so we can rename it n. Similarly in the second sum we have $f_{n-2}x^{n-2}$, so we define $m = n - 2$ to change this to $f_m x^m$; the sum now runs from $m = 0$, and once again we rename the dummy index from m to n.

The second sum on the right hand sum is already a generating function. The sum on the left is a generating function without the first two terms $f_0 x^0$ and $f_1 x^1$, while the first sum on the right is a generating function without its first term.

$$f(x) - f_0 x^0 - f_1 x^1 = x \left(f(x) - f_0 x^0 \right) + x^2 f(x)$$

Now let's rearrange.

$$f(x) - 1 - x = f(x)x - x + f(x)x^2$$

Solving the algebraic equation we easily find an explicit expression for the generating function.

$$f(x) = \frac{1}{1 - x - x^2}$$

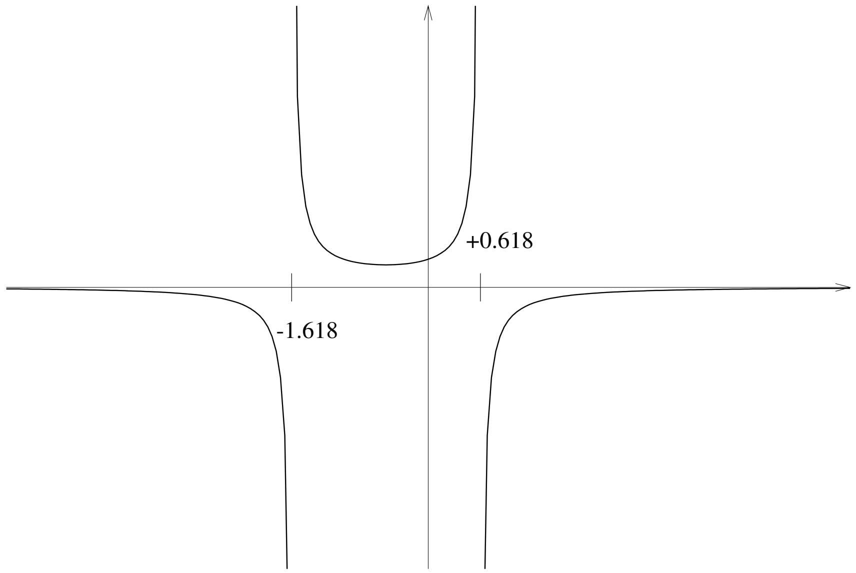

Figure 1: The generating function for the Fibonacci sequence. Note the divergences at $-\gamma \approx -1.618$ and $-\gamma' \approx 0.618$.

Using standard methods for plotting functions (that you learned in high school) we can plot the generating function of the Fibonacci series.

The zeros of the quadratic in the denominator are easily found to be $-\gamma$ and $-\gamma'$ where $\gamma \equiv \frac{1+\sqrt{5}}{2} = \cos^{-1}(\frac{\pi}{5})$ is the famous 'golden ratio' and $\gamma' =\equiv \frac{1-\sqrt{5}}{2} = -\frac{1}{\gamma}$.

We can now return to our original problem. In order to find an explicit formula for the n^{th} Fibonacci element, we need only to rewrite the generating function as an infinite polynomial and pick out the coefficients. To do this we use a 'partial fraction expansion'

$$f(x) = \frac{1}{(x+\gamma)(x+\gamma')} = \frac{1}{a-b}\left(\frac{a}{1-ax} - \frac{b}{1-bx}\right)$$

where $a + b = -ab = 1$. Utilizing the formula for the sum of a geometric progression $\frac{1}{1-ax} = \sum_{n=0}^{\infty}(ax)^n$ and comparing term by term, we find

$$f_n = \frac{1}{\sqrt{5}}\left(\gamma^{n+1} - (\gamma')^{n+1}\right) \tag{3}$$

the desired explicit formula for the n^{th} Fibonacci element.

Most people when seeing this formula for the first time are amazed that this combination of irrational numbers yields an integer at all. When that impression wears off, a feeling of being tricked sets in. The two irrational numbers in the numerator contain exactly a factor of $\sqrt{5}$, which is exactly what is being eliminated by the denominator; but if it is all a trick why can't a formula without a $\sqrt{5}$ be devised? So we are now surprised by our prior lack of surprise! The explicit equation is *so* astounding that you are strongly encouraged to run to a computer and try it out. Please remember to round the result to the nearest integer in order to compensate for finite precision calculations.

Now that we have become convinced of the great utility of generating functions, we will slightly adapt them for use in DSP. The z-transform is conventionally defined as:

$$S(z) = \mathrm{zT}(s_n) = \sum_{n=-\infty}^{\infty} s_n z^{-n} \tag{4}$$

and you probably noticed two modifications but in fact there are three!

First, since in DSP digital signals in the time representation extend to infinity in both directions (unlike infinite sequences in math where the index runs from zero to infinity), we needed to make the sum run from minus infinity rather than from zero. Second, the DSP convention is to use negative powers z^{-n} rather than x^n; but since the sum runs over both negative and positive indexes this is merely a convention. (Using z instead of z^{-1} is equivalent to interchanging s_n with s_{-n}.) However, there is a third difference, implied by our calling the variable z rather than x. We will allow z to be a complex variable rather than merely a real one.

This is a major difference. While the generating function enabled the study of sequences using the tools of real analysis, the definition of the zT over the complex plane makes available even more powerful tools of complex functions.

Unlike the generating function we saw above, which is defined over the real axis, $S(z)$ is defined over the complex plane, called the z-plane. In this complex plane every point z represents a digital signal $s_n = z^n$. Any complex variable z can be written in polar form

$$z = re^{\mathrm{i}\omega}$$

where r is the magnitude, and ω the angle. So, the signal represented by z can be thought of as $s_n = r^n e^{\mathrm{i}\omega n}$.

Complex sinusoids correspond to $z = e^{\mathrm{i}\omega t}$ on the unit circle, since there $r = 1$ and so $s_n = e^{\mathrm{i}\omega n}$. For zero angle $z = 1$ and so $s_n = 1^n = 1$ the signal is constant (DC). For 180 degrees $z = -1$ and so $s_n = (-1)^n$ the Nyquist signal. When $z = re^{\mathrm{i}\omega}$ lies inside the unit circle $r < 1$ and we obtain a *decaying* exponential $s_n = r^n e^{\mathrm{i}\omega n}$. Growing exponentials correspond to z outside the unit circle. Other signals can be written as the combination of several points in the complex plane.

What happens if we calculate the zT only on the unit circle in the z-plane?

Evaluate the zT only for $z = e^{\mathrm{i}\omega}$ as a function of the angle ω, we find

$$s(\omega) = S(z)\big|_{z=e^{\mathrm{i}\omega}} = \sum_{n=-\infty}^{\infty} s_n z^{-n} = \sum_{n=-\infty}^{\infty} s_n e^{-\mathrm{i}\omega n} \tag{5}$$

which is precisely the DFT. So, the zT reduces to the DFT if evaluated on the unit circle. Put in another way, the zT is an extension of the DFT to outside the unit circle.

> For other nonunity magnitudes we can always write $r = e^{\lambda}$ so that $z = e^{\lambda + \mathrm{i}\omega}$ and
>
> $$S(z) = \sum_{n=-\infty}^{\infty} s_n z^{-n} = \sum_{n=-\infty}^{\infty} s_n e^{-(\lambda+\mathrm{i}\omega)n} \tag{6}$$
>
> which is a digital version of the **L**aplace **T**ransform (LT). we won't need the Laplace transform, since it is useful for analog signals. The LT expands analog signals in terms of exponentially increasing or damped sinusoids. Its expression is
>
> $$f(s) = \int_{-\infty}^{\infty} f(t) e^{-st} dt \tag{7}$$
>
> where s is understood to be complex (defining the s-plane). Sinusoids correspond to purely imaginary s, decaying exponentials to positive real s, growing exponentials to negative real s. Just as the zT extends the DFT, the LT generalizes the FT, since the FT is simply the LT along the imaginary s axis.

Note that the terminology *z transform* is embarrassing. The FT and the DFT are true transforms, since they transform functions into functions and sequences into sequences. The Fourier Series is not a transform since it maps periodic analog signals into sequences. Similarly the zT inputs a sequence and returns a complex function. This change of form from sequence to function should disqualify the zT from being called a *transform*, but for some reason doesn't. But don't worry - outside DSP the term 'z transform' is entirely unknown. Also, the name z is simply used as a common way of implying that a variable is complex (rather than real). It's not that Prof. Z discovered this transform just like Fourier discovered his!

The most important property of the zT concerns time shifts. Assume that we know the zT of a signal s, what can we say about the zT of the delayed signal $\hat{z}^{-1}s$? (Remember that $\hat{z}^{-1}$ is the *delay* operator, that is, if $y = \hat{z}^{-1}x$ then $y_n = x_{n-1}$.) It turns out that there is a simple relationship (which explains why we called the delay operator $\hat{z}^{-1}$ in the first place!).

We start by substituting s_{n-1} instead of s_n in the formula for the zT.

$$\mathrm{zT}(\hat{z}^{-1}s) = \sum_{n=-\infty}^{\infty} s_{n-1}\, z^{-n}$$

Now we do our regular trick of playing with indexes, but this time we don't have to move the sum indexes since both are infinity!

$$\mathrm{zT}(\hat{z}^{-1}s) = z^{-1} \sum_{n=-\infty}^{\infty} s_n z^{-n}$$

We immediately recognize the zT on the right hand side and conclude

$$\mathrm{zT}(\hat{z}^{-1}s) \ = \ z^{-1}\,\mathrm{zT}(s) \ .$$

Be careful here. The first z in this equation is part of zT meaning - take the z transform. The second $\hat{z}^{-1}$ is the *delay operator*. The third z^{-1} is a complex number that simply multiplies the value of the zT of s. Remember that the zT is defined for every point z in the complex plane. The value of the zT of the delayed signal at the point z in the complex plane is $S(z)$ times z^{-1}.

Accordingly the factor of z^{-1} can be thought of as a unit delay *operator*, as indeed we originally defined it. The origin of the symbol that was arbitrary then is now understood; delaying the signal by one digital unit of time can be accomplished by multiplying it by z^{-1} in the z domain. This interpretation is the basis for much of the use of the zT in DSP.

For example, consider a radioactive material with half-life τ years. At the beginning of an experiment $n = 0$ we have 1 unit of mass $m = 1$ of this material; after one half-life $n = 1$ the mass has dropped to $m = \frac{1}{2}$ units, $\frac{1}{2}$ having been lost. At digital time $n = 2$ its mass has further dropped to $m = \frac{1}{4}$ after losing a further $\frac{1}{4}$, etc. After an infinite wait

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \ldots = 1$$

all of the material has been lost (actually converted into another material). The mass left as a function of time measured in half-lives is

$$m_n = \frac{1}{2}^n$$

an exponentially decreasing signal. Now a scientist measures the amount of mass at some unknown time n and wishes to predict (or is it postdict?) what the mass was one half-life back in time. All that need be done is to double the amount of mass measured, which is to use the operator z^{-1} with z being identified as $\frac{1}{2}$. This example might seem a bit contrived, but we shall see later that many systems when left alone tend to decrease exponentially in just this manner.

We leave as exercises what the zT does to the *time advanced* signal $\hat{z}s$ (that's an easy question), and what it does to the *time reversed* signal Rev$s$.

We have been ignoring a question that always must be raised for infinite series. Does the zT *converge*? When there are only a finite number of terms in a series there is no problem with performing the summation, but with an infinite number of terms the terms must decay fast enough with n for the sum not to explode. For complex numbers with large magnitudes the terms will get larger and larger with n, and the whole sum becomes meaningless.

By now you may have become so accustomed to infinities that you may not realize the severity of this problem. The problem with divergent infinite series is that the very idea of adding terms may be called into question. We can see that unconvergent sums can be meaningless by studying the following enigma that purports to prove that $\infty = -1$! Define

$$S = 1 + 2 + 4 + 8 + \ldots$$

so that S is obviously infinite. By pulling out a factor of 2 we get

$$S = 1 + 2(1 + 2 + 4 + 8 + \ldots)$$

and we see that the expression in the parentheses is exactly S. This implies that $S = 1 + 2S$, which can be solved to give $S = -1$. The problem here is that the infinite sum in the parentheses is meaningless, and in particular one cannot rely on normal arithmetical laws (such as $2(a+b) = 2a+2b$) to be meaningful for it. It's not just that I is infinite; I is truly meaningless and by various regroupings, factorings, and the like, it can seem to be equal to anything you want.

The only truly well-defined infinite series are those that are *absolutely convergent.* The series

$$S = \sum_{n=0}^{\infty} a_n$$

is absolutely convergent when

$$A = \sum_{n=0}^{\infty} |a_n|$$

converges to a finite value. If a series S seems to converge to a finite value but A does not, then by rearranging, regrouping, and the like you can make S equal to just about anything.

Since the zT terms are $a_n = s_n z^n$, our first guess might be that $|z|$ must be very small for the sum to converge absolutely. Note, however, that the sum in the zT is from negative infinity to positive infinity; for absolute convergence we require

$$A = \sum_{n=-\infty}^{\infty} |s_n||z|^n = \sum_{n=-\infty}^{-1} |s_n||z|^n + \sum_{n=0}^{\infty} |s_n||z|^n = \sum_{n=1}^{\infty} |s_{-n}||\zeta|^n + \sum_{n=0}^{\infty} |s_n||z|^n$$

where we defined $\zeta \equiv z^{-1}$. If $|z|$ is small then $|\zeta|$ is large, and consequently small values of $|z|$ can be equally dangerous. In general, the **R**egion **O**f **C**onvergence (ROC) of the z transform will be a ring in the z-plane with the origin at its center (see Figure 2). This ring may have $r = 0$ as its lower radius (and so be disk-shaped), or have $r = \infty$ as its upper limit, or even be the entire z-plane. When the signal decays to zero for both $n \to -\infty$ and $n \to \infty$ the ring will include the unit circle.

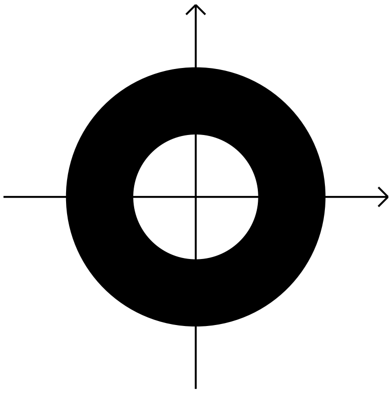

Figure 2: In general, the region of convergence (ROC) of the z transform is a ring in the z-plane with the origin at its center.

As a simple exercise - what happens to the ROC when we apply the delay operator (i.e., if we know the ROC of zTs what is the ROC of zT$\hat{z}^{-1}s$? A harder question is what happens to the ROC when we apply the time reversal operator?