Catalan numbers

Frazer Jarvis

Catalan numbers appear in many places throughout mathematics, and are a fairly frequent subject for articles in this magazine. Indeed, a very nice introduction to their properties was given in this magazine by Vun and Belcher [3], and a historical account of their discovery in Chinese mathematics (before Catalan!) has also appeared in a recent article by Larcombe [1]. The article by Vun and Belcher gives some combinatorial situations in which Catalan numbers play a role, such as the problem of dissecting a polygon into triangles. In this note, we'll briefly review the main properties of Catalan numbers, and give another (possibly new) derivation of the generating function.

The main definition for Catalan numbers that we will use is the following.

Definition 1 The nth *Catalan number*, denoted C_n, is the number of ways of multiplying together n symbols.

This isn't a particularly enlightening definition, and we will see the sequence arising in other more interesting situations. But let's just think about this definition and what it means.

Suppose $n = 3$ for simplicity, and we want to multiply a by b by c. We can essentially do this in two ways, first multiplying a and b, and then by c, or by multiplying b and c and then by a. Symbolically, these two operations are given by $(ab)c$ and $a(bc)$. As there are two ways to multiply 3 objects, we see that $C_3 = 2$.

Note that we are insisting that a, b and c remain in the same order: given three matrices A, B and C, we will have $(AB)C = A(BC)$, but in general, any permutation of the order in which A, B and C appear will give rise to a different product.

To multiply 4 objects, a, b, c and d, there are five different ways:

$$(((ab)c)d), \quad ((a(bc))d), \quad ((ab)(cd)), \quad (a((bc)d)), \quad \text{and } (a(b(cd))).$$

This shows that $C_4 = 5$.

How to compute the nth Catalan number

There are various ways of obtaining a simple formula for C_n, one of which is by using the generating function: we will give a method for computing this in the next section. The derivation we give now is based on one given by Singmaster [2] (in turn based on an idea of D.André from 1878), and relies on the idea of a *reverse Polish string*. When we do a sum like $3 + 4 \times 6$, we can easily scan the expression to see which part of the sum we should evaluate first ($4 \times 6 = 24$ in the example), and which part should be evaluated second ($3 + 24 = 27$): we can scan the expression much faster than we can do the calculation, and so the way in which the sum is presented is not important. But computers can carry out the arithmetic (almost) instantly, and if we want the computer to do the sum as quickly as possible – very useful if variants of the calculation must be done many times – we have to think about how to present the data to the computer so that it knows the order to do the calculations, preferably using as little memory as possible. One solution is the reverse Polish method, in which the sum would be presented as a string

$$3\ 4\ 6\ \times\ +$$

in which the computer starts by reading in the values 3, 4 and 6 to memory locations. Then the '$\times$' means "multiply the last two entries": now the memory locations consist of "3" and "24". Finally, the '$+$' means "add the previous two entries", so that 3 and 24 are added and the final answer 27 is output. Brackets are never required in reverse Polish notation.

Remember that C_n is the number of ways of multiplying n symbols. Let's multiply a, b and c. A reverse Polish string will be a string consisting of these symbols (in this order), and involving another symbol X, meaning "multiply the last two things in the string".

Consider the string $abXcX$. We first read in a and b, then we multiply to form ab, next read in c, and then multiply these two things to get $(ab)c$. The string $abcXX$ means: read in a, then b, then c; now multiply the last two things so that the string has a and bc, and finally multiply these to get $a(bc)$.

Remark 1 In fact, we can form the reverse Polish string for any product simply by deleting all left parentheses, and substituting X for all right parentheses. Thus the string $(a((bc)d))$ has reverse Polish string $abcXdXX$. Let's check this works: we read in a, then b and then c, multiply the last two so that the string consists of a and bc, then read in d, multiply the last two to get a and $(bc)d$, and the final multiplication gives $a((bc)d)$ as required.

You might like to write down the reverse Polish strings for all five of the products of 4 symbols, and go through the interpretation of each product in terms of operations on the objects of the string.

Let's write S for one of the symbols a, b,.... Then the two ways of multiplying three objects are represented by the strings $SSSXX$ and $SSXSX$.

A string representing multiplication of n objects must have n copies of the symbol S. Also, as the operation X reduces the number of objects on the string by 1, replacing two objects with their product, and we have to go from n objects to 1 product, the string must have $n-1$ copies of the symbol X. There are therefore $2n-1$ symbols in the string, with n of these being S, and the other $n-1$ being X.

However, not every possible string represents a valid way of multiplication. We can only apply the symbol X when there are at least 2 objects in the string. For $n = 2$, the only valid string is SSX; both of the strings XSS and SXS would involve multiplication of fewer than 2 symbols! It is easy to see that a string of Ss and Xs is legitimate precisely when the number of Ss in the first k symbols of the string is more than the number of Xs, for every value of k.

We can represent the string by means of a graph: if the next symbol is S, then we move 1 unit upwards in the y-direction, and if it is X, we move 1 unit downwards.

For example, the string $SSSXSXX$ of the remark has the following graph:
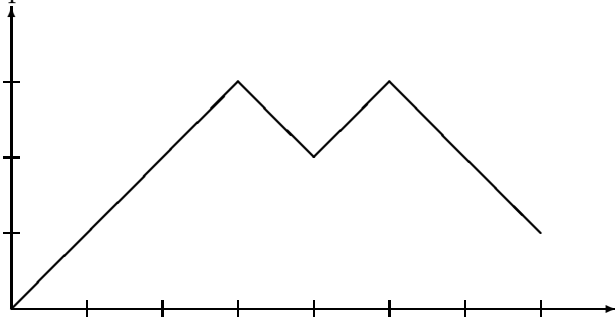
The line always begins at $(0,0)$ and ends at $(2n-1, 1)$.

The condition that a string be legitimate is that the graph always lies above the x-axis. So the first move *must* be from $(0,0)$ to $(1,1)$. Now we have to count the number of walks from $(1,1)$ to $(2n-1,1)$ which do not touch (or cross) the axis. There are $\binom{2n-2}{n-1}$ walks from $(1,1)$ to $(2n-1,1)$ in total, got by choosing $n-1$ steps upwards, and making the remaining $n-1$ go downwards; we will count the number that do touch the axis and subtract to get the number we want.

We have the following result:

> The number of walks from $(1, 1)$ to $(2n − 1, 1)$ which touch the
> axis is the same as the number of walks from $(1, 1)$ to $(2n−1, −1)$.

We'll show that there is a correspondence between these two sorts of walks. The idea is very simple. As soon as a walk touches the axis, we reflect the rest of the walk in the axis. So if we have a walk such as the following:
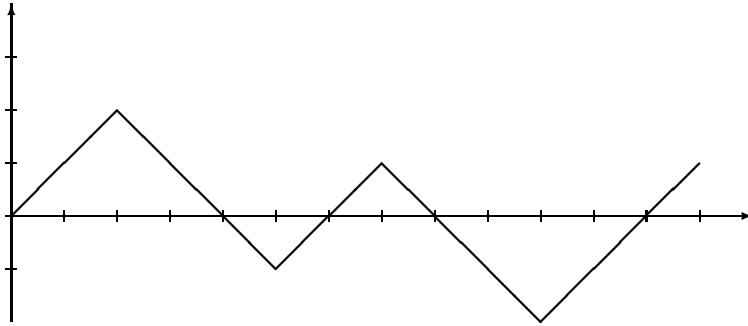
we reflect the part of the graph after time $t = 4$ to get:
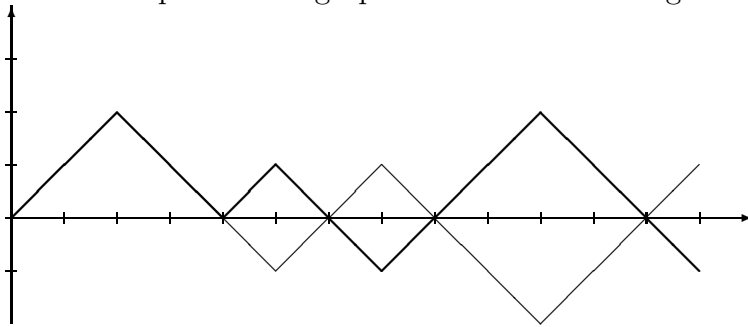
It is easy to construct a converse: given any walk from $(1, 1)$ to $(2n − 1, −1)$, it must hit the x-axis at some point. If we reflect the graph from that point onwards, then we get a walk from $(1, 1)$ to $(2n − 1, 1)$ which touches the axis.

Now any walk from $(1, 1)$ to $(2n − 1, −1)$ must go upwards $n − 2$ times and downwards n times, and so there are $\binom{2n-2}{n}$ such paths. It follows that the number of walks from $(1, 1)$ to $(2n − 1, 1)$ which do *not* touch the x-axis

is

$$\binom{2n-2}{n-1} - \binom{2n-2}{n} = \frac{(2n-2)!}{(n-1)!(n-1)!} - \frac{(2n-2)!}{n!(n-2)!}$$
$$= \frac{(2n-2)!}{(n-1)!(n-2)!}\left(\frac{1}{n-1} - \frac{1}{n}\right)$$
$$= \frac{(2n-2)!}{(n-1)!(n-2)!}\frac{1}{n(n-1)}$$
$$= \frac{1}{n}\frac{(2n-2)!}{(n-1)!(n-1)!}$$
$$= \frac{1}{n}\binom{2n-2}{n-1}.$$

These kinds of graph also appeared in the article of Vun and Belcher: they may be interpreted as representing the problem of counting the number of ways in which a tied election between two candidates can be held in which one candidate was always ahead of the rival, except before the first vote was cast, and after the last vote was cast.

The generating function

If we have a sequence of numbers, $a_0, a_1, a_2, \ldots$, then we can form the series

$$a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots.$$

This series is known as the *generating function* for the sequence. Often the generating function allows you to perform calculations with the terms in the sequence, and thus to give easier ways to prove certain identities. There is a lovely book by Wilf [4], just about generating functions, and how to manipulate them to get interesting results.

One interesting feature is that it is often possible to give other ways to write generating functions. For example, the sequence 1, 2, 3,...has the generating function
$$1 + 2x + 3x^2 + 4x^3 + \cdots,$$

but we might recognise this as the expansion of $\frac{1}{(1-x)^2}$. This single function therefore encodes the sequence of natural numbers (positive integers). More complicated sequences can also be similarly encoded; the Fibonacci numbers, defined by $F_0 = F_1 = 1$, $F_{n+1} = F_n + F_{n-1}$ for $n \geq 1$, arise as the coefficients of the function $\frac{1}{1-x-x^2}$. (To prove this, let $F(x) = 1 + x + 2x^2 + 3x^3 + 5x^4 + \cdots$ be the generating function, and write down $xF(x)$ and $x^2F(x)$. Now add them,

and note that you get $F(x)-1$. Now rearrange the equality $x^2F(x)+xF(x) = F(x) - 1$ to get the above expression for $F(x)$.)

Here's a way of finding the generating function for the Catalan numbers, which is not in [4].

The generating function is

$$C(x) = C_0 + C_1 x + C_2 x^2 + C_3 x^3 + C_4 x^4 + \cdots,$$

where C_n, the nth Catalan number, is the number of ways of bracketing products together, as mentioned above.

Thus

$$
\begin{aligned}
C(x) &= x + x^2 + 2x^3 + 5x^4 + \cdots \\
&= x + x^2 + [x^3 + x^3] + [x^4 + x^4 + x^4 + x^4 + x^4] + \cdots \\
&= x + (x.x) + ((x.x)x) + (x(x.x)) \\
&\quad + (((x.x)x)x) + ((x(x.x))x) + ((x.x)(x.x)) + (x((x.x)x)) + (x(x(x.x))) + \cdots
\end{aligned}
$$

where each term coresponds to multiplying x by itself using all the possible bracketings. Observe that, apart from the first term, if we strip off the outer pair of brackets, every term is naturally the product of two smaller terms in the series for $C(x)$:

$$
\begin{aligned}
C(x) &= x + x.x + (x.x)x + x(x.x) \\
&\quad + ((x.x)x)x + (x(x.x))x + (x.x)(x.x) + x((x.x)x) + x(x(x.x)) + \cdots
\end{aligned}
$$

and a term like $((x.x)x)x$ is the product of the two terms $((x.x)x)$ and x. In fact, it is easy to see that, apart from the first term, the other terms are exactly $C(x)^2$ (all terms apart from the first are a product of two smaller terms, and any product of two smaller terms will arise as a way of multiplying x by itself a suitable number of times, so will appear in $C(x)$). Thus

$$C(x) = x + C(x)^2,$$

so that $C(x)^2 - C(x) + x = 0$. We can solve this with the quadratic formula to give

$$C(x) = \frac{1 \pm \sqrt{1 - 4x}}{2}.$$

In fact, we must choose the minus sign here, otherwise the coefficients of the powers of x in the generating function of $C(x)$ are all negative, whereas we want $C(x)$ to be the generating function of the Catalan numbers, all of which are positive. Indeed, if we expand the square root $\sqrt{1 - 4x} = (1 - 4x)^{1/2}$ as a series $1 + a_1 x + a_2 x^2 + \cdots$ using the binomial formula, we have

$$\sqrt{1 - 4x} = 1 - 2x - 2x^2 - \cdots,$$

6

and so we need to choose the minus sign to get positive coefficients in $C(x)$. Then
$$C(x) = \frac{1 - \sqrt{1 - 4x}}{2}.$$
The binomial formula tells us that the coefficient of x^n (if $n > 0$) in this series is

$$
\begin{aligned}
C_n &= -\frac{1}{2} \left\{ \frac{\frac{1}{2}(-\frac{1}{2})(-\frac{3}{2}) \cdots (-\frac{(2n-3)}{2})}{n!} (-4)^n \right\} \\
&= \frac{1}{2} \left\{ \frac{\frac{1}{2}(\frac{1}{2})(\frac{3}{2}) \cdots (\frac{(2n-3)}{2})(n-1)!}{n!(n-1)!} (2^2)^n \right\} \\
&= \frac{1}{2} \left\{ \frac{\frac{1}{2}[\frac{1}{2}.1.\frac{3}{2}.2 \cdots .(n-2)(\frac{(2n-3)}{2})(n-1)]}{n!(n-1)!} 2^{2n} \right\} \\
&= \frac{1}{2} \left\{ \frac{\frac{1}{2}(1.2.3.4 \cdots .(2n-4)(2n-3)(2n-2)}{n!(n-1)!} 2^2 \right\} \\
&= \frac{(2n-2)!}{n!(n-1)!} \\
&= \frac{1}{n} \binom{2n-2}{n-1}.
\end{aligned}
$$

References

[1] Peter Larcombe, *The 18th century Chinese discovery of the Catalan numbers*, Mathematical Spectrum 32 (1999/2000) 5–6

[2] David Singmaster, *An elementary evaluation of the Catalan numbers*, American Math. Monthly 85 (1978) 366–368

[3] Isaac Vun, Paul Belcher, *Catalan numbers*, Mathematical Spectrum 30 (1997/8) 3–5

[4] Herbert S.Wilf, *Generatingfunctionology*, Academic Press (1994), (*see also* http://www.cis.upenn.edu/∼wilf/)

Frazer Jarvis *is a member of the Department of Pure Mathematics at the University of Sheffield, and his research interests are centred on algebraic number theory. Outside mathematics, he is a keen pianist and sings with the Sheffield Philharmonic Chorus.*