The Average Height of Catalan Trees by Counting Lattice Paths

Nachum Dershowitz   Christian Rinderknecht

In memoriam Philippe Flajolet, friend and colleague

Structured documents, like books, articles, and web pages, are composed of chapters, sections, paragraphs, figures, appendices, indices, etc. The occurrences of these components are mutually constrained; for instance, it is understood that a section is part of a chapter and that appendices are located at the end of a document. This hierarchical layout is meant to facilitate reading, and it supports the search for specific items of information. When considering computer systems, these data must be uniformly encoded by means of a formal language.

Consider, for instance, an email message. It contains at least the sender's address, a subject or title, the recipient's address, and a body of text. These elements correspond to *nodes* arranged in a structure called a *Catalan tree*, a.k.a. an ordered tree or rooted plane tree. For example, the email

```
From: Me
Subject: Homework
To: You

A deadline is a due date for a homework.
```

can be modeled by the tree in Figure 1, where the topmost node ("email") is called the *root* and the framed pieces of text are *leaves*. Note that, for historical reasons, computer scientists grow their trees upside down, with the root at the top. The inner (non-leaf) nodes hold "metadata", or "markup", that is, information about the nature of the data contained in the subtree.

Catalan trees are a pervasive data structure in computer science, in that they are a natural representation for hierarchical data. For example, in XML (eXtensible Markup Language), textual information is stored in leaves, and, consequently, its retrieval requires the traversal of the tree from the root to a leaf. The *height* of a tree is the number of nodes on a maximal path from

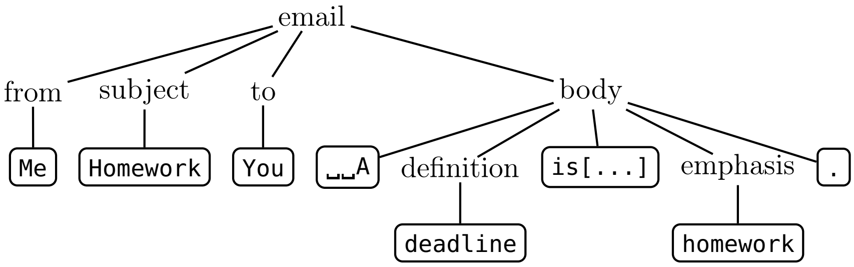


Figure 1: An email viewed as a Catalan tree

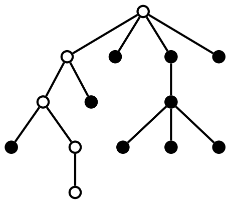


Figure 2: Catalan tree of height 5 and size 13.

root to leaf; for example, travel down the path with nodes depicted as $\circ$ in the tree of height 5 in FIGURE 2.

In general, the maximum cost of a search is proportional to the height of the tree, and the determination of the average height becomes relevant when performing a series of random searches [16]. The mathematical study of this average quantity often relies on advanced analytical tools, and the purpose of the present note is to propose a partial simplification of these approaches by using elementary combinatorics.

The Analytical Derivation

We measure the *size* of a tree by the number of its edges; for example, the tree in FIGURE 2 is of size 13. Let h_n be the average height of Catalan trees of size n and H_n^h the number of Catalan trees of size n and height h. We then have $h_n = S_n/C_n$, where $S_n := \sum_{h \geqslant 1} h\, H_n^h$, and $C_n := \binom{2n}{n}/(n+1)$ is the number of Catalan trees of size n. The height of a tree with n edges can range from 2 (all leaves directly below the root) to $n+1$ (one straight path from root to a lone leaf).

To gain purchase on the sum S_n, we may define $H_n^{<h}$ as being the number of trees with n edges and height less than h. Then $H_n^h = H_n^{<h+1} - H_n^{<h}$. Of

course, we have $H_n^{<h} = H_n^{<n+2} = C_n$, if $h > n+1$. Formulas can be further simplified by letting $H_n^{\geqslant h}$ be the number of trees with n edges and height greater than or equal to h. Now we have:

$$S_n = \sum_{h \geqslant 1} h\left(H_n^{<h+1} - H_n^{<h}\right) = \sum_{h \geqslant 1} h\left(H_n^{\geqslant h} - H_n^{\geqslant h+1}\right) = \sum_{h \geqslant 1} H_n^{\geqslant h}. \qquad (1)$$

Knuth, de Bruijn, and Rice [11] published a landmark paper in 1972, where they obtained the asymptotic approximation of the average height h_n. They started by modeling the problem with a generating function [17] that satisfies a recurrence equation whose solution expresses the generating function in terms of continued fractions of Fibonacci polynomials. Integration over complex numbers is then utilized to obtain the formula

$$H_n^{\geqslant h} = \sum_{k \geqslant 1}\left[\binom{2n}{n+1-kh} - 2\binom{2n}{n-kh} + \binom{2n}{n-1-kh}\right]. \qquad (2)$$

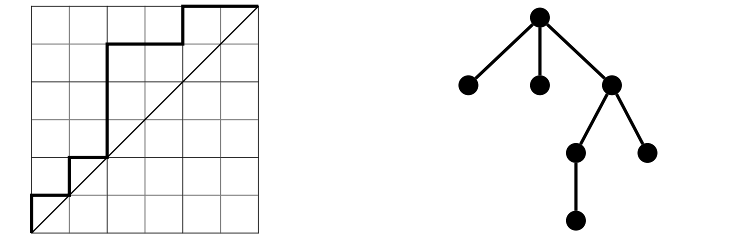
The authors conclude by employing real and complex analysis to obtain asymptotic expansions of $H_n^{\geqslant h}$, S_n, and h_n. As we will see, the main term is $h_n \sim \sqrt{\pi n}$, where $f(n) \sim g(n)$ means $\lim_{n\to\infty} f(n)/g(n) = 1$, wherever f and g are defined.

The purpose of the present note is to show how to circumvent heavy analytic techniques in the derivation of equation (2). Instead, we propose an elementary combinatorial proof based on the enumeration of the Dyck paths of a certain height, which are in bijection with Catalan trees of a related height. We find this bijective proof to be more intuitive, in particular to computer scientists, for whom the result matters for the analysis of algorithms. Technically, our approach is in tune with Mohanty [12], as well as Dershowitz and Zaks [2].

Counting Catalan trees

Before we determine $H_n^{\geqslant h}$, let us solve a related and easier question: deriving the number C_n of Catalan trees with n edges, called the *Catalan number*.

In 1984, Kemp [9, p. 64] (see also [5]) derived equation (2) by analytical means too, but, instead of working directly with Catalan trees, he used certain lattice paths in an integer grid. *Monotonic lattice paths* [12, 8] are made up of two kinds of steps, oriented upwards and oriented rightwards, starting at $(0,0)$ with an upward step. *Dyck paths* of length $2n$ are monotonic paths ending at (n,n) that never venture below the diagonal; an example for $n = 6$ is shown in FIGURE 3a.



(a) Dyck path of length 12. (b) Catalan tree with 6 edges.

Figure 3: Bijection between Dyck paths and Catalan trees.

A bijection with Dyck paths Crucially, there is a bijection between Dyck paths of length $2n$ and Catalan trees with n edges [10].

This bijection is shown on an example in FIGURE 3. To construct the lattice path in FIGURE 3a from the tree in FIGURE 3b, we imagine that the tree is a roadmap and our avatar plans a tour starting at the root as follows: we take the rightmost unvisited road (from the avatar's viewpoint), else we backtrack: in the end, we have taken each road twice: there, and back again. More technically, in FIGURE 4, we follow the dotted arrows: each downward arrow in the tree corresponds to a step up $\uparrow$ (called a *rise*) in the lattice, and an upward arrow in the tree to a step right $\rightarrow$ (called a *fall*). In the tree, the series is $\downarrow \uparrow \downarrow \uparrow \downarrow \downarrow \downarrow \uparrow \uparrow \downarrow \uparrow \uparrow$, which becomes $\uparrow \rightarrow \uparrow \rightarrow \uparrow \uparrow \uparrow \rightarrow \rightarrow \uparrow \rightarrow \rightarrow$ in the lattice. If we follow the latter from the start at the bottom left corner $(0, 0)$, we obtain the path in FIGURE 3a. This kind of traversal is called *preorder*, or "document" order, because it is the way we would read the document represented by the tree, from cover to cover. Note as well that there are always $n + 1$ nodes if, and only if, there are n edges in the tree, because there is precisely one edge per node going up, save for the topmost node (*root*).

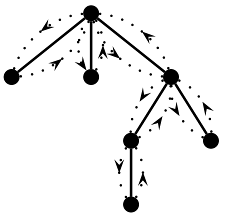


Figure 4: Preorder traversal

The inclusion-exclusion principle The previous bijection allows us to count the Catalan trees with n edges by counting instead the Dyck paths of length $2n$.

It is easy to count all the monotonic paths of length $2n$ because there are as many as choices of n rises amongst $2n$ steps, that is, $\binom{2n}{n}$. To count only the Dyck paths, we need to subtract the number of paths that start with a rise but cross below the diagonal at some point.

This approach is a simple instance of the method known as the *inclusion-exclusion principle*, whereby the direct and difficult enumeration of a set is replaced by an easier enumeration of a strict superset and the subtraction of the cardinality of a strict subset, so that the resulting sets are equal.

An example of a path that is not a Dyck path is shown in FIGURE 5, drawn in bold. The first point reached below the diagonal is used to plot a dotted line parallel to the diagonal back to the y-axis. All the steps on the path from that point back to $(0,0)$ are then changed into their counterpart: a rise is replaced by a fall and vice-versa. The resulting segment is drawn as connected dashed lines. This operation is called a *reflection* [13]. The crux of the matter is that we can reflect each monotonic path crossing the diagonal into a distinct path from $(1,-1)$ to (n,n). These reflected paths can, in turn, be reflected back into their original counterpart when they reach the dotted line. In other words, the mapping is bijective. (Another intuitive and visual approach to the same result has been published by Callan [1].) Consequently, there are as many monotonic paths from $(0,0)$ to (n,n) that cross the diagonal as there are monotonic paths from $(1,-1)$ to (n,n). The latter are readily enumerated: $\binom{2n}{n-1}$. In conclusion, the number of Dyck paths of length $2n$ is

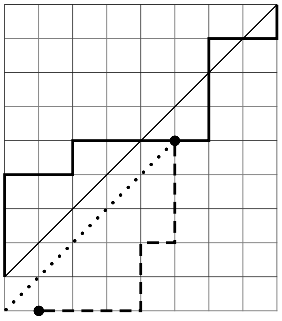


Figure 5: Reflection of a prefix with respect to $y = x - 1$.

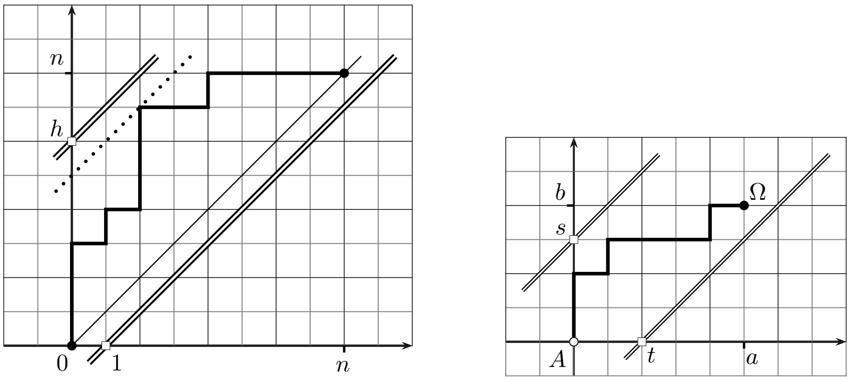
$$\begin{aligned} C_n &= \binom{2n}{n} - \binom{2n}{n-1} = \binom{2n}{n} - \frac{(2n)!}{(n-1)!(n+1)!} \qquad (3) \\ &= \binom{2n}{n} - \frac{n}{n+1}\frac{(2n)!}{n!n!} = \binom{2n}{n} - \frac{n}{n+1}\binom{2n}{n} = \frac{1}{n+1}\binom{2n}{n}. \end{aligned}$$

Using Stirling's formula for the asymptotic equivalence, we draw the conclusion:

$$C_n = \frac{1}{n+1}\binom{2n}{n} \sim \frac{4^n}{n\sqrt{\pi n}}, \quad \text{as } n \to \infty. \qquad (4)$$

A Combinatorial Proof

In 1996, Sedgewick and Flajolet [15, 6] derived the enumerations of Catalan trees by height, also using analytic combinatorics, but they employed real analysis to obtain the asymptotic approximation of $H_n^{\geqslant h}$. They write [15, p. 260]:



(a) Dyck path of length $2n$ and height $h-1$.

(b) Path from A to Ω avoiding the boundaries $y = x + s$ and $y = x - t$.

Figure 6: Paths avoiding diagonal boundaries.

> *This analysis is the hardest nut that we are cracking in this book. It combines techniques for solving linear recurrences and continued fractions, generating function expansions, especially by the Lagrange inversion theorem, and binomial approximations and Euler-Maclaurin summations.*

To avoid the aforementioned advanced techniques used to derive equation (2), we use again a bijection between Dyck paths and Catalan trees, but, this time, the key point is that Catalan trees of size n and height h are in bijection with Dyck paths of length $2n$ and height $h-1$. This simple observation allows us to reason about the height of the Dyck paths and transfer our findings back to Catalan trees.

With the determination of $H_n^{<h}$ in mind, let us consider a Dyck path of length $2n$ and height $h-1$, as in FIGURE 6a. The double lines are boundaries that may not be attained by the path. This is in fact a special case of a general monotonic path between two diagonal boundaries, as shown in FIGURE 6b, where s denotes the vertical distance from A, and t, the horizontal distance from A. It is well known that the number of monotonic paths from $A(0,0)$ to $\Omega(a,b)$ avoiding the boundaries $y = x+s$ and $y = x-t$ is

$$|\mathcal{L}(a,b;t,s)| = \sum_{k\in\mathbb{Z}} \left[\binom{a+b}{b+k(s+t)} - \binom{a+b}{b+k(s+t)+t}\right]. \qquad (5)$$

The proof by Mohanty [12, p. 6] of this formula is based on the reflection principle and the principle of inclusion and exclusion, which we used earlier. We quote his proof here verbatim, because it is rarely found in print nowadays.

Proof (Mohanty [12]). For brevity, call the boundaries $x = y + t$ and $x = y - s$, $\mathcal{L}^+$ and $\mathcal{L}^-$, respectively. Denote by A_1 the set of paths that reach $\mathcal{L}^+$, by A_2 the set of paths that reach $\mathcal{L}^+$, $\mathcal{L}^-$ in that order, and in general by A_i the set of paths reaching $\mathcal{L}^+$, $\mathcal{L}^-$, $\mathcal{L}^+$, ... (i times) in the specified order. Similarly, let B_i be the set of paths reaching $\mathcal{L}^-$, $\mathcal{L}^+$, $\mathcal{L}^-$, ... (i times) in the specified order. An application of the usual inclusion-exclusion method yields

$$|\mathcal{L}(a,b;t,s)| = \binom{a+b}{b} + \sum_{i\geqslant 1}(-1)^i\left(|A_i| + |B_i|\right), \tag{6}$$

where $|A_i|$ and $|B_i|$ are evaluated by using the reflection principle repeatedly. For example, consider A_3. Since every path in A_3 must reach $\mathcal{L}^+$, A_3 when reflected about $\mathcal{L}^+$ becomes the set of paths from $(t,-t)$ to (a,b) each of which reaches $\mathcal{L}^+$ after reaching $\mathcal{L}^-$. Another reflection about $\mathcal{L}^-$ would make A_3 equivalent to the set of paths from $(-s-t, s+t)$ to (a,b) that reach $\mathcal{L}^+$, which in turn can be written as $R(a+s+t, b-s-t; 2s+3t)$. [Note: $R(a,b;t)$ is the set of paths from $(0,0)$ to (a,b) reflected about $\mathcal{L}^+$.] Thus, since $|R(a,b;t)| = \binom{a+b}{a-t}$, we have

$$|A_3| = \binom{a+b}{a-s-2t},$$

and, more generally,

$$|A_{2j}| = \binom{a+b}{a+j(s+t)} \quad \text{and} \quad |A_{2j+1}| = \binom{a+b}{a-j(s+t)-t}.$$

The expressions for $|B_{2j}|$, $|B_{2j+1}|$, $j = 0,1,2,\ldots$, with $|A_0|$, $|B_0|$ being $\binom{a+b}{b}$, are obtained by interchanging a with b and s with t. Substitution of these values in (6) yields (5) after some simplifications. □

Resuming our argument, if we match the subfigures in Figure 6, we find $s = h$, $t = 1$, $a = b = n$, hence $a + b = 2n$ and $b + k(s+t) = n + k(h+1)$, which we plug into formula (5) and change h into $h - 1$:

$$H_n^{<h} = \sum_{k\in\mathbb{Z}}\left[\binom{2n}{n+kh} - \binom{2n}{n+1+kh}\right].$$

After splitting the sum into $k < 0$, $k = 0$, and $k > 0$, then changing the sign of k in the first case, using $\binom{p}{q} = \binom{p}{p-q}$ in the second, and lastly gathering the remaining sums ranging over $k \geqslant 1$, we reach

$$H_n^{<h} = -\sum_{k\geqslant 1}\left[\binom{2n}{n+1-kh} - 2\binom{2n}{n-kh} + \binom{2n}{n-1-kh}\right]$$
$$+\binom{2n}{n} - \binom{2n}{n-1}.$$

Recognizing C_n from equation (3), we simplify as follows:

$$C_n - H_n^{<h} = \sum_{k\geqslant 1}\left[\binom{2n}{n+1-kh} - 2\binom{2n}{n-kh} + \binom{2n}{n-1-kh}\right].$$

Finally, recalling that $H_n^{\geqslant h} = C_n - H_n^{<h}$, we arrive at

$$H_n^{\geqslant h} = \sum_{k\geqslant 1}\left[\binom{2n}{n+1-kh} - 2\binom{2n}{n-kh} + \binom{2n}{n-1-kh}\right],$$

which is none other than our target, equation (2).

In this way, we have achieved our goal merely by enumerating lattice paths, and hopefully have, in the process, made this classic result less daunting.

Asymptotics

We could stop here, but we would like to give a hint as to how the asymptotic approximation is carried out. The approximation will give us a practical handle on the expected height of Catalan trees, which in turn tells us what to expect by way of performance of algorithms, like search, that traverse down paths in arbitrary trees.

Equation (1) entails $S_n = \sum_{h\geqslant 1} H_n^{\geqslant h}$; therefore

$$S_n = \sum_{k'\geqslant 1} d(k')\left[\binom{2n}{n+1-k'} - 2\binom{2n}{n-k'} + \binom{2n}{n-1-k'}\right],$$

where $d(k')$ is the number of positive divisors of k', but complex analysis is needed [11, 4]. Another way is to express the binomials in terms of $\binom{2n}{n-kh}$:

$$\binom{2n}{n-m+1} = \frac{(2n)!}{(n-m+1)!\,(n+m-1)!}$$

$$
\begin{aligned}
&= \frac{(2n)!\,(n+m)}{(n-m)!\,(n-m+1)(n+m)!} = \frac{n+m}{n-m+1}\binom{2n}{n-m},\\
\binom{2n}{n-m-1} &= \frac{(2n)!}{(n-m-1)!\,(n+m+1)!}\\
&= \frac{(2n)!\,(n-m)}{(n-m)!\,(n+m)!\,(n+m+1)} = \frac{n-m}{n+m+1}\binom{2n}{n-m}.
\end{aligned}
$$

Therefore,

$$\binom{2n}{n-m+1} - 2\binom{2n}{n-m} + \binom{2n}{n-m-1} = 2\,\frac{2m^2-(n+1)}{(n+1)^2-m^2}\binom{2n}{n-m}.$$

Let $F_n(m) = (2m^2-n)/(n^2-m^2)$. We have

$$S_n = 2\sum_{h\geqslant 1}\sum_{k\geqslant 1} F_{n+1}(kh)\binom{2n}{n-kh}.$$

From equation (4) and $h_n = S_n/C_n$, we deduce $h_n = (n+1)S_n/\binom{2n}{n}$, hence we must approximate $(n+1)F_{n+1}(m)$ and $\binom{2n}{n-m}/\binom{2n}{n}$. On the one hand, we have

$$F_{n+1}(m) \sim \frac{2m^2-n}{n^2} \sim \frac{2m^2-n}{n(n+1)},$$

so $(n+1)F_{n+1}(kh) \sim 2k^2h^2/n - 1$. On the other hand, Sedgewick and Flajolet [15, 4.6, 4.8] show

$$\binom{2n}{n-m}\bigg/\binom{2n}{n} \sim e^{-m^2/n}.$$

Assuming that the tails (the implicit error terms) of the two previous approximations decrease exponentially, we have

$$h_n \sim \sum_{h\geqslant 1}\sum_{k\geqslant 1}(4k^2h^2/n-2)e^{-k^2h^2/n} = \sum_{h\geqslant 1} H(h/\sqrt{n}),$$

where $H(x) = \sum_{k\geqslant 1}(4k^2x^2-2)e^{-k^2x^2}$. Finally, Sedgewick and Flajolet [15, §5.9], on the one hand, and Graham, Knuth, and Patashnik [7, §9.6], on the other hand, use real analysis to conclude

$$h_n \sim \sum_{h\geqslant 1} H(h/\sqrt{n}) \sim \sqrt{n}\int_0^\infty H(x)dx \sim \sqrt{\pi n}.$$

The end of this derivation is difficult because the error terms in the bivariate asymptotic approximations must be carefully checked, so it is unlikely to be simplified further.

Remarkably, the main term $\sqrt{\pi n}$ in the asymptotic value for height can also be obtained by simple lattice-path arguments [3], as follows.

A Purely Combinatorial Derivation

We are going to proceed in two steps: first, we will bound the average height in terms of the average distance of a random node from the root; second, we will determine the latter, yielding the result only by combinatorial means.

Average height

We have already seen the correspondence between lattice paths and Catalan trees, in which a rise reaching the lth diagonal corresponds to a node at level l in the tree, counting levels from root level 0. A simple bijection between paths will show that for every node on level l of a tree of height h and size n, there is a corresponding node on either level $h - l$ or $h - l - 1$ in another tree of the same height and size.

Consider the Dyck path in FIGURE 7, in bijection with a tree with $n = 8$ edges and height $h = 4$. Let us find the last (rightmost) point on the path where it reaches its full height (the dotted line of equation $y = x + h - 1$), which we call the *apex* of the path (marked A in the figure). The immediately following fall leads to B and it is drawn with a double line. Let us rotate the segment from $(0, 0)$ to A, and the segment from B to (n, n) by 180°. The invariant fall (A, B) now connects the rotated segments. This way, what was the apex becomes the origin and vice-versa, making this a height-preserving bijection between paths. See FIGURE 8.

The point is that every rise to level l in FIGURE 7, representing a node on level l, ends up reaching level $h - l$ or $h - l - 1$ in FIGURE 8, depending on whether it was to the left (segment before A) or right (segment after B)

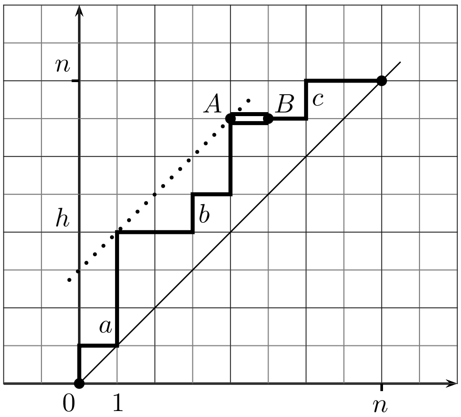


Figure 7: A Dyck path of length $2n$ and height $h - 1$

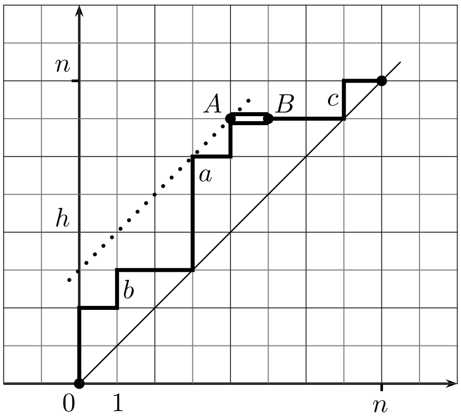


Figure 8: Dyck path in bijection with FIGURE 7

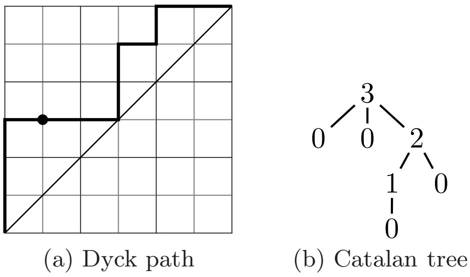
of the apex. In the example in the figure, the rise a reaches level 1, and its counterpart after the transformation rises to level $4 - 1 = 3$; the rise b reached level 2 and still does so because $4 - 2 = 2$; the rise c also reached level 2, but because it was to the right of the apex, it reaches now level $4 - 2 - 1 = 1$. It follows from this bijection that *the average height of trees with n nodes is within one of twice the average level of a node.*

We now have to determine the average level of a node in order to conclude. For this, we investigate the average path length of a tree.

Average path length

The *path length* of a Catalan tree is the sum of the lengths of the paths from the root. In order to study the average path length, we will follow Dershowitz and Zaks [2] in finding first the average number of nodes of degree d at level l in a Catalan tree with n edges, where the *degree of a node* is the number of its children (the number of nodes immediately below it).

Degree-based bijection The first step of our method for finding the average path length consists in finding an alternative bijection between Catalan trees and Dyck paths. In FIGURE 3b, we can see a Catalan tree equivalent to the Dyck path in FIGURE 3a, built from the preorder traversal of that tree. FIGURE 9b shows the same tree, where the contents of the nodes are their degree. The preorder traversal (of the degrees) is $(3, 0, 0, 2, 1, 0, 0)$. Since the last degree is always 0 (a leaf), we remove it and settle for $(3, 0, 0, 2, 1, 0)$. Another equivalent Dyck path may be obtained by mapping the degrees of that list into as many occurrences of rises (↑) and one fall (→), so, for in-



(a) Dyck path (b) Catalan tree

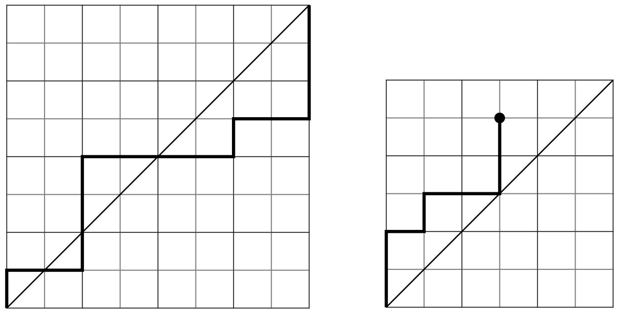
Figure 9: Degree-based bijection

stance, 3 is mapped to ↑ ↑ ↑ → and 0 to →. In the end, $(3,0,0,2,1,0)$ is mapped into ↑ ↑ ↑ → → → ↑ ↑ → ↑ → →, which corresponds to the Dyck path in FIGURE 9a. It is easy to convince ourself that we can reconstruct the tree from the Dyck path, so we indeed have a bijection.

The reason for this new bijection is that we need to find the average number of Catalan trees whose root has a given degree. This number will help us in finding the average path length, following an idea of Ruskey [14]. From the bijection, it is clear that the number of trees whose root has degree $r = 3$ is the number of Dyck paths made of the segment from $(0,0)$ to $(0,r)$, followed by one fall (see the dot at $(1,r)$ in FIGURE 9a), and then all monotonic paths above the diagonal until the upper right corner (n,n). Therefore, we need to determine the number of such paths.

Path reversal Let us add to our tool box one more bijection which often proves useful: *reversal*. It simply consists in reversing the order of the steps making up a path. Consider for example FIGURE 10a. Of course, the composition of two bijections being a bijection, the composition of a reversal and a reflection is bijective, hence the monotonic paths above the diagonal from $(1,r)$ to (n,n) are in bijection with the monotonic paths above the diagonal from $(0,0)$ to $(n-r,n-1)$. For example, FIGURE 10b shows the reversal and reflection of the Dyck path of FIGURE 9a after the point $(1,3)$, distinguished by the black disk (•).

Counting trees by root degree Recalling that Catalan trees with n edges are in bijection with Dyck paths of length $2n$, we now know that the number of Catalan trees with n edges and whose root has degree r is the number of monotonic paths above the diagonal from the point $(0,0)$ to $(n-r,n-1)$. We can find this number using the same technique we used for the total number C_n of Dyck paths. The principle of inclusion and exclusion says that we



(a) Reversal of FIGURE 5

(b) Reversal and reflection of FIGURE 9a after $(1, 3)$

Figure 10: Reversals and reflections

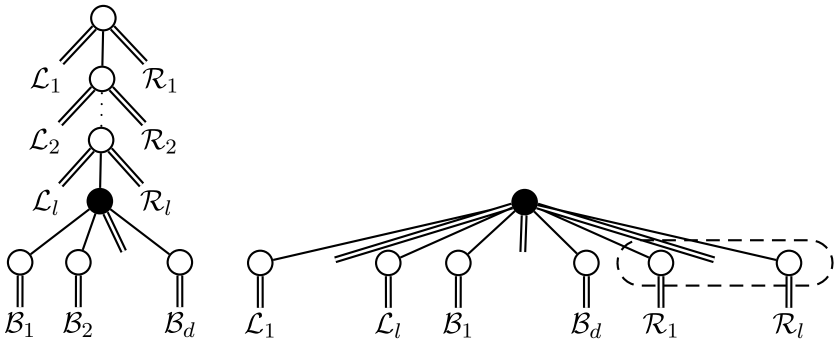
should count the total number of paths with the same extremities and retract the number of paths that cross the diagonal. The former is $\binom{2n-r-1}{n-1}$, which enumerates the ways to interleave $n-1$ rises (↑) and $n-r$ falls (→). The latter number is the same as the number of monotonic paths from $(1, -1)$ to $(n-r, n-1)$, as shown by reflecting the paths up to their first crossing, that is, $\binom{2n-r-1}{n}$; in other words, that is the number of interleavings of n rises with $n-r-1$ falls. Finally, imitating the derivation of equation (4), the number $\mathcal{R}_n(r)$ of trees with n edges and root of degree r is

$$\mathcal{R}_n(r) = \binom{2n-r-1}{n-1} - \binom{2n-r-1}{n}. \tag{7}$$

Counting trees by node level and degree Let $\mathcal{N}_n(l, d)$ be the number of Catalan trees with n edges at level l and of degree d. Ruskey [14] found a neat bijection to relate it to $\mathcal{R}_n(r)$ by the following equation:

$$\mathcal{N}_n(l, d) = \mathcal{R}_{n+l}(2l + d). \tag{8}$$

FIGURE 11a depicts the general pattern of a Catalan tree with node (•) of level l and degree d. The double edges denote a set of edges, so the $\mathcal{L}_i$, $\mathcal{R}_i$ and $\mathcal{B}_i$ actually represent forests. In FIGURE 11b, we see a Catalan tree in bijection with the former, from which it is made by lifting the node of interest (•) to become the root, the forests $\mathcal{L}_i$ with their respective parents are attached below it, then the $\mathcal{B}_i$, and, finally, the $\mathcal{R}_i$ for which new parents are needed (inside a dashed frame in the figure). Clearly, the new root is of degree $2l + d$ and there are $n + l$ edges. Importantly, the transformation



(a) n edges, ($\bullet$) is of degree d and at level l

(b) $n + l$ edges, root of degree $2l + d$

Figure 11: Bijection

can be inverted for any tree (it is injective and surjective), so it is indeed a bijection. From (7) and (8), we deduce

$$\mathcal{N}_n(l, d) = \binom{2n - d - 1}{n + l - 1} - \binom{2n - d - 1}{n + l}. \tag{9}$$

Average level of a node Let $\mathbb{E}[P_n]$ be the *average path length* of a Catalan tree with n edges. We have

$$\mathbb{E}[P_n] := \frac{1}{C_n} \sum_{l=0}^{n} l \sum_{d=0}^{n} \mathcal{N}_n(l, d), \tag{10}$$

because there are C_n trees and the double summation is the sum of the path lengths of all the trees with n edges. If we average again by the number of nodes, *i.e.*, $n+1$, we obtain the average level of a node in a random Catalan tree. In particular, equation (9) entails that the total number of nodes at level l in all Catalan trees with n edges is

$$\sum_{d=0}^{n} \mathcal{N}_n(l, d) = \sum_{d=0}^{n} \binom{2n - d - 1}{n + l - 1} - \sum_{d=0}^{n} \binom{2n - d - 1}{n + l}. \tag{11}$$

Let us consider the first sum:

$$\sum_{d=0}^{n} \binom{2n - d - 1}{n + l - 1} = \sum_{i=n-1}^{2n-1} \binom{i}{n + l - 1} = \sum_{i=n+l-1}^{2n-1} \binom{i}{n + l - 1}.$$

We have the derivation

$$\begin{aligned}
\binom{n+m}{n+1} &= \binom{n+m-1}{n} + \binom{n+m-1}{n+1} \\
&= \binom{n+m-1}{n} + \left[\binom{n+m-2}{n} + \binom{n+m-2}{n+1}\right] \\
&= \binom{n+m-1}{n} + \binom{n+m-2}{n} + \cdots + \left[\binom{n}{n} + \binom{n}{n+1}\right], \\
\binom{n+m}{n+1} &= \sum_{j=0}^{m-1} \binom{n+j}{n}.
\end{aligned}$$

This identity is equivalent to $\sum_{i=j}^{k} \binom{i}{j} = \binom{k+1}{j+1}$, so $j = n+l-1$ and $k = 2n-1$ yields

$$\sum_{d=0}^{n} \binom{2n-d-1}{n+l-1} = \binom{2n}{n+l}.$$

Furthermore, replacing l by $l+1$ gives $\sum_{d=0}^{n} \binom{2n-d-1}{n+l} = \binom{2n}{n+l+1}$, so we can now resume from equation (11) and find the total number of nodes at level l in all Catalan trees with n edges to be

$$\sum_{d=0}^{n} \mathcal{N}_n(l,d) = \binom{2n}{n+l} - \binom{2n}{n+l+1}. \tag{12}$$

Using equation (12) in definition (10), we draw

$$\begin{aligned}
\mathbb{E}[P_n] \cdot C_n &= \sum_{l=0}^{n} l \left[\binom{2n}{n+l} - \binom{2n}{n+l+1}\right] \\
&= \sum_{l=1}^{n} l \binom{2n}{n+l} - \sum_{l=0}^{n-1} l \binom{2n}{n+l+1} \\
&= \sum_{l=1}^{n} l \binom{2n}{n+l} - \sum_{l=1}^{n} (l-1) \binom{2n}{n+l} \\
&= \sum_{l=1}^{n} \binom{2n}{n+l} = \sum_{i=n+1}^{2n} \binom{2n}{i}.
\end{aligned}$$

The remaining summation is easy to crack because it is the sum of one half of an even row in Pascal's triangle, which is symmetric: the first half equals the second half, only the central element remaining – there are an odd number of entries in an even row. This is readily proven as follows: $\sum_{j=0}^{n-1}\binom{2n}{j} = \sum_{j=0}^{n-1}\binom{2n}{2n-j} = \sum_{i=n+1}^{2n}\binom{2n}{i}$. Therefore

$$\sum_{i=0}^{2n}\binom{2n}{i} = 2\sum_{i=n+1}^{2n}\binom{2n}{i} + \binom{2n}{n},$$

and we can continue as follows:

$$\frac{\mathbb{E}[P_n]}{n+1} = \frac{1}{2}\left[\sum_{i=0}^{2n}\binom{2n}{i} - \binom{2n}{n}\right]\bigg/\binom{2n}{n} = \frac{1}{2}\left[\binom{2n}{n}^{-1}\sum_{i=0}^{2n}\binom{2n}{i} - 1\right].$$

The remaining sum is perhaps the most famous combinatorial identity because it is a corollary of the venerable *binomial theorem*, which states that, for all real numbers x and y, and all positive integers n, we have the following equality:

$$(x+y)^n = \sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k.$$

Setting $x = y = 1$ yields the identity $2^n = \sum_{k=0}^{n}\binom{n}{k}$, which finally unlocks our last step, recalling the approximation (4):

$$\frac{\mathbb{E}[P_n]}{n+1} = \frac{1}{2}\left[4^n\bigg/\binom{2n}{n} - 1\right] \sim \frac{1}{2}\sqrt{\pi n}. \tag{13}$$

Conclusion Recalling that we proved that the average height of trees with n nodes is within one of twice the average level of a node, equation (13) entails

$$H_n \sim 2\,\frac{\mathbb{E}[P_n]}{n+1} \sim \sqrt{\pi n}.$$

References

[1] David Callan. Pair them up! A visual approach to the Chung-Feller theorem. *The College Mathematics Journal*, 26(3):196–198, May 1995.

[2] Nachum Dershowitz and Shmuel Zaks. Applied tree enumerations. In *Proceedings of the Sixth Colloquium on Trees in Algebra and Programming*, volume 112 of *Lecture Notes in Computer Science*, pages 180–193, Berlin, Germany, 1981. Springer.

[3] Nachum Dershowitz and Shmuel Zaks. The Cycle Lemma and Some Applications. *European Journal of Combinatorics*, 11(1):35–40, 1990.

[4] Philippe Flajolet, Xavier Gourdon, and Philippe Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144:3–58, 1995.

[5] Philippe Flajolet, Markus Nebel, and Helmut Prodinger. The scientific works of Rainer Kemp (1949–2004). *Theoretical Computer Science*, 355(3):371–381, April 2006.

[6] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, January 2009.

[7] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics*. Addison-Wesley, third edition, 1994.

[8] Katherine Humphreys. A history and a survey of lattice path enumeration. *Journal of Statistical Planning and Inference*, 140(8):2237–2254, August 2010. Special issue on Lattice Path Combinatorics and Applications.

[9] Reiner Kemp. *Fundamentals of the Average Case Analysis of Particular Algorithms*. Wiley-Teubner Series in Computer Science. John Wiley & Sons, B. G. Teubner, 1984.

[10] David A. Klarner. Correspondence between plane trees and binary sequences. *Journal of Combinatorial Theory*, 9:401–411, 1970.

[11] Donald E. Knuth, Nicolaas G. de Bruijn, and Stephen O. Rice. *The Average Height of Planted Plane Trees*, pages 15–22. Academic Press, December 1972. Republished in *Selected Papers on the Analysis of Algorithms*, CSLI Lecture Notes 102, Stanford University, CA, pp. 215–223, 2000.

[12] Sri Gopal Mohanty. *Lattice Path Counting and Applications*, volume 37 of *Probability and Mathematical Statistics*. Academic Press, New York, USA, January 1979.

[13] Marc Renault. Lost (and found) in translation: André's actual method and its application to the generalized ballot problem. *American Mathematical Monthly*, 155(4):358–363, April 2008.

[14] Frank Ruskey. A simple proof of a formula of Dershowitz and Zaks. *Discrete Mathematics*, 43(1):117–118, 1983.

[15] Robert Sedgewick and Philippe Flajolet. *An Introduction to the Analysis of Algorithms.* Addison-Wesley, 1996.

[16] Jeffrey Scott Vitter and Philippe Flajolet. *Average-Case Analysis of Algorithms and Data Structures*, volume A of *Handbook of Theoretical Computer Science*, pages 431–524. Elsevier Science, 1990.

[17] Herbert S. Wilf. *Generatingfunctionology.* Academic Press, 1990.

Summary The average height of Catalan trees of a given size is a structural parameter important in the analysis of algorithms, as it measures the expected maximum cost of a search in a tree. This parameter has been studied first with generating functions and complex variable theory, yielding an asymptotic approximation. Later on, real analysis was used instead of complex analysis. We have further reduced the conceptual difficulty by replacing generating functions with the enumeration of monotonic lattice paths, whose graphical representations make the derivation much more intuitive.