Self-describing sequences and the Catalan family tree

Zoran Šunić

Department of Mathematics
Texas A&M University
College Station, TX 77843-3368, USA

Submitted: Mar 19, 2002; Accepted: May 20, 2003; Published: May 29, 2003
MR Subject Classifications: 05A15, 05C05, 11Y55

Abstract

We introduce a transformation of finite integer sequences, show that every sequence eventually stabilizes under this transformation and that the number of fixed points is counted by the Catalan numbers. The sequences that are fixed are precisely those that describe themselves — every term t is equal to the number of previous terms that are smaller than t. In addition, we provide an easy way to enumerate all these self-describing sequences by organizing them in a Catalan tree with a specific labelling system.

Prefix ordered sequences and rooted labelled trees

The following connection between prefix ordered sequences and rooted labelled trees is well known and we briefly mention only the instance which is useful for our considerations.

Let $\mathcal{A}$ be the set of finite integer sequences $a = (a_0, a_1, \dots)$ with the property that $0 \leq a_i \leq i$, for all indices. We order the sequences in $\mathcal{A}$ by the *prefix* relation, i.e.,

$$(a_0, a_1, \dots, a_n) \preceq (b_0, b_1, \dots, b_m)$$

if $n \leq m$ and $a_i = b_i$, for $i = 0, \dots, n$. The sequences in $\mathcal{A}$ can be organized in a rooted labelled tree $\mathcal{T}$ which reflects the prefix order relation. The root of the tree $\mathcal{T}$ is labelled by 0. Every vertex that is at distance n from the root has $n+2$ children labelled by $0, 1, \dots, n, n+1$ (see Figure 1). The vertices whose distance to the root is n form the n-th *level* of the tree $\mathcal{T}$, which is also called the n-th *generation*. For every vertex v at the level n in the tree $\mathcal{T}$ there exist a unique path of length n from the root to v. The labels of the vertices on this path form a unique sequence $(a_0, a_1, \dots, a_n)$ in $\mathcal{A}$ that corresponds to the vertex v and this sequence is called the *full name* of v. The correspondence

$$v \leftrightarrow \text{the full name of } v$$

provides a bijection between the vertices in $\mathcal{T}$ and the sequences in $\mathcal{A}$. Under this bijection, the vertices from the n-th generation in $\mathcal{T}$ correspond to the sequences of length $n+1$ in

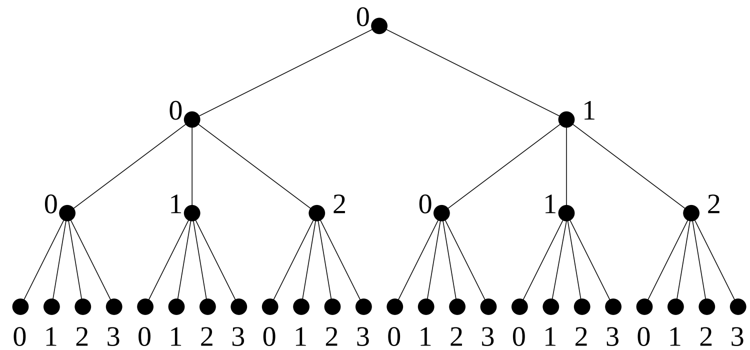

Figure 1: The rooted labelled tree $\mathcal{T}$ up to the third generation

$\mathcal{A}$. The set of vertices in the n-th generation is denoted by $\mathcal{T}_n$ and the corresponding set of sequences by $\mathcal{A}_n$.

The sequence $a = (a_0, a_1, \ldots, a_n)$ is a prefix of the sequence $b = (b_0, b_1, \ldots, b_m)$ if and only if the vertex v_a with full name a is on the unique path between the root and the vertex v_b with full name b, i.e., if and only if the vertex v_a is an ancestor of the vertex v_b. Consider a graph endomorphism α of $\mathcal{T}$ that fixes the root (and therefore also preserves the levels). Such an endomorphism corresponds to a transformation of sequences $\alpha : \mathcal{A} \to \mathcal{A}$ that preserves the length of the sequences and also their prefix order, i.e.,

$$a \preceq b \qquad \text{implies} \qquad \alpha a \preceq \alpha b,$$

for all sequences a and b in $\mathcal{A}$.

In the sequel, we often deliberately blur the distinction between the vertices in $\mathcal{T}$ and the corresponding sequences in $\mathcal{A}$. Similarly, we do not distinguish tree endomorphisms of $\mathcal{T}$ fixing the root from sequence transformations that preserve the length and the prefix order. This mistake actually improves our presentation.

Let α be an endomorphism of $\mathcal{T}$. Since every generation in $\mathcal{T}$ is finite, the α *orbit*

$$\alpha^* u = \{\alpha^i u \mid i \geq 0\}$$

of every vertex u of $\mathcal{T}$ is finite. Thus, starting from any vertex, repeated applications of α produce *periodic points*, i.e., points a for which $\alpha^k a = a$ for some $k > 0$. The *period* of the periodic point a is the smallest k for which $\alpha^k a = a$. The points of period 1 are *fixed points* and the points of period dividing 2 are *double points*. Obviously, if u and v are periodic points of α and u is a prefix of v then the period of u divides the period of v.

It is easy sometimes to estimate how long it takes before a periodic point is reached. We make use of the *lexicographic ordering* $\leq$ of the sequences in $\mathcal{A}_n$ (note the difference with the prefix ordering $\preceq$). Namely, for $a = (a_0, a_1, \ldots, a_n)$ and $b = (b_0, b_1, \ldots, b_n)$, set $a < b$ if $a_i < b_i$ at the first index where a and b differ.

Theorem 1. *Let α be an endomorphism of the tree $\mathcal{T}$ and assume that, for some $n \geq 1$, there exists $k \geq 1$ such that, for every vertex u in generation n, either*

$$u \leq \alpha^k u \leq \alpha^{2k} u \leq \dots$$

or

$$u \geq \alpha^k u \geq \alpha^{2k} u \geq \dots .$$

Then, starting from any point in generation n, repeated applications of α lead to a periodic point of period dividing k in $O(n^2)$ steps.

Proof. We show that $\beta = \alpha^k$ reaches a fixed point in no more than

$$1 + 2 + \dots + n = n(n+1)/2$$

steps.

Start with any vertex u in generation n. Without loss of generality we may assume

$$u \leq \beta u \leq \beta^2 u \leq \dots .$$

After the first application of β the initial segment up to index 1 of βu is fixed under β. After the next two steps the entry at index 2 will be fixed. Proceeding in the same fashion we see that the initial segment of $\beta^{1+2+\dots+i} u$ up to index i is fixed under β. Indeed, once the initial segment up to index $i-1$ is fixed the entry at index i can go up no more than i times (from 0 to i) before it stabilizes. Thus, $\beta^{1+2+\dots+n} u$ is fixed under β. □

Self-describing sequences

We define an endomorphism $\delta : \mathcal{A} \to \mathcal{A}$ transforming sequences in $\mathcal{A}$ by

$$(\delta a)_i = \#\{j \mid j < i,\ a_j < a_i\}.$$

Thus, for each term t in the sequence a, $(\delta a)_i$ counts the number of previous terms that are smaller than t. The transformation δ makes perfect sense even for sequences out of $\mathcal{A}$, but the image is in $\mathcal{A}$ and it stays there under further iterations. A sequence that is fixed under δ is called a *self-describing sequence*. Therefore, the sequence $a = (a_0, a_1, \dots)$ is self-describing if

$$\#\{j \mid j < i,\ a_j < a_i\} = a_i,$$

for all indices, i.e., every term t is equal to the number of previous terms that are smaller than t.

The Catalan family tree

We describe now a rooted labelled subtree of $\mathcal{T}$, denoted by $\mathcal{C}$ and called *the Catalan family tree* or just the *Catalan family*. The root vertex 0 belongs to $\mathcal{C}$. It has two children named 0 and 1 and we consider 0 the older sibling. The oldest sibling in this family always has 2 children, the second oldest 3, the third oldest 4, and so on. The oldest child of a member of the family x gets named after the oldest sibling of x, the second oldest child after the second oldest sibling, and so on, until x uses its own name for its second to last child and n for the youngest one, where n is the generation number of the children (the level in the tree). The diagram in Figure 2 depicts the family members of $\mathcal{C}$ up to the third generation.

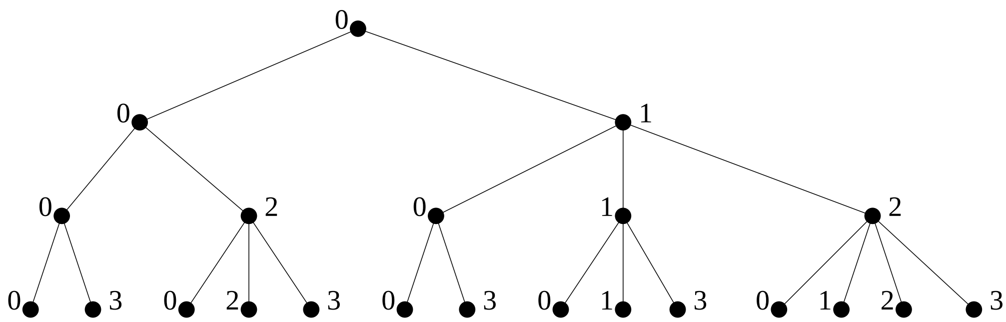

Figure 2: The Catalan family tree $\mathcal{C}$ up to the third generation

The connection

We establish now a connection between the self-describing sequences and the Catalan family tree.

Theorem 2. *The full names of the members of the Catalan family are precisely the self-describing sequences. In other words, they are the fixed points of the endomorphism δ.*

Moreover, repeated applications of δ to any sequence in $\mathcal{A}$ eventually produce a member of the Catalan family, i.e. a fixed point of δ. The number of applications needed to reach such a point is $O(n^2)$.

All statements of the theorem are implied by Theorem 1 and the following lemma.

Lemma 1. *If a is a member of the Catalan family then $a = \delta a$. Otherwise, $a < \delta a$.*

Proof. The proof is by induction on the generation number n. The statement is true for $n = 0$ and $n = 1$. Assume that the statement is true for all vertices up to the n-th generation.

Let

$$a = (a_0, a_1, \dots, a_n, x)$$

be a $(n + 1)$-st generation member of the Catalan family. We consider two cases.

If $x = n + 1$ then

$$\#\{j \mid j < n + 1,\ a_j < x\} = \#\{j \mid j < n + 1,\ a_j < n + 1\} = n + 1 = x,$$

and a is a fixed point of δ.

If $x \neq n+1$, then $a_n \geq x$ and there exists an n-th generation member of the Catalan family whose full name is

$$a' = (a_0, a_1, \ldots, a_{n-1}, x),$$

namely the one after whom a was named. We have

$$\#\{j \mid j < n+1,\ a_j < x\} = \#\{j \mid j < n,\ a_j < x\} = x,$$

where the first equality comes from the fact that $a_n \geq x$ and the second from the inductive hypothesis, since $\delta a' = a'$.

Thus all members of the Catalan family are fixed under δ.

Now, let

$$a = (a_0, a_1, \ldots, a_n, x)$$

be a full name of a vertex in $\mathcal{T}$ in the n-th generation that is not a member of the Catalan family $\mathcal{C}$. If any proper prefix of a is not in $\mathcal{C}$ we obtain the claim directly from the inductive hypothesis. Thus we may assume that

$$a'' = (a_0, a_1, \ldots, a_n)$$

is a member of the Catalan family. Since a is not in $\mathcal{C}$ we have $a_n \neq x$ and $n+1 \neq x$. We consider two cases.

If $a_n > x$ then $a' = (a_0, a_1, \ldots, a_{n-1}, x)$ is not in $\mathcal{C}$ and

$$\#\{j \mid j < n+1,\ a_j < x\} = \#\{j \mid j < n,\ a_j < x\} > x,$$

where the equality comes from the fact that $a_n > x$ and the inequality from the inductive hypothesis.

If $a_n < x < n+1$ then

$$\#\{j \mid j < n+1,\ a_j < x\} = \#\{j \mid j < n,\ a_j < x\} + 1 \geq x+1,$$

where the equality comes from the fact that $a_n < x$ and the inequality from the inductive hypothesis. The equality in the last case is possible only when $a' = (a_0, a_1, \ldots, a_{n-1}, x)$ is in $\mathcal{C}$. □

We proceed by counting the self-describing sequences with fixed length. In addition, we obtain a result on the distribution of names in $\mathcal{C}$. Recall that the n-th Catalan number is equal to

$$c_n = \frac{1}{n+1}\binom{2n}{n}.$$

A recursive definition of the Catalan numbers is given by

$$c_0 = 1,$$
$$c_{n+1} = c_0c_n + c_1c_{n-1} + \cdots + c_nc_0.$$

Theorem 3. *The number of self-describing sequences in $\mathcal{A}_n$, i.e., the number of n-th generation members of the Catalan family is the $(n+1)-th$ Catalan number c_{n+1}.*

Moreover, for $r = 0, \ldots, n$, the number of n-th generation members of the Catalan family whose name is r is equal to $c_r c_{n-r}$.

Proof. Denote by z_n the number of n-th generation members of the Catalan family whose name is 0. More generally, for $r = 0, \ldots, n$ denote by $f_{n,r}$ the number of n-th generation members of the Catalan family whose name is r. Finally, denote by g_n the number of n-th generation members of the Catalan family.

Since the oldest child of every member of the Catalan family is named 0, we have, for all n,

$$z_{n+1} = g_n.$$

Since the youngest sibling in the r-th generation is always named r and the oldest 0 we also have, for all r,

$$f_{r,r} = f_{r,0} = z_r.$$

For some fixed r, consider the set of $f_{r,r}$ r-th generation members named r together with all their descendants in $\mathcal{C}$ whose names are greater or equal to r. This forest of $f_{r,r}$ identical subtrees of $\mathcal{C}$ contains all members of $\mathcal{C}$ whose name is r. Moreover, each tree in this forest looks exactly like the Catalan family tree, except that all labels are increased by r. Indeed, each r-th generation member of $\mathcal{C}$ named r has two children, named r and $r+1$, the oldest sibling always has two children, the second oldest three, etc. Thus, for any n and $r = 0, \ldots, n$, the number $f_{n,r}$ of n-th generation members of $\mathcal{C}$ named r is $f_{r,r}$ times larger than the number of $(n-r)$-th generation members of $\mathcal{C}$ named 0, i.e.,

$$f_{n,r} = f_{r,r} f_{n-r,0} = z_r z_{n-r}.$$

Since $z_0 = 1$ and

$$\begin{aligned} z_{n+1} = g_n &= f_{n,0} + f_{n,1} + \cdots + f_{n,n} \\ &= z_0 z_n + z_1 z_{n-1} + \cdots + z_n z_0 \end{aligned}$$

we conclude that, for all n, z_n is the $n-th$ Catalan number. The statements of the theorem follow now easily from the relations $g_n = z_{n+1}$ and $f_{n,r} = z_r z_{n-r}$. □

Connection to other Catalan trees and objects

It is well known that the Catalan numbers appear naturally under many circumstances. The exercises on Catalan numbers in [Sta99] provide a trove of examples, along with references, in which Catalan numbers count the number of objects of particular type and size. The self-describing sequences provide yet another example that we now relate to some other objects counted by the Catalan numbers.

Consider the sequences in $\mathcal{A}$ with the property that $a_{i+1} \leq a_i + 1$, for all indices (see the Exercise 6.19.u in [Sta99]). Such sequences are called *sequences with unit increase.*

The rooted labelled tree that corresponds to the set of sequences with unit increase looks the same as the Catalan family tree, just with a different labelling and we obtain an easy bijective correspondence between the self-describing sequences and the sequences with unit increase. We could use this bijective connection to show that the Catalan numbers count the number of self-describing sequences. Instead, we provided a direct proof of Theorem 3 and the reason is that there is an important difference in the distribution of labels in the Catalan family tree and the tree of the sequences with unit increase.

Theorem 4. *For $r = 0, \ldots, n$, the number of n-th generation vertices in the tree of sequences with unit increase labelled by r is*

$$\frac{r+1}{n+1}\binom{2n-r}{n}.$$

Proof. Let $a = (a_0, a_1, \ldots, a_n)$ be a sequence with unit increase. Following Exercise 6.19.u in [Sta99], we define, for $i = 0, \ldots, n-1$,

$$b_i = a_i - a_{i+1} + 1.$$

Construct a sequence of n 1's and $n - a_n$ negative 1's by replacing each b_i, $i = 0, \ldots, n-1$ by one 1 followed by b_i negative 1's. The newly obtained sequence has non-negative partial sums. The correspondence between the sequences in $\mathcal{A}_n$ with unit increase that end by r and the sequences of n 1's and $n - r$ negative 1's with non-negative partial sums is bijective. It is shown in [Bai96] that the number of sequences with non-negative partial sums that consist of n 1's and k negative 1's is equal to

$$\frac{n+1-k}{n+1}\binom{n+k}{n}$$

and this implies our claim. □

In passing, we make a slightly more general remark. Namely, for a fixed positive integer m, consider the sequences with the property that $a_0 = 0$ and $0 \leq a_{i+1} \leq a_i + m$, for all indices. Such sequences are called *sequences with m-increase*. We can easily construct the rooted labelled tree that corresponds to such sequences. For a sequence $(a_0, a_1, \ldots, a_n)$ with m-increase, define, for $i = 0, \ldots, n-1$,

$$b_i = a_i - a_{i+1} + m.$$

Following the same approach as before, construct a sequence of n m's and $n - a_n$ negative 1's by replacing each b_i, $i = 0, \ldots, n-1$ by one m followed by b_i negative 1's. The newly obtained sequence has non-negative partial sums and the correspondence between the sequences $(a_0, a_1, \ldots, a_n)$ with m-increase that end by r and the sequences of n 1's and $mn - r$ negative 1's with non-negative partial sums is bijective. Such sequences are discussed in [FS01], where simple recursive formulae for their number is provided.

Unfortunately, closed formulae are not provided yet, but we note that the number of n-th generation sequences with m-increase is given by $c_m(n+1)$ where

$$c_m(n) = \frac{1}{mn+1}\binom{(m+1)n}{n}.$$

The last displayed number is the generalization of the Catalan numbers which counts, for example, the number of rooted $(m+1)$-ary trees with n interior vertices.

It is worth nothing that Julian West [Wes95] recursively constructs a rooted labelled tree whose root is labelled by 2 and each vertex labelled by x has x children labelled by $2, 3, \ldots, x+1$. This tree, which West calls a Catalan tree, looks again exactly like the Catalan family tree, but with different labels. In fact, the tree of the sequences with unit increase can be obtained from the Catalan tree constructed by Julian West by decreasing all labels by 2.

Similarly, in the spirit of the Julian West construction, for any positive integer m, construct a rooted labelled tree whose root is labelled by $m+1$ and each vertex labelled by x has x children labelled by $m+1, m+2, \ldots, m+x$. The tree of sequences with m-increase can be obtained from this tree by decreasing all labels by $m+1$.

Mirror symmetry and mutually describing sequences

We introduce another endomorphism $\gamma : \mathcal{A} \to \mathcal{A}$ transforming sequences in $\mathcal{A}$ by

$$(\gamma a)_i = \#\{j \mid j < i,\ a_j \geq a_i\}.$$

Clearly $\gamma = \mu\delta$ where μ is the *mirror involution* of $\mathcal{A}$ given by

$$(\mu a)_i = i - a_i.$$

We call μ the mirror involution of $\mathcal{A}$ since μ mirrors the tree $\mathcal{T}$ through its vertical axis of symmetry.

The endomorphism γ is studied in [Šun02]. Clearly, γ has no fixed points other than the sequence (0). However, γ has a lot of double points. If a is a double point of γ then so is $b = \gamma a$. Moreover, then $\gamma b = a$ and the sequences a and b mutually describe each other.

Theorem 5 ([Šun02]). *Repeated applications of γ to any sequence in $\mathcal{A}$ eventually produce a double point of γ. The number of application needed to reach a double point in $\mathcal{A}_n$ is $O(n^2)$ and there are more than 2^n such points.*

The sequence that counts the number of double points of γ in the n-th generation starts as follows

$$1, 2, 4, 10, 26, 70, 216, \ldots$$

This sequence does not appear in the Encyclopedia of Integer Sequences [SP95] nor in the online version [Slo] as of January 2002. It is interesting that we have such a good understanding of the fixed points of δ, via the Catalan family tree, but we are still not able to count the number of double points of the mirror related endomorphism $\gamma = \mu\delta$.

Some other endomorphisms leading to fixed or double points are studied in [Šun02]. For one of them, the set of double points of length n is in bijective correspondence with the Young tableaux of size n.

Acknowledgements

Thanks to Richard Stanley and Louis Shapiro for their interest and input.

References

[Bai96] D. F. Bailey, *Counting arrangements of* 1*'s and* −1*'s*, Math. Mag. **69** (1996), no. 2, 128–131.

[FS01] Darrin D. Frey and James A. Sellers, *Generalizing Bailey's generalization of the Catalan numbers*, Fibonacci Quart. **39** (2001), no. 2, 142–148.

[Slo] N. J. A. Sloane, `http://www.research.att.com/~njas/sequences/`.

[SP95] N. J. A. Sloane and Simon Plouffe, *The encyclopedia of integer sequences*, Academic Press Inc., San Diego, CA, 1995.

[Sta99] Richard P. Stanley, *Enumerative combinatorics. Vol. 2*, Cambridge University Press, Cambridge, 1999, With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.

[Šun02] Zoran Šunić, *Young tableaux and other mutually describing sequences*, Journal of Integer Sequences **5** (2002), no. 1, Article 02.1.5.

[Wes95] Julian West, *Generating trees and the Catalan and Schröder numbers*, Discrete Math. **146** (1995), no. 1-3, 247–262.