Enumerating words with forbidden factors

Some things that could be better known

Richard Pinch

Institute of Mathematics and its Applications

07 February 2019

In this talk . . .

- Some facts, well-known and less-well-known, about recurrence relations and power series
- Counting walks on graphs
- Counting words with forbidden factors
- An application

In this talk . . .

- Some facts, well-known and less-well-known, about recurrence relations and power series
- Counting walks on graphs
- Counting words with forbidden factors
- An application

In this talk . . .

- Some facts, well-known and less-well-known, about recurrence relations and power series
- Counting walks on graphs
- Counting words with forbidden factors
- An application

In this talk . . .

- Some facts, well-known and less-well-known, about recurrence relations and power series
- Counting walks on graphs
- Counting words with forbidden factors
- An application

A curiosity?

$$\frac{1}{89} = 0.011235\ldots$$

A curiosity?

$$\frac{1}{89} = 0.011235955056179\ldots$$

A curiosity?

$$
\begin{array}{rl}
\dfrac{1}{89} = & 0.0112358 \\
& \quad\quad\quad\quad\quad 13 \\
& \quad\quad\quad\quad\quad\quad 21 \\
& \quad\quad\quad\quad\quad\quad\quad 34 \\
& \quad\quad\quad\quad\quad\quad\quad\quad 55 \\
& \quad\quad\quad\quad\quad\quad\quad\quad\quad \ddots
\end{array}
$$

A curiosity?

Let F_n denote the nth Fibonacci number (with $F_0 = 0$).

$$\frac{1}{89} = \sum_{n=0}^{\infty} F_n 10^{-n-1} .$$

A curiosity?

This is simply the power series expansion

$$\frac{z}{1 - z - z^2} = \sum_{n=0}^{\infty} F_n z^n$$

evaluated at $z = \frac{1}{10}$. The occurence of the polynomial $1 - z - z^2$ should come as no surprise, as it is the driving polynomial of the Fibonacci sequence.

Not a curiosity

The power series with coefficients which satisfy a linear recurrence relation are precisely the formal power series attached to rational functions: indeed, the denominator of the rational function will be the driving polynomial of the linear recurrence.

Not a curiosity

The polynomials over a field K form a K-algebra with dual the space of infinite sequences.
The linear recurrent sequences form the *restricted* dual, and hence have a coalgebra structure.
Indeed, the polynomials and the linear recurrent sequences form bialgebras and in two different ways.

Generating functions

Let (r_n) be a sequence defined by a recurrence relation

$$r_{n+k} = c_{k-1} r_{n+k-1} + \cdots + c_0 r_n$$

for $n \geq 0$, with $c_0 \neq 0$, subject to given initial conditions on $r_0, \ldots, r_{k-1}$. Let $f(x)$ be the auxiliary polynomial

$$f(x) = x^k - \left(c_{k-1} x^{k-1} + \cdots + c_0\right) .$$

Let the roots of f be $\alpha_1, \ldots \alpha_k$: we shall assume for simplicity of this exposition that f has no repeated roots.

Generating functions

The generating function as a formal power series is equal to a rational function of the form

$$\sum_{n=1}^{\infty} r_n z^n = \frac{p(z)}{\tilde{f}(z)}$$

where p is a polynomial of degree $\leq k-1$, determined by the initial conditions, and $\tilde{f}$ is the *reciprocal* polynomial

$$\tilde{f}(z) = \prod_{i=1}^{k} (1 - \alpha_i z) = (-1/c_0) z^k f(1/z) .$$

Traces

We may also express any such generating function as

$$\sum_{i=1}^{k} \frac{\beta_i}{1 - \alpha_i z},$$

that is, as a *trace*

$$\operatorname{tr}\left(\frac{\beta}{1 - \alpha z}\right) .$$

Impulse response sequence

The *impulse response sequence* has initial values $0, \ldots, 0, 1$, so that every sequence is a linear combination of this and its left shifts. The corresponding generating function has initial term z^{k-1}.

We have

$$\operatorname{tr}\frac{1/f'(\alpha)}{(1-\alpha z)} = z^{k-1} + O(z^k)$$

as the generating function of the IRS in trace form.

Walks on graphs

Let $G = (V, E)$ be a directed graph with possibly multiple edges, and A the adjacency matrix with entry A_{ij} counting the number of edges from vertex v_i to vertex v_j. Then the entries of the matrix power A^n, for $n \geq 0$, count the number of directed paths of length n between vertices.

Walks on graphs

We can regard the formal power series

$$\sum_{n \geq 0} A^n z^n$$

as either a formal power series over the matrix ring or as a matrix whose entries are formal power series.

Generating function

Let $G = (V, E)$ be a directed graph with possibly multiple edges, and
It is then easy to see that

$$(I - Az)\left(\sum_{n\geq 0} A^n z^n\right) = I$$

and since $I - Az$ is invertible, this matrix formal power series is an expression for the inverse $(I - Az)^{-1}$.

Formal power series

We are considering the formal term z^n as encoding paths of length n. Let B be a matrix in which every entry is a polynomial in z with integer coefficients and with zero constant term: so that Az is an example of such a matrix. Then the sum

$$\sum_{n \geq 0} B^n$$

converges (formally speaking in the z-adic topology) to a formal power series in z over the ring of matrices.

Counting paths

If we want to count the number of paths of length n from a given vertex i to vertex j, then we write

$$e_i^\top \left(\sum_{n \geq 0} A^n z^n \right) e_j$$

where e_i is the coordinate vector for the i-th position.

Counting paths

We have the generating function

$$e_i^\top (I - Az)^{-1} e_j$$

and expanding $(I - Az)^{-1}$ we obtain

$$\frac{1}{\det(1 - Az)} e_i^\top \mathrm{adj}(I - Az) e_j$$

which is a rational function with the characteristic polynomial of A as denominator.

Forbidden factors

We give as a worked example the problem of counting words over a finite alphabet with forbidden factors: that is, words with no consecutive substring drawn from a finite forbidden set.

An application

The specific application that motivated us to look at this problem comes from the problem of counting binary words in *prefix normal binary words*: words over the alphabet $\{\mathbf{0}, \mathbf{1}\}$ with the property that no factor has more occurrences of the symbol $\mathbf{1}$ than the prefix of the same length. This problem is addressed, for example, by Burcsi et al. They obtain an upper bound of the form $2^{n-\lg n+1}$ by observing that a binary word in prefix form for which the first k symbols are an initial run of $\mathbf{1}^{k-1}$ followed by a $\mathbf{0}$ necessarily has the property that it has no factor of the form $\mathbf{1}^k$.

de Bruijn graphs

We wish to enumerate words whose factors of length k come from a permissible set S. The *de Bruijn graph* or state machine G_S has vertex set S. There is a directed edge from a vertex $xY \in S$ to a vertex $Yz \in S$, where x, z denote single letters and Y denotes a string of length $k-1$: following an edge represents the movement one place of a k-long window into the word. This reduces the problem to one of enumerating walks in the graph G_S.

de Bruijn graphs

We wish to enumerate words whose factors of length k come from a permissible set S. The *de Bruijn graph* or state machine G_S has vertex set S. There is a directed edge from a vertex $xY \in S$ to a vertex $Yz \in S$, where x, z denote single letters and Y denotes a string of length $k-1$: following an edge represents the movement one place of a k-long window into the word. This reduces the problem to one of enumerating walks in the graph G_S.

Words without runs

We illustrate the application of power series to obtain this upper bound for binary words with no factor $\mathbf{1}^k$.

The de Bruijn graph has $2^k - 1$ vertices, labelled by all binary strings except $\mathbf{1}^k$, and each vertex xY has two arrows to vertices $Y\mathbf{0}$ and $Y\mathbf{1}$: except for the vertex $\mathbf{01}^{k-1}$ which has only a single arrow, to $\mathbf{1}^{k-1}\mathbf{0}$.

Words without runs

Fix $k \geq 2$ and consider the number $F(n)$ of binary words of length n with no runs of k consecutive 1. We know that F satisfies a linear recurrence relation which we can obtain from its de Bruijn diagram. In this case, it is easy to see that F satisfies the recurrence

$$F(n) = F(n-1) + \cdots + F(n-k)$$

with initial conditions $F(i) = 2^i$, $i = 0, \ldots, k-1$.

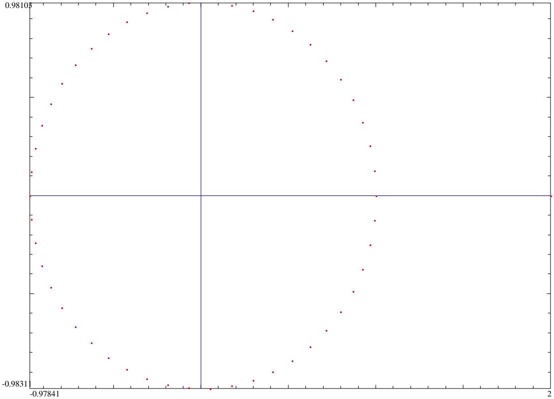
The characteristic equation

Let $f(z) = z^k - (z^{k-1} + \cdots 1)$ be the auxiliary polynomial of this recurrence. It is clear that f has no roots of absolute value greater than or equal to 2. Let α_1 denote the largest real root.

Theorem

- *The dominant root satisfies* $\alpha_1 < 2 - \frac{1}{2^k}$
- f *has no repeated roots*
- *All roots other than* α_1 *are inside the unit circle.*

A picture



The characteristic equation

This already gives us the asymptotic behaviour of F. For a more explicit bound we need to locate the other roots and to estimate $f'(\alpha)$ as α runs over the roots of f.

Theorem

The roots of f within the unit circle are at distance of order $1/k$ from the $k-1$ non-trivial k-th roots of unity.

Conclusion

We conclude that $f'(\alpha) \sim k$, for α one of the roots inside the unit circle. Hence we can obtain an estimate strong enough to give Theorem 15 of [BFLRS]

$$\sum_{k=1}^{n} F_k(n-k) < 2^{n-\lg n+1} ,$$

Summary

- Recurrent sequences and the associated power series have a rich natural algebraic structure
- This structure is useful in a wide range of applications in combinatorics

Summary

- Recurrent sequences and the associated power series have a rich natural algebraic structure
- This structure is useful in a wide range of applications in combinatorics

The end

Questions?

Comments!