GEOMETRIC DISTRIBUTIONS AND FORBIDDEN SUBWORDS

Helmut Prodinger*
Department of Algebra and Discrete Mathematics, Technical University of Vienna, Austria
(*Submitted July 1993*)

In a recent paper [1] Barry and Lo Bello dealt with the moment generating function of the geometric distribution of order k. I want to draw the attention of the *Fibonacci Community* to several related papers that were apparently missed by the authors and also to provide a straightforward derivation of their result.

Since the moment generating function $M(t)$ is related to the probability generating function $f(z)$ by $M(t) = f(e^t)$, it is sufficient to consider $f(z)$.

We code a success trial by **1** and a failure by **0**, thereby obtaining a *word* consisting of the letters **0** and **1**. A sequence of n trials is thus represented by a *word* of length n over the *alphabet* $\{\mathbf{0}, \mathbf{1}\}$. In a natural way we attach a *weight* ω to each word x by replacing **1** by p and **0** by q and then multiplying as usual. For instance, the word **0110** has the weight p^2q^2. We consider *languages* (sets of words) L and their *generating function* $\ell(z)$. The latter is defined to be

$$\ell(z) = \sum_{x \in L} \omega(x) z^{|x|}, \tag{1}$$

where $|x|$ is the *length* (number of letters) of the word x. This generating function can be obtained simply by formally replacing the letter **1** by pz and **0** by qz in the language L and replacing the so-called *concatenation* of words by the usual product and the (disjoint) *union* by the usual addition so that, for instance, $L = \{\mathbf{0}, \mathbf{010}, \mathbf{0110}\}$ has the generating function $\ell(z) = qz + pq^2z^3 + p^2q^2z^4$.

Instead of considering $\mathbb{P}\{X = n\}$, it is easier to consider $\mathbb{P}\{X > n\}$; that means the probability that n trials did not produce k consecutive successes, or the probability that a random word of n letters does not contain the (contiguous) subword $\mathbf{1}^k$. We consider the language of these words. A compact notion of it is

$$(\mathbf{1}^{<k}\,\mathbf{0})^*\mathbf{1}^{<k}, \tag{2}$$

where $\mathbf{1}^{<k} = \{\varepsilon, \mathbf{1}, \mathbf{11}, \ldots, \mathbf{1}^{k-1}\}$, with ε being the empty word. This expresses the fact that words without the (contiguous) subword $\mathbf{1}^k$ can be written as several blocks of less than k ones, separated by zeros. Let us recall that the asterisk L^* describes sequences of L. More formally, $L^* = \bigcup_{n\geq 0} L^n$, and L^n means the concatenation of n copies of L, which can be defined recursively by $LL = \{xy \mid x \in L, y \in L\}$ and $L^n = L^{n-1}L$ and $L^0 = \{\varepsilon\}$. Quite nicely, the generating function of L^* is obtained by $\frac{1}{1-\ell(z)}$. Now, to the language $\mathbf{1}^{<k}\,\mathbf{0}$ the generating function

$$(1 + pz + (pz)^2 + \cdots + (pz)^{k-1}) \cdot qz = \frac{1 - p^k z^k}{1 - pz} qz \tag{3}$$

is associated, and thus we have, furthermore,

*This note was written while the author visited the University Paris 6; he is thankful for the warm hospitality he encountered there.

$$g(z)=\sum_{n\geq 0}\mathbb{P}\{X>n\}z^n=\frac{1}{1-qz\frac{1-p^kz^k}{1-pz}}\cdot\frac{1-p^kz^k}{1-pz}=\frac{1-p^kz^k}{1-z+qp^kz^{k+1}}. \tag{4}$$

From this we also obtain the probability generating function

$$\begin{aligned} f(z)&:=\sum_{n\geq 0}\mathbb{P}\{X=n\}z^n=\sum_{n\geq 0}\left(\mathbb{P}\{X>n-1\}-\mathbb{P}\{X>n\}\right)z^n \\ &=1+z\sum_{n\geq 1}\mathbb{P}\{X>n-1\}z^{n-1}-\sum_{n\geq 0}\mathbb{P}\{X>n\}z^n \\ &=1-(1-z)g(z)=\frac{1-z+qp^kz^{k+1}-1+p^kz^k+z-p^kz^{k+1}}{1-z+qp^kz^{k+1}} \\ &=\frac{p^kz^k(1-pz)}{1-z+qp^kz^{k+1}}. \end{aligned} \tag{5}$$

This derivation completely avoided unpleasant recursions. For such very useful combinatorial constructions and their automatic translation into generating functions, we refer to the survey [2] and a few earlier survey papers of Flajolet cited therein.

The probability generating function (5) appeared first in [10].

Guibas and Odlyzko in a series of papers ([3], [4], [5]) dealt with general forbidden subwords, not just $\mathbf{1}^k$. These papers were surveyed in [8] and [9]. Rewriting things accordingly, formula (6.44) in [9] gives

$$f(z)=\frac{(pz)^k}{(pz)^k+(1-z)C(z)}, \tag{6}$$

where the polynomial $C(z)$ (the "correlation polynomial") depends on the forbidden pattern and is

$$C(z)=1+(pz)+\cdots+(pz)^{k-1}=\frac{1-(pz)^k}{1-pz} \tag{7}$$

in this special instance.

Knuth used similar arguments in [7]. He considered strings of **0**, **1**, **2**, where **0** and **2** appear with probability 1/4 and **1** appears with probability 1/2 and the string $\mathbf{1}^k\mathbf{2}$ is forbidden. Also, he considered the zeros of the "auxiliary equation"

$$1-z+qp^kz^{k+1}=0. \tag{8}$$

For example, there is a "dominant" solution $\rho=\rho_k$ which can be approximated by "bootstrapping": Starting from $z=1+qp^kz^{k+1}$, a first approximation is $\rho\approx 1$. Inserting this on the right-hand side and expanding, we find $\rho\approx 1+qp^k$, and after one more step,

$$\rho\approx 1+qp^k+(k+1)q^2p^{2k}, \tag{9}$$

etc. Kirschenhofer and Prodinger also used this type of argument in [6].

With this dominant singularity it is also easy to find the asymptotics of $\mathbb{P}\{X=n\}$ for fixed k, as $n\to\infty$. We have

$$f(z)=\frac{p^kz^k(1-pz)}{1-z+qp^kz^{k+1}}\sim\frac{A_k}{1-z/\rho}\quad\text{as } z\to\rho. \tag{10}$$

This can be explained informally by saying that *locally* only one term of the *partial fraction decomposition* of the *rational function f(z)* is needed to describe its behavior in a vicinity of the dominant singularity ρ.

Here, A_k is a constant that can be found by the traditional techniques to compute the partial fraction decomposition of a rational function.

Thus, the coefficient of z^n in $f(z)$ (i.e., $\mathbb{P}\{X = n\}$) behaves as $A_k \cdot \rho^{-n}$ (the coefficient of z^n in $\frac{A_k}{1-z/\rho}$). The constant A_k behaves as $A_k \approx qp^k$ for $k \to \infty$.

Such asymptotic considerations are to be found in many textbooks and survey articles, notably in [9].

REFERENCES

1. M. J. J. Barry & A. J. Lo Bello. "The Moment Generating Function of the Geometric Distribution of Order *k*." *The Fibonacci Quarterly* **31.2** (1993):178-80.
2. P. Flajolet & J. Vitter. "Analysis of Algorithms and Data Structures." In *Handbook of Theoretical Computer Science*, Vol. A, pp. 432-524. Amsterdam: North Holland, 1990.
3. L. J. Guibas & A. M. Odlyzko. "Maximal Prefix-Synchronized Codes." *SIAM J. Appl. Math.* **35** (1978):401-18.
4. L. J. Guibas & A. M. Odlyzko. "Long Repetitive Patterns in Random Sequences." *Z. Wahrscheinlichkeitstheorie und verwandte Gebiete* **53** (1980):241-62.
5. L. J. Guibas & A. M. Odlyzko. "String Overlaps, Pattern Matching, and Nontransitive Games." *J. Comb. Theory A* **30** (1981):183-208.
6. P. Kirschenhofer & H. Prodinger. "A Coin Tossing Algorithm for Counting Large Numbers of Events." *Mathematica Slovaca* **42** (1992):531-45.
7. D. E. Knuth. "The Average Time for Carry Propagation." *Indagationes Mathematicæ* **40** (1978):238-42.
8. A. M. Odlyzko. "Enumeration of Strings." In *Combinatorial Algorithms on Words*, pp. 205-28. Ed. A. Apostolico & Z. Galil. New York: Springer, 1985.
9. A. M. Odlyzko. "Asymptotic Enumeration Methods." In *Handbook of Combinatorics.* Amsterdam: North Holland, to appear.
10. A. N. Philippou, C. Georghiou, & G. N. Philippou. "A Generalized Geometric Distribution and Some of Its Properties." *Statistics and Probability Letters* **1.4** (1983):171-75.

AMS Classification Number: 05A15