GEOMETRIC DISTRIBUTIONS AND FORBIDDEN SUBWORDS

Helmut Prodinger

Department of Algebra and Discrete Mathematics,
Technical University of Vienna

Abstract. We describe in this tutorial note how the geometric distribution of order k can be seen in context of enumeration problems of strings.

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 functing $M(t)$ 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 $\mathbf{1}$ and a failure by $\mathbf{0}$, thereby obtaining a *word* consisting of the letters $\mathbf{0}$ and $\mathbf{1}$. 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 $\mathbf{1}$ by p and $\mathbf{0}$ by q and then multiplying as usual. For instance, the word $\mathbf{0110}$ has the weight p^2q^2. We consider *languages* (sets of words) L and their *generating function* $l(z)$. The latter is defined to be

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

where $|x|$ is the *length* (number of letters) of the word x. This generating function can be simply obtained by formally replacing the letter $\mathbf{1}$ by pz and $\mathbf{0}$ 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 $l(z)=qz+pq^2z^3+p^2q^2z^4$.

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

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

1991 *Mathematics Subject Classification.* 05A15.

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

Typeset by $\mathcal{AMS}$-TEX

where $\mathbf{1}^{<k} = \{\varepsilon, \mathbf{1}, \mathbf{11}, \dots, \mathbf{1}^{k-1}\}$, 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 zeroes. Let us recall that the asterisk L^* describes sequences of L. More formally, $L^* = \bigcup_{n\ge 0} L^n$, and L^n means the concatenation of n copies of L, which can be recursively defined by $LL = \{xy | x \in L, y \in L\}$ and $L^n = L^{n-1}L$ and $L^0 = \{\varepsilon\}$. Quite nicely, the generating function of L^* is obtained by $\frac{1}{1-l(z)}$. Now, to the language $\mathbf{1}^{<k}\mathbf{0}$ the generating function

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

is associated, and thus we have furthermore

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

From this we also obtain the probability generating function

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

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

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

in this special instance.

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

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

For example, there is a "dominant" solution $\rho = \rho_k$ which can be approximated by "bootstrapping": Starting from $z = 1 + qp^k z^{k+1}$, 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

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

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 and A.J. Lo Bello, *The moment generating function of the geometric distribution of order k*, The Fibonacci Quarterly **31** (1993), 178–180.
2. P.Flajolet and J.Vitter, *Analysis of algorithms and data structures*, Handbook of Theoretical Computer Science, Vol.A., North Holland, 1990, pp. 432–524.
3. L.J.Guibas and A.M.Odlyzko, *Maximal prefix-synchronized codes*, SIAM J. Appl. Math. **35** (1978), 401–418.
4. L.J.Guibas and A.M.Odlyzko, *Long repetitive patterns in random sequences*, Z. Wahrscheinlichkeitstheorie und verwandte Gebiete **53** (1980), 241–262.
5. L.J.Guibas and A.M.Odlyzko, *String overlaps, pattern matching, and nontransitive games*, J. Comb. Theory A **30** (1981), 183–208.
6. P.Kirschenhofer and H.Prodinger, *A coin tossing algorithm for counting large numbers of events*, Mathematica Slovaca **42** (1992), 531–545.
7. D.E.Knuth, *The average time for carry propagation*, Indagationes Mathematicæ **40** (1978), 238–242.
8. A.M.Odlyzko, *Enumeration of strings*, Combinatorial Algorithms on Words, A.Apostolico and Z.Galil, eds., Springer, 1985, pp. 205–228.
9. A.M.Odlyzko, *Asymptotic enumeration methods*, in: Handbook of Combinatorics,, North Holland (to appear).
10. A.N.Phillippou, C.Georghiu, and G.N.Philippou, *A Generalized Geometric Distribution and Some of its Properties*, Statistics and Probability Letters **1.4** (1983), 171–175.

Department of Algebra and Discrete Mathematics
Technical University of Vienna
Wiedner Hauptstrasse 8–10
A-1040 Vienna
Austria

AMS classification: 05A15