SET PARTITIONS, WORDS, AND APPROXIMATE COUNTING WITH BLACK HOLES

HELMUT PRODINGER

Abstract. Words satisfying the restricted growth property $w_k \le 1+\max\{w_0,\dots,w_{k-1}\}$ are in correspondence with set partitions. Underlying the geometric distribution to the letters, these words are enumerated with respect to the largest letter occurring, which corresponds to the number of blocks in the set partition. It turns out that on average, this parameter behaves like $\log_{1/q} n$, where q is the parameter of the geometric distribution.

1. Introduction

Set partitions of $\{1,\dots,n\}$ can be coded by words $w_1\dots w_n$, where the letters are positive integers, and (with $w_0 = 0$) the *restricted growth property* $w_k \le 1 + \max\{w_0,\dots,w_{k-1}\}$ holds for $1 \le k \le n$, compare [8].

In [9], a study of such words was started where the letters where equipped with geometric probabilities pq^{k-1} for $k = 1, 2, \dots$ and $p + q = 1$. If letters are drawn independently, then for $P(n)$, the probability that a word of length n has the restricted growth property, an explicit and an asymptotic formula where found:

$$P(n) = p\sum_{j=0}^{n-1}(-1)^j\binom{n-1}{j}q^j(p;q)_j = \sum_{j=0}^{n}(-1)^j\binom{n}{j}(p;q)_j,$$

with $(x;q)_n = (1-x)(1-xq)\dots(1-xq^{n-1})$. This notation is used also for $(x;q)_\infty$ when the product is extended to infinity, see [1]. Further, with $b = \log p/\log q$, $Q = \frac{1}{q}$, $L := \log Q$, $\chi_k = \frac{2\pi i k}{L}$,

$$P(n) = \frac{(p;q)_\infty}{L(q;q)_\infty}\Gamma(b)n^{-b} + n^{-b}\frac{(p;q)_\infty}{L(q;q)_\infty}\sum_{k\neq 0}\Gamma(b+\chi_k)e^{-2\pi i k\cdot\log_Q n} + O(n^{-b-1}).$$

Note that the series is a Fourier series and represents a (small) periodic function.

A convenient way to imagine the evolution of these probabilities is by a state diagram as in Figure 1:

Date: March 20, 2012.

1991 *Mathematics Subject Classification.* 05A16, 05A15, 05A18.

Key words and phrases. Set partitions, words, geometric probabilities, approximate counting, Rice's method.

This author is supported by an incentive grant from the NRF of South Africa. This note was prepared while he was a guest at the Academia Sinica in Taipei, Taiwan. The hospitality is gratefully acknowledged.

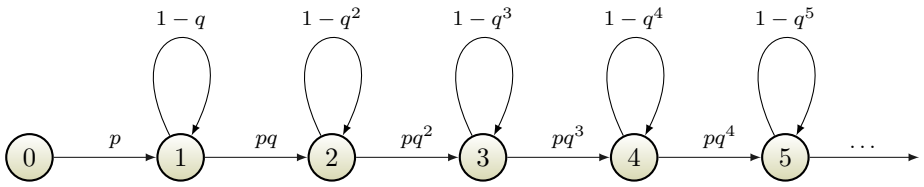

FIGURE 1. State diagram of the evolution of the probabilities of geometrically distributed words satisfying the restricted growth property.

The states (labelled $0, 1, \dots$) represent the largest letter seen so far, and the restricted growth property only allows to stay in such a state or advance to the state with label one higher. The edges are labelled with the probabilities to either remain in a state or advance to the next one. Note that the sum of them, $1-q^k+pq^k = 1-q^{k+1} < 1$, which means that in each state there is a chance to violate the restricted growth property. We call this *falling into a black hole*. Reading further letters does not help; there is no escape from a black hole.

In the recent paper [8], the analysis was extended, by computing, *inter alia*, the probability $P(n,k)$ to end up in state k after n random letters. Naturally, $P(n) = \sum_k P(n,k)$. Although it was not mentioned explicitly, the method to derive $P(n,k)$ is one that is very common when one analyzes *Digital Search Trees*, see [3]. It is worthwhile to mention that k represents the number of blocks of the set partition corresponding to the word with the restricted growth property.

The state diagram as mentioned is, however, not uncommon in the literature. It resembles the one appearing in *approximate counting*:

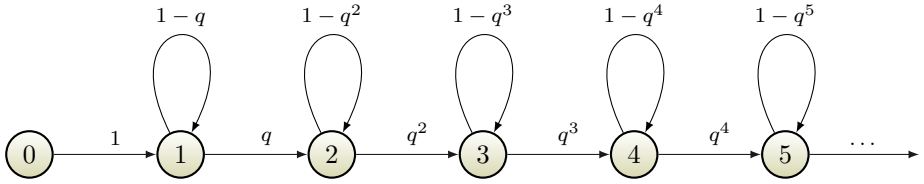

FIGURE 2. State diagram of the evolution of the counter in *approximate counting*. Normally, the (deterministic) initial step from state 0 to state 1 is not shown.

In approximate counting, one is interested in the state one is in after n random steps. This state number is interpreted as the value of a certain counter. This model was first analyzed by Flajolet in [2] and remains popular to this day; [6] is a recent example, and it contains many backward pointers to the literature. Flajolet derived explicit expressions to reach state k after n random steps, and computed expectation and variance of this parameter.

Since we know so much about approximate counting and the situation is so similar, it should be possible to gain some additional insight to the analysis of the restricted growth model. This is indeed the motivation for the present note: While approximate counting might be known to some people in Theoretical Computer Science, it is most

likely new to the Combinatorics Community who reads the Australasian Journal of Combinatorics. As a general reference for the methods we are using we cite the book [5] which represents the state of the art in *analytic combinatorics* and its applications to the *analysis of algorithms*.

The discussion given thus far motivates the catchy nickname *approximate counting with black holes*; for completeness, one might think about an extra state (the black hole) in which one falls from each other state k with the appropriate probability q^{k+1}, when the restricted growth property is violated for the first time. And, naturally, one cannot escape from the black hole.

The quantities

$$\frac{P(n,k)}{P(n)}$$

sum to 1 and define a probability distribution. We will in the next section (re)derive an explicit formula for $P(n,k)$ and then study the average value of the probability distribution just described. This answers the question "Which state does one reach, on average, after n random steps, subject to the condition that the random word satisfies the restricted growth property?" Higher moments could also be studied, but we refrain from doing that in order to keep this note short and crisp.

We cite here the average value C_n in approximate counting: Flajolet only gave it for $p = q = \frac{1}{2}$, but the analysis is easily extended:

$$C_n = \log_Q n - \alpha + \frac{\gamma}{L} + \frac{1}{2} - \frac{1}{L}\sum_{k\neq 0}\Gamma(\chi_k)e^{-2\pi i k\cdot\log_Q n} + O(\frac{1}{n}),$$

where γ is Euler's constant and

$$\alpha = \sum_{k\geq 1}\frac{q^k}{1-q^k}.$$

Within this range of accuracy, it does not matter whether one has one extra step in the beginning or not.

2. Analysis of approximate counting with black holes

In [9] is was proved that

$$P(n,k) = pq^{k-1}P(n-1,k-1) + (1-q^k)P(n-1,k), \quad P(0,0) = 1.$$

This recursion is immediately understood when looking at the state diagram and how one can reach state k using the nth letter.

In the relatively recent paper [7], we collected many old and new results and techniques. We will rederive a formula for $P(n,k)$ following a technique presented there. Set

$$F(z,u) = \sum_{n,k} z^n u^k P(n,k),$$

then the recursion translates into the functional equation

$$F(z,u) - 1 = pzuF(z,qu) + zF(z,u) - zF(z,qu).$$

We want to solve this equation by *iteration*. However, since $F(z,u) = 1 + \cdots$, this does not work. Thus we define $G(z,u) := F(z,u) - 1$ and rewrite the equation:

$$G(z,u) = pzu + pzuG(z,qu) + zG(z,u) - zG(z,qu),$$

which can now be iterated:

$$\begin{aligned} G(z,u) &= \frac{pzu}{1-z} + \frac{z(pu-1)}{1-z} G(z,qu) \\ &= \frac{pzu}{1-z} + \frac{z(pu-1)}{1-z}\left[\frac{pqzu}{1-z} + \frac{z(pqu-1)}{1-z} G(z,q^2u)\right] \\ &= \frac{pzu}{1-z} + \frac{z(pu-1)}{1-z}\left[\frac{pqzu}{1-z} + \frac{z(pqu-1)}{1-z}\left[\frac{pq^2zu}{1-z} + \frac{z(pq^2u-1)}{1-z} G(z,q^3u)\right]\right] \\ &= \cdots = pu \sum_{j\ge 0} \frac{z^{j+1}(-1)^j (pu;q)_j q^j}{(1-z)^{j+1}}. \end{aligned}$$

From this we find for $n \ge 1$

$$[z^n]G(z,u) = pu \sum_{j=0}^{n-1} \binom{n-1}{j} (-1)^j (pu;q)_j q^j = \sum_{j=0}^{n} \binom{n}{j} (-1)^j (pu;q)_j.$$

Apart from the normalization by dividing this by $P(n)$, this is a probability generating function.

Furthermore, we have for $k \ge 1$

$$\begin{aligned} [z^n u^k]G(z,u) &= P(n,k) \\ &= p^k \sum_{j=0}^{n-1} \binom{n-1}{j} (-1)^j [u^{k-1}](u;q)_j q^j \\ &= p^k \sum_{j=0}^{n-1} \binom{n-1}{j} (-1)^j q^j \sum_{t=0}^{k-1} \frac{(-1)^t q^{\binom{t}{2}} q^{(k-1-t)j}}{(q;q)_t (q;q)_{k-1-t}} \\ &= p^k \sum_{j=0}^{n-1} \binom{n-1}{j} (-1)^j \sum_{t=0}^{k-1} \frac{(-1)^t q^{\binom{t}{2}} q^{(k-t)j}}{(q;q)_t (q;q)_{k-1-t}} \\ &= p^k \sum_{t=0}^{k-1} \frac{(-1)^t q^{\binom{t}{2}}}{(q;q)_t (q;q)_{k-1-t}} (1-q^{k-t})^{n-1}; \end{aligned}$$

in the second line a formula due to Rothe [1] has been used.

This formula is equivalent to Theorem 3.1 from [8] and provides an explicit formula for $P(n,k)$.

To find the expected value as promised, one starts from

$$[z^n]G(z,u) = \sum_{j=0}^{n} \binom{n}{j} (-1)^j (pu;q)_j,$$

differentiates this with respect to u, and then plugs in $u = 1$, with the result

$$g(n) := \sum_{j=0}^{n} \binom{n}{j} (-1)^j (p;q)_j \left[-\alpha_p + \sum_{l \geq j} \frac{pq^l}{1-pq^l} \right].$$

Here, we use the abbreviation

$$\alpha_p = \sum_{l \geq 0} \frac{pq^l}{1-pq^l}.$$

We rewrite this as

$$\begin{aligned} g(n) &= \sum_{j=0}^{n} \binom{n}{j} (-1)^j \frac{(q^b;q)_\infty}{(q^{b+j};q)_\infty} \left[\frac{q^{b+j}}{1-q^{b+j}} - \alpha_p + \sum_{l \geq 1} \frac{q^{l+b+j}}{1-q^{l+b+j}} \right] \\ &= \sum_{j=0}^{n} \binom{n}{j} (-1)^j \frac{(q^b;q)_\infty}{(q^{b+j};q)_\infty} \frac{q^{b+j}}{1-q^{b+j}} \\ &\quad + \sum_{j=0}^{n} \binom{n}{j} (-1)^j \frac{(q^b;q)_\infty}{(q^{b+j};q)_\infty} \left[-\alpha_p + \sum_{l \geq 1} \frac{q^{l+b+j}}{1-q^{l+b+j}} \right] =: E_1 + E_2. \end{aligned}$$

There is a convenient technique to handle such alternating sums; it is called *Rice's method*, and is described at length in [4]:

$$E_1 = -\frac{1}{2\pi i} \int_{\mathcal{C}} \frac{\Gamma(n+1)\Gamma(-z)}{\Gamma(n+1-z)} \frac{(q^b;q)_\infty}{(q^{b+1+z};q)_\infty} \frac{q^{b+z}}{(1-q^{b+z})^2} dz,$$

where the curve $\mathcal{C}$ encloses the poles at $0, 1, \ldots, n$ and no others. To find asymptotics, one extends the curve and has, as a compensation, to subtract the extra residues that one encounters. In our case the relevant ones are at $z = -b - \chi_k$, $k \in \mathbb{Z}$, and we must find the residues of

$$\frac{\Gamma(n+1)\Gamma(-z)}{\Gamma(n+1-z)} \frac{(q^b;q)_\infty}{(q^{b+1+z};q)_\infty} \frac{q^{b+z}}{(1-q^{b+z})^2}$$

at these poles. With $w = z + b + \chi_k$, we must expand

$$\frac{\Gamma(n+1)\Gamma(-w+b)}{\Gamma(n+1-w+b)} \frac{(q^b;q)_\infty}{(q^{w+1};q)_\infty} \frac{q^w}{(1-q^w)^2}$$

to two terms around $w = 0$. So we compute

$$\begin{aligned} &[w^{-1}] \frac{\Gamma(n+1)\Gamma(-w+c)}{\Gamma(n+1-w+c)} \frac{(p;q)_\infty}{(q;q)_\infty} [1 - L\alpha w] \frac{1}{L^2 w^2} \\ &= -L\alpha \frac{\Gamma(n+1)\Gamma(c)}{\Gamma(n+1+c)} \frac{(p;q)_\infty}{(q;q)_\infty} \frac{1}{L^2} + [w^1] \frac{\Gamma(n+1)\Gamma(-w+c)}{\Gamma(n+1-w+c)} \frac{(p;q)_\infty}{(q;q)_\infty} \frac{1}{L^2} \\ &= -\alpha \frac{\Gamma(n+1)\Gamma(c)}{\Gamma(n+1+c)} \frac{(p;q)_\infty}{(q;q)_\infty} \frac{1}{L} - \frac{\Gamma(n+1)\Gamma'(c)}{\Gamma(n+1+c)} \frac{(p;q)_\infty}{(q;q)_\infty} \frac{1}{L^2} \\ &\quad + \frac{\Gamma(n+1)\Gamma(c)\psi(n+1+c)}{\Gamma(n+1+c)} \frac{(p;q)_\infty}{(q;q)_\infty} \frac{1}{L^2}, \end{aligned}$$

with $\psi = \Gamma'/\Gamma$. We have chosen the letter c to represent $b + \chi_k$.

Asymptotically, as $n \to \infty$:

$$-\alpha n^{-c}\Gamma(c)\frac{(p;q)_\infty}{(q;q)_\infty}\frac{1}{L} - n^{-c}\Gamma'(c)\frac{(p;q)_\infty}{(q;q)_\infty}\frac{1}{L^2} + n^{-c}\Gamma(c)\log_Q n\frac{(p;q)_\infty}{(q;q)_\infty}\frac{1}{L}.$$

These terms must be summed over $k \in \mathbb{Z}$, and we see that in the main term the same periodic oscillation appears that was already present in the study of $P(n)$ itself.

After normalization (division by the asymptotic equivalent of $P(n)$, mentioned earlier):

$$\log_Q n - \alpha + \textit{oscillation}.$$

We also have to consider the term

$$E_2 = \sum_{j=0}^{n}\binom{n}{j}(-1)^j\frac{(q^b;q)_\infty}{(q^{b+j};q)_\infty}\left[-\alpha_p + \sum_{l\ge 1}\frac{q^{l+b+j}}{1-q^{l+b+j}}\right],$$

which means that we have to compute the residues of

$$\frac{\Gamma(n+1)\Gamma(-z)}{\Gamma(n+1-z)}\frac{(q^b;q)_\infty}{(q^{b+1+z};q)_\infty}\frac{1}{1-q^{b+z}}\left[-\alpha_p + \sum_{l\ge 1}\frac{q^{l+b+z}}{1-q^{l+b+z}}\right]$$

at $z=-b$ and also at $z=-b-\chi_k$. However, the expression at the square bracket has zeros there, and thus there are no residues originating from E_2.

The oscillatory function is given by

$$-\frac{1}{L}\frac{\sum_{k\in\mathbb{Z}}\Gamma'(b+\chi_k)e^{-2\pi ik\cdot\log_Q n}}{\sum_{k\in\mathbb{Z}}\Gamma(b+\chi_k)e^{-2\pi ik\cdot\log_Q n}}.$$

As it stands, it is not small, but one can pull out the main term which originates from the terms for $k=0$ in numerator and denominator and thus has represented it as

$$-\frac{1}{L}\psi(b) + \omega(\log_Q n),$$

where $\omega(x)$ is now a tiny oscillation that occurs so frequently in the Analysis of Algorithms, see [5].

Summarizing, we have our main result.

Theorem 1. *The average state reached in approximate counting with black holes is after n random steps given by*

$$\log_Q n - \alpha - \frac{1}{L}\psi(b) + \omega(\log_Q n) + O(\frac{1}{n});$$

the error term originates from the neglected poles at $b-1+\chi_k$. The periodic function $\omega(x)$ (of period 1) has small amplitude, due to the rapid decay of the Gamma function and its derivatives along vertical lines.

Note that in the symmetric case $p=q$, we have $b=1$, and $\psi(1)=-\gamma$.

3. Conclusion

One could do many other things, following the numerous papers on approximate counting as role models. However, we are not going to do that, as the motivation to write this note was to link two areas (set partitions and approximate counting) that *a priori* do not seem to have much in common.

Further research will concentrate on the modified growth property

$$w_k \le d + \max\{w_0, \dots, w_{k-1}\}.$$

For $d \ge 2$, it seems unlikely that explicit enumerations will work. But an asymptotic analysis should be still within reach, although with more advanced methods.

References

[1] G. Andrews, R. Askey, and R. Roy. *Special Functions*, volume 71 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1999.

[2] P. Flajolet. Approximate counting: a detailed analysis. *BIT*, 25:113–134, 1985.

[3] P. Flajolet and R. Sedgewick. Digital search trees revisited. *SIAM Journal on Computing*, 15:748–767, 1986.

[4] P. Flajolet and R. Sedgewick. Mellin transforms and asymptotics: Finite differences and Rice's integrals. *Theoretical Computer Science*, 144:101–124, 1995.

[5] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, Cambridge, 2009.

[6] M. Fuchs, C.-K. Lee, and H. Prodinger. Approximate counting via the Poisson-Laplace-Mellin method. *DMTCS, proceedings AofA12*, yy:xxx–xxx, 2012.

[7] G. Louchard and H. Prodinger. Generalized approximate counting revisited. *Theoretical Computer Science*, 391:109–125, 2008.

[8] T. Mansour and M. Shattuck. Set partitions as geometric words. *Australasian Journal of Combinatorics*, xx:xxx–xxx, 2012.

[9] K. Oliver and H. Prodinger. Words coding set partitions. *Applicable Analysis and Discrete Mathematics*, 5:55–59, 2011.

Helmut Prodinger, Mathematics Department, Stellenbosch University, 7602 Stellenbosch, South Africa.

E-mail address: hproding@sun.ac.za