Some explicit formulae for the distributions of words

Hayato Takahashi (Random Data Lab. Inc.)*

In [1–11, 15] generating functions of the distributions of words are given as rational functions, however except for simple cases, it is difficult to expand rational functions into power series [3] and we cannot obtain explicit formulae for the distributions of words from rational generating functions. In this article, we give some explicit formulae for the distributions of words. Parts of the paper have been presented in [12–14].

Let $X^n := X_1 \cdots X_n$ be random variables that take value in finite alphabet and $N(w_1, \ldots, w_l; X^n)$ the number of the appearances of the words $w_1, \ldots, w_l$ in an arbitrary position of X^n, i.e.

$$N(w_1, \ldots, w_l; X^n) := \left(\sum_{i=1}^{n-|w_1|+1} I_{w_1}(X_i \cdots X_n), \ldots, \sum_{i=1}^{n-|w_l|+1} I_{w_l}(X_i \cdots X_n) \right),$$

where $|w_j|$ is the length of w_j and $I_{w_j}(X_i \cdots X_n) = 1$ if $X_i \cdots X_{i+|w_j|-1} = w_j$ else 0 for all i, j. For example $N(10, 11; 1011101) = (2, 2)$. A word x is called overlapping if there is a word z such that x appears at least 2 times in z and $|z| < 2|x|$ otherwise x is called nonoverlapping. A pair of words x, y is called overlapping if there is a word z such that x and y appear in z and $|z| < |x| + |y|$. A finite set of words S is called nonoverlapping if every pair (x, y) for $x, y \in S$ are not overlapping, otherwise, S is called overlapping. For example, sets of words, $\{11\}$, $\{10, 01\}$, and $\{00, 11\}$ are overlapping, and $\{10\}$ and $\{00111, 00101\}$ are nonoverlapping.

Theorem 1 ([12, 14]). *Let $X_1 X_2 \cdots X_n$ be i.i.d. random variables that take value in finite alphabet $\mathcal{A}$ and P an i.i.d. probability on $\mathcal{A}^n$. Let $w_1, \ldots, w_l$ be the set of nonoverlapping words, $m_i = |w_i|$, and $P(w_i)$ the probability of w_i for $i = 1, \ldots, l$. Then*

$$P(N(w_1, \ldots, w_l; X^n) = (s_1, \ldots, s_l))$$
$$= \sum_{\substack{k_1, \ldots, k_l: \\ s_1 \le k_1, \ldots, s_l \le k_l \\ \sum_i m_i k_i \le n}} (-1)^{\sum_i k_i - s_i} \binom{n - \sum_i m_i k_i + \sum_i k_i}{s_1, \ldots, s_l, k_1 - s_1, \ldots k_l - s_l} \prod_{i=1}^{l} P^{k_i}(w_i).$$

For simplicity, in the following, we consider binary i.i.d. random variables. We enumerate increasing sequence of words. Then we enumerate overlapping word 0^m for $m = 0, 1, \ldots$. Suppose that w_1 is a prefix of w_2 and $(k_1, k_2) = N(w_1, w_2; X^n)$. Then k_1 is the number of appearances of w_1 and w_2. To avoid duplication, we modify the function N.

$$N'(w_1, \ldots, w_l; X^n) := (k_1, k_2, \ldots, k_l) \text{ where}$$
$$k_1 = s_1 - s_2, k_2 = s_2 - s_3, \ldots, k_l = s_l \text{ and } (s_1, \ldots, s_l) = N(w_1, \ldots, w_l; X^n).$$

* e-mail: hayato@h-takahashi.sakura.ne.jp
web: http://h-takahashi.sakura.ne.jp
2000 Mathematics Subject Classification: 05A15, 68R15, 62E15, 68R05.
Keywords: exact distribution, words, combinatorics.

Theorem 2. *Let P be an i.i.d. probability on $\{0,1\}^n$ and $w_1 = 10^m, w_2 = 10^{m+1}, \ldots, w_{n-m} = 10^{n-1}$. Let $(k_1, \ldots, k_{n-m}) = N'(w_1, \ldots, w_{n-m}; X^n))$ and $P_n(t) := P(t = \sum_i ik_i)$. Then*

$$P(N(0^m; X^n) = t) = (P_{n+1}(t) - P(0)P_n(t))P^{-1}(1), \text{ and}$$

$$P_n(t) = \sum_{\substack{r,k_1,\ldots,k_{n-m}: \\ \sum_i (m+i)k_i \le n,\ 0 \le r \le \sum_i k_i \\ t = k1 + 2k_2 + \cdots + (n-m)k_{n-m} - r}} (-1)^r \binom{n - \sum_i (m+i)k_i + \sum_i k_i}{k_1, \ldots, k_{n-m}} \binom{\sum_i k_i}{r} \prod_{i=1}^{n-m} P^{k_i}(w_i).$$

Remark 1. *Let $m = 1$ and P be the fair coin-flipping in Theorem 2. Then $P_n(t) = \binom{n}{t} 2^{-n}$ for all $t \le n$.*

References

[1] F. Bassino, J. Clément, and P. Micodème. Counting occurrences for a finite set of words: combinatorial methods. *ACM Trans. Algor.*, 9(4):Article No. 31, 2010.

[2] V. Berthé and M. Rigo. *Combinatorics, words and symbolic dynamics.* Encyclopedia of Mathematics and Its Applications 159. Cambridge University Press, 2016.

[3] W. Feller. *An Introduction to probability theory and its applications Vol. 1.* Wiley, 3rd edition, 1970.

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

[5] I. Goulden and D. Jackson. *Combinatorial Enumeration.* John Wiley, 1983.

[6] L. Guibas and A. Odlyzko. String overlaps, pattern matching, and nontransitive games. *J. Combin. Theory Ser. A*, 30:183–208, 1981.

[7] P. Jacquet and W. Szpankowski. *Analytic Pattern Matching.* Cambridge University Press, 2015.

[8] M. Lothaire. *Applied Combinatorics on words.* Encyclopedia of Mathematics and Its Applications 105. Cambridge University Press, 2005.

[9] M. Régnier and W. Szpankowski. On pattern frequency occurrences in a markovian sequence. *Algorithmica*, 22(4):631–649, 1998.

[10] S. Robin and J. J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.*, 36(1):179–193, 1999.

[11] S. Robin, F. Rodolphe, and S. Schbath. *DNA, words and models.* Cambridge University Press, english edition, 2005.

[12] H. Takahashi. The explicit formulae for the distributions of nonoverlapping words and its applications to statistical tests for pseudo random numbers. Arxiv 2105.05172.

[13] H. Takahashi. Inclusion-exclusion principles on partially ordered sets and the distributions of the number of pattern occurrences in finite samples, Sep. 2018. Mathematical Society of Japan, Statistical Mathematics Session, Okayama Univ. Japan.

[14] H. Takahashi. The explicit formula for the distributions of nonoverlapping words. *IEICE IT2021-123*, (428):234–236, Mar 2022.

[15] M. S. Waterman. *Introduction to computational biology.* Chapman & Hall, New York, 1995.