A unified approach to explicit formulae for the distributions of runs

Hayato Takahashi

Random Data Lab. Inc.

Feb. 11, 2023
Workshop "Number theory and Ergod theory"

Problem: the number of the occurrences of words in finite strings

The number of the occurrences of words in finite strings plays important role in infomation theory, genome analysis, statistics, AI, etc.

Example: The words 10 and 00 appear in 100010010 three times.

Run: $0^m, m = 2, 3, \ldots.$

We study the enumeration (and distribution) of the number of the occurrences of runs with several types of counting in finite strings.

Remark: The distributions of the number of the occurrence of letters 1 and 0 are given by binomial distribution.

Contents

Known results, generating functions

In Regnier et.al [12], Bassino et.al [1], and Robin [13], the number of the occurrences of words given as generating functions.

$f(n,k,w)$: the number of $x_1 \cdots x_n$ in which w appears k times. Then

$$\sum_{n,k} f(n,k,w) z_1^n z_2^k = \frac{g(z_1,z_2)}{h(z_1,z_2)}.$$

g, h: polynomial.

Known results, generating functions 2, example

Example： Guibas and Odlyzko [5]

$$\sum_n f(n,0,10)z^n = \frac{1}{(1-z)^2} = (\sum z^n)^2 = \sum (n+1)z^n.$$

$$f(n,0,10) = n+1 \text{ for all } n = 1,2,\ldots$$

$$000, 001, 011, 111 \text{ and } f(3,0,10) = 4.$$

Known results, generating functions 3

$f(n, k, w)$: the number of $x_1 \cdots x_n$ in which w appears k times. Then

$$\sum_{n,k} f(n, k, w) z_1^n z_2^k = \frac{g(z_1, z_2)}{h(z_1, z_2)}, \ g, h\text{: polynomial.}$$

注１：既存の方法では長さ n に関する再帰的な関係から生成関数を導出したために上式で n を固定した有限次元の生成関数で表すことはできない．Generating functions are derived by induction on length n and we do not have a finite order generating function.

注２: 生成関数をベキ級数展開すれば $f(n, k, w)$ が求まるが一般には有利関数のべき級数展開は簡単な場合を除いて難しい．It is difficult to expand rational function into power series except for simple cases.

注３: 生成関数から $f(n, k, w)$ の近似解や再帰的計算法を導くことができる．We have approximation of $f(n, k, w)$ from generating function. Some reccurrence formula for $f(n, k, w)$ are derived from generating function.

Main theorem 1: Distributions of nonoverlapping words

Theorem (Takahashi [17, 14])

Let X_1^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

$$P(N(w_1, \ldots, w_l; X_1^n) = (s_1, \ldots, s_l))$$

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

Outline of Proof

Let

$$A(k) := \binom{n - mk + k}{k} P^k(w). \tag{1}$$

Example $k = 2$.

$$x_1 \cdots x_n = \cdots \underbrace{w} \cdots \underbrace{w} \cdots$$
$$x_1 \cdots x_{n-2m+2} = \cdots \ \alpha \ \cdots \ \alpha \cdots$$

$B(t)$: the probability that nonoverlapping words w appear k times.
Then

$$A(k) = \sum_{k \le t} B(t) \binom{t}{k}.$$

Outline of Proof

Let $F_A(z) := \sum_k A(k)z^k$ and $F_B(z) := \sum_k B(k)z^k$. Then

$$\begin{aligned} F_A(z) &= \sum_k z^k \sum_{k\le t} B(t)\binom{t}{k} \\ &= \sum_t B(t) \sum_{k\le t} \binom{t}{k} z^k \\ &= \sum_t B(t)(z+1)^t = F_B(z+1), \end{aligned}$$

and

$$F_B(z) = F_A(z-1).$$

moments

Let

$$A_{t,s} := \sum_r \binom{s}{r} r^t (-1)^{s-r}.$$

$A_{t,s}$ is the number of surjective functions from $\{1, 2, \dots, t\} \to \{1, 2, \dots, s\}$ for $t, s \in \mathbb{N}$, see pp.100 Problem 1 Riordan 1958.

Theorem (Takahashi [17, 14])

Let w be a nonoverlapping word.

$$\forall t \; E(N^t(w; X^n)) = \sum_{s=1}^{\min\{T,t\}} A_{t,s} \binom{n - s|w| + s}{s} P^s(w),$$

where $T = \max\{t \in \mathbb{N} \mid n - t|w| \geq 0\}$.

Known resuls, runs, various type of counting

Fu et.al (1994) [3] showed the distributions of the following statistics by Markov imbedding method.

(i) $E_{n,m}$, the number of 0^m of size exactly m. Mood(1940) [9].

(ii) $G_{n,m}$, the number of 0^m of size greater than or equal to m.

(iii) $N_{n,m}$, the number of nonoverlapping consecutive 0^m. Feller [2].

(iv) $M_{n,m}$, the number of overlapping consecutive 0^m.

(v) L_n, the size of the longest run of 0s.

Example: $E_{8,2} = 1$, $G_{8,2} = 2$, $N_{8,2} = 3$, $M_{8,2} = 4$, and $L_8 = 4$ for

run 00 and 10000100.

Known resuls, other explicit formulae

Explicit formulae for the distributions of runs are given in

$G_{n,m}$: Makri et.al (2007) [8]

$N_{n,m}$: Hirano (1986) [6], Phillipou et.al (1986) [11], Godbole (1990) [4], Muselli (1996) [10]

$M_{n,m}$: Ling (1988) [7].

L_n: Makri et.al (2007) [8]

Definition

(i) $\bar{E}_{n,m}$, the number of 0^m of size exactly m that start with 1.

(ii) $\bar{G}_{n,m}$, the number of 0^m of size greater than or equal to m that start with 1.

(iii) $\bar{N}_{n,m}$, the number of nonoverlapping consecutive 0^m that start with 1.

(iv) $\bar{M}_{n,m}$, the number of overlapping consecutive 0^m that start with 1.

$E_{8,2} = \bar{E}_{8,2} = 1$, $G_{8,2} = \bar{G}_{8,2} = 2$, $N_{8,2} = \bar{N}_{8,2} = 3$, $M_{8,2} = \bar{M}_{8,2} = 4$ for

run 00 and 10000100.

$E_{10,2} = 2$, $G_{10,2} = 3$, $N_{10,2} = 4$, $M_{10,2} = 5$,
$\bar{E}_{10,2} = 1$, $\bar{G}_{10,2} = 2$, $\bar{N}_{10,2} = 3$, $\bar{M}_{10,2} = 4$ for

run 00 and 0010000100.

Main Theorem2: Distributions of runs

Theorem

Let P be an i.i.d. probability on $\{0,1\}^n$.

(i) $P(E_{n,m}=t) = (P(\bar{E}_{n+1,m}=t) - P(0)P(\bar{E}_{n,m}=t))/P(1)$.

$$P(\bar{E}_{n,m}=t) = \sum_{\substack{k_1,k_2: \\ (m+1)k_1+(m+2)k_2\le n, \\ t\le k_1+k_2}} (-1)^{k_1-t}\binom{n-(m+1)k_1-(m+2)k_2+k_1+k_2}{k_1,k_2} \times \binom{k_1+k_2}{t} P^{k_1}(10^m)P^{k_2}(10^{m+1}).$$

Main Theorem 2

Theorem (Continue)

(ii) $P(G_{n,m} = t) = (P(\bar{G}_{n+1,m} = t) - P(0)P(\bar{G}_{n,m} = t))/P(1).$

$$P(\bar{G}_{n,m} = t) = \sum_{k:\ t \le k, (m+1)k \le n} (-1)^{k-t} \binom{n-(m+1)k+k}{t, k-t} P^k(10^m).$$

Main Theorem 3

Theorem (Continue)

(iii) $P(N_{n,m} = t) = (P(\bar{N}_{n+1,m} = t) - P(0)P(\bar{N}_{n,m} = t))P^{-1}(1)$.

$P(\bar{N}_{n,m} = t) =$

$$\sum_{\substack{r,k_1,\ldots,k_T: \\ \sum_i (mi+1)k_i \le n,\ 0 \le r \le \sum_i k_i \\ t = \sum_i ik_i - r}} (-1)^r \binom{n - \sum_i (mi+1)k_i + \sum_i k_i}{k_1, \ldots, k_{n-m}} \binom{\sum_i k_i}{r}$$

$$\times \prod_{i=1}^{T} P^{k_i}(10^{im}).$$

T is the maximum integer such that $Tm + 1 \le n$.

Main Theorem 4

Theorem (Continue)

(iv) $P(M_{n,m} = t) = (P(\bar{M}_{n+1,m} = t) - P(0)P(\bar{M}_{n,m} = t))P^{-1}(1)$.

$$P(\bar{M}_{n,m} = t) = \sum_{\substack{r,k_1,\ldots,k_{n-m}: \\ \sum_i (m+i)k_i \leq n,\ 0 \leq r \leq \sum_i k_i \\ t = \sum ik_i - r}} (-1)^r \binom{n - \sum_i (m+i)k_i + \sum_i k_i}{k_1, \ldots, k_{n-m}}$$

$$\times \binom{\sum_i k_i}{r} \prod_{i=1}^{n-m} P^{k_i}(10^{m+i-1}).$$

(v) $P(L_n = t) = P(N_{n,t+1} = 0) - P(N_{n,t} = 0)$.

Lemma (Takahashi [18])

Let

$$E_{n,m,t} = \{x_1^n \mid E_{n,m}(x_1^n) = t\} \text{ and } \bar{E}_{n,m,t} = \{x_1^n \mid \bar{E}_{n,m}(x_1^n) = t\}.$$

Then

$$P(\bar{E}_{n+1,m,t}) = P(0)P(\bar{E}_{n,m,t}) + P(1)P(E_{n,m,t}). \tag{2}$$

$(G_{n,m,t}, \bar{G}_{n,m,t})$, $(N_{n,m,t}, \bar{N}_{n,m,t})$, *and* $(M_{n,m,t}, \bar{M}_{n,m,t})$ *are defined by similar manner and (2) is true for them respectively.*

Proof) Let $\bar{E}^0_{n+1,m,t} = \{0x_1^n \mid \bar{E}_{n+1,m}(0x_1^n) = t\}$ and $\bar{E}^1_{n+1,m,t} := \{1x_1^n \mid \bar{E}_{n+1,m}(1x_1^n) = t\}$. Then

$$\bar{E}^0_{n+1,m,t} = \{0x_1^n \mid x_1^n \in \bar{E}_{n,m,t}\},\ \bar{E}^1_{n+1,m,t} = \{1x_1^n \mid x_1^n \in E_{n,m,t}\}, \text{ and} \tag{3}$$

$$\bar{E}_{n+1,m,t} = \bar{E}^0_{n+1,m,t} \cup \bar{E}^1_{n+1,m,t}. \tag{4}$$

By (3) and (4), we have (2). The proof of the latter part is similar. □

Definitions

$\mathbf{N}(w_1, \ldots, w_l; X_1^n)$: the number of the overlapping appearances of $w_1, w_2, \ldots, w_l$ in X_1^n.

Suppose that w_1 and w_2 are nonoverlapping,
$w_1 \sqsubset w_2$ and $\mathbf{N}(w_1, \ldots, w_l; X_1^n) = (s_1, \ldots, s_l)$.

Then
s_1 is the number of the appearances of w_1 and w_2.

$$\mathbf{N}'(w_1, \ldots, w_l; X_1^n) := (s_1 - s_2, s_2 - s_3, \ldots, s_l)$$
$$\text{if } \mathbf{N}(w_1, \ldots, w_l; X_1^n) = (s_1, s_2, \ldots, s_l).$$

Example: $\mathbf{N}(100, 1000; 1010001) = (1, 1)$ and
$\mathbf{N}'(100, 1000; 1010001) = (0, 1)$.

Lemma (Takahashi [15, 16, 18])

Let $w_1 \sqsubset w_2 \cdots \sqsubset w_l$ *be an increasing sequence of nonoverlapping words,*

$$A(k_1,\dots,k_l) := \binom{n - \sum_i m_i k_i + \sum_i k_i}{k_1,\dots,k_l} \prod_{i=1}^{l} P^{k_i}(w_i),$$

$$B(k_1,\dots,k_l) := P(\mathbf{N}'(w_1,\dots,w_l; X^n) = (k_1,k_2,\dots,k_l)),$$

$$F_A(z_1,\dots,k_l) := \sum_{\substack{k_1,\dots,k_l: \\ \sum_i m_i k_i \le n}} A(k_1,\dots,k_l) z^{k_1} \cdots z^{k_l}, \ \textit{and}$$

$$F_B(z_1,\dots,z_l) := \sum_{\substack{k_1,\dots,k_l: \\ \sum_i m_i k_i \le n}} B(k_1,\dots,k_l) z^{k_1} \cdots z^{k_l}.$$

Then

$$F_A(z_1, \ldots, z_l) = F_B(z_1 + 1, z_1 + z_2 + 1, \ldots, \sum_i z_i + 1) \text{ and} \qquad (5)$$

Set $z_1 = X, z_2 = X(X+1), \ldots, z_l = X(X+1)^{l-1}$ in (6). Then

$$F_A(X, X(X+1), \ldots, X(X+1)^{l-1}) = F_B(X+1, (X+1)^2, \ldots, (X+1)^l)$$

$$\begin{aligned} F_A(Y-1, (Y-1)Y, \ldots, (Y-1)Y^{l-1}) &= F_B(Y, Y^2, \ldots, Y^l) \\ &= \sum_{\substack{k_1, \ldots, k_l: \\ \sum_i m_i k_i \leq n}} B(k_1, \ldots, k_l) Y^{\sum i k_i}. \end{aligned}$$

Generalization

Our theorem is true for arbitrary alphabet.
$X_1, X_2, \ldots, X_n$: i.i.d. r.v.$\sim (R, \mathcal{B}, Q)$.
Event $A_0 \subset \mathbb{R}$ and $Q(A_0) = Q(X_i \in A_0)$.
Example: The run $A_0 A_0$ occurs one time in the event $A_0^c A_0^c A_0 A_0 A_0^c$.

Corollary

The probability of statistics (i)–(v) of run A_0 is obtained by setting $P(0) = Q(A_0)$ and $P(1) = Q(A_0^c)$ in Main theorem.

Example: $X_i \in \{0, 1, 2, \ldots\}$. The probability of runs of 0 are obtained by setting $P(1) = 1 - Q(0)$ and $P(0) = Q(0)$ in Main theorem.

Reference I

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

[3] J. C. Fu and M. V. Koutras.
Distribution theory of runs: a Markov chain approach.
J. Amer. Statist. Assoc., 89(427):1050–1058, 1994.

[4] A. P. Godbole.
Specific formulae for some success run distributions.
Statist. Probab. Lett., 10:119–124, 1990.

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

Reference II

[6] K. Hirano.
Some properties of the distributions of order k.
pages 43–53, 1986.
Fibonacci Numbers and their Applications, A. N. Phillipou, A. F. Horadam and G. E. Bergum eds, Reidel.

[7] K. D. Ling.
On binomial distributions of order k.
Statist. Probab. Letters, 6:247–250, 1988.

[8] F. S. Makri, A. N. Philippou, and Z. M. Psillakis.
Shortest and longest length of success runs in binary sequences.
J. Statist. Plan. Inference, 137:2226–2239, 2007.

[9] A. M. Mood.
The distribution theory of runs.
Ann. Math. Statist, 11(4):367–392, 1940.

[10] M. Muselli.
Simple expressions for success run distributions in Bernoulli trials.
Statist. Probab. Lett., 31:121–128, 1996.

Reference III

[11] A. N. Phillipou and F. S. Makri.
Success, runs and longest runs.
Statist. Probab. Lett., 4:211–215, 1986.

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

[13] 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.

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

[15] 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.

Reference IV

[16] H. Takahashi.
The distributions of sliding block patterns in finite samples and the inclusion-exclusion principles for partially ordered sets.
RIMS Kôkyûroku, Kyoto University, 2116:1–9, 2019.
arxiv:1811.12037v1.

[17] H. Takahashi.
The explicit formula for the distributions of nonoverlapping words.
IEICE Technical Report IT2021-123, 121(428):234–236, Mar 2022.

[18] H. Takahashi.
Explicit formula for the distributions of runs.
IEICE Technical Report IT2022-65, 122(355):208–210, Jan 2023.