The distributions of sliding block patterns in finite samples and the inclusion-exclusion principles for partially ordered sets

Hayato Takahashi[1]

Let $X \in A^n$ with finite alphabet A and $w \in A^*$. Let $|w|$ be the length of the word w. We consider the following random variable,

$$N_w := \sum_{i=1}^{n} I_{X_i^{i+|w|-1}=w} \text{ where } I_{X_i^{i+|w|-1}=w} = 1 \text{ if } X_i^{i+|w|-1} = w \text{ else } 0.$$

We also call this statistics sliding block patterns. In particular if we count the occurrence of multiple words, it is called suffix tree.

The distributions of sliding block patterns have been shown via generating functions based on induction of sample size, see [1, 2, 3, 5, 4].

In this paper we show the distributions of sliding block patterns for Bernoulli processes with finite alphabet, which is not based on the induction on sample size. We show a new inclusion-exclusion formula in multivariate generating function form on partially ordered sets, and show a simpler expression of generating functions of the number of pattern occurrences in finite samples.

We say that a word w is overlapping if there is a word x with $|w| < |x| < 2|w|$ and w appears in x at least 2 times, and w is called non-overlapping if there is no such x. We write $x \sqsubset y$ if x is a prefix of y.

Theorem 1 *Let P be an i.i.d. process of fixed sample size n of finite alphabet. Let $s_1 \sqsubset s_2 \sqsubset \cdots \sqsubset s_l$ be an increasing non–overlapping words of finite alphabet, i.e., s_i is a prefix of s_j and $m_i < m_j$, where m_i is the length of s_i, for all $i < j$. Let $P(s_i)$ be the probability of s_i for $i = 1, \ldots, l$. Let*

$$\begin{aligned}
A(k_1, \ldots, k_l) &= \binom{n - \sum_i m_i k_i + \sum_i k_i}{k_1, \ldots, k_l} \prod_{i=1}^{l} P^{k_i}(s_i), \\
B(k_1, \ldots, k_l) &= P(\sum_{i=1}^{n} I_{X_i^{i+m_i-1}=s_j} = k_j,\ j = 1, \ldots, l), \\
F_A(z_1, \ldots, z_l) &= \sum_{k_1,\ldots,k_l} A(k_1, \ldots, k_l) z^{k_1} \cdots z^{k_l},\ \textit{and} \\
F_B(z_1, \ldots, z_l) &= \sum_{k_1,\ldots,k_l} B(k_1, \ldots, k_l) z^{k_1} \cdots z^{k_l}.
\end{aligned} \tag{1}$$

Then

$$F_A(z_1, z_2, \ldots, z_l) = F_B(z_1 + 1, z_1 + z_2 + 1, \ldots, z_1 + \cdots + z_l + 1).$$

With slight modification of Theorem 1, we can compute the number of the occurrence of the overlapping increasing words. For example, let us consider increasing self-overlapping words 11, 111, 1111 and the number of their occurrences. Let 011, 0111, 01111 then these words are increasing non-self-overlapping words. The number of occurrences 11, 111, 1111 in sample of length n is equivalent to the number of occurrences 011, 0111, 01111 in sample of length $n + 1$ that starts with 0. We can apply Theorem 1 to derive the distribution of increasing overlapping words with this manner.

In [5], expectation, variance, and CLTs for the sliding block pattern are shown. We show the general higher moments for non-overlapping words.

[1]Random Data Lab. Email: hayato.takahashi@ieee.org

Theorem 2 *Let w be a non-overlapping pattern.*

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

$$A_{t,s} = \sum_r \binom{s}{r} r^t (-1)^{s-r}, \; T = \max\{t \in \mathbb{N} \mid n - t|w| \geq 0\}.$$

In the above theorem, $A_{t,s}$ is the number of surjective functions from $\{1, 2, \ldots, t\} \to \{1, 2, \ldots, s\}$ for $t, s \in \mathbb{N}$, see [6].

In [5], it is shown that central limit theorem holds for sliding block patterns,

$$P(\frac{N_w - E(N_w)}{\sqrt{V_w}} < x) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}x^2} dx,$$

where w is non-overlapping pattern, $E(N_w) = (n - |w| + 1)P(w)$ and $V(N_w) = E(N_w) + (n - 2|w| + 2)(n - 2|w| + 1)p^2(w) - E^2(N_w)$.

Let

$$N'_w := \sum_{i=1}^{\lfloor n/|w| \rfloor} I_{X_{i*|w|}^{(i+1)*|w|-1} = w}.$$

N'_w obeys binomial law if the process is i.i.d. We call N'_w *block-wise sampling*. As an application of CLT approximation, we compare power functions of sliding block sampling N_w and block-wise sampling N'_w. We consider the following test for sliding block patterns: We write $E_\theta = E(N_w)$ and $V_\theta = V(N_w)$ if $P(w) = \theta$. Null hypothesis: $P(w) = \theta^*$ vs alternative hypothesis $P(w) < \theta^*$. Reject null hypothesis if and only if $N_w < E_{\theta^*} - 5\sqrt{V_{\theta^*}}$. The likelihood of the critical region is called power function, i.e., $Pow(\theta) := P_\theta(N_w < E_{\theta^*} - 5\sqrt{V_{\theta^*}})$ for $\theta \leq \theta^*$.

We construct a test for block-wise sampling: Null hypothesis: $P(w) = \theta^*$ vs alternative hypothesis $P(w) < \theta^*$. Reject null hypothesis if and only if $N'_w < E'_{\theta^*} - 5\sqrt{V'_{\theta^*}}$, where $E'_\theta = \lfloor n/|w| \rfloor \theta$ and $V'_\theta = \lfloor n/|w| \rfloor \theta(1 - \theta)$. The following table shows powers of tests for sliding block patterns and block wise sampling at $\theta = 0.2, 0.18, 0.16$ under the condition that $\theta^* = 0.25, |w| = 2,$ and $n = 500$.

θ	0.2	0.18	0.16
Power of Sliding block	0.316007	0.860057	0.995681
Power of Block wise	0.000295	0.002939	0.021481

Acknowledgement

The author thanks for a helpful discussion with Prof. S. Akiyama and Prof. M. Hirayama at Tsukuba University. This work was supported by the Research Institute for Mathematical Sciences, an International Joint Usage/Research Center located in Kyoto University.

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

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

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

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

[6] J. Riordan. *Introduction to combinatorial analysis*. John Wiley, 1958.