# The explicit formulae for the distributions of words

Hayato Takahashi

Random Data Lab. Inc.

Aug. 22, 2023
ICIAM 2023 Waseda Univ.

# Problem: counting words in finite strings

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

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

Explicit formulae for distributions of words are only known for several types of counting of runs (e.g. 00).

Those for nonoverlapping words (e.g. 10) are even unknown!

# Known results, generating functions

The probability of the number of occurrences of words given as rational generating functions (e.g. Bassino et.al [4], Regnier et.al [17], and Robin [18]).

$f(n, k, w)$: probability that $w$ appears $k$ times in sample size $n$. 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.

Remark: We have approximations and recurrence formulae from rational generating functions, however it is difficult (or very tedious task) to obtain explicit formulae from them in general.

# Known results for runs

For $x \in \{0,1\}^n$, let
(i) $E_{n,m}(x)$, the number of $0^m$ of size exactly $m$ in $x$ [5], (explicit formulae are not known except for Markov imbedding method.)
(ii) $G_{n,m}(x)$, the number of $0^m$ of size greater than or equal to $m$ in $x$ [5, 3],
(iii) $N_{n,m}(x)$, the number of non-overlapping $0^m$ in $x$ [6, 9, 14, 3, 5],
(iv) $M_{n,m}(x)$, the number of overlapping $0^m$ in $x$ [11, 3, 10, 5, 7]
(v) $L_n(x)$, the size of the longest run of 0s in $x$ [12, 16, 3, 5],
(vi) $T_k(x)$, the stopping time $t$ such that $0^k$ first appear in $x = x_1 \cdots x_t$ [2, 15, 19], and
(vii) $N_{n,m,\mu}(x)$, the enumeration of $0^m$ such that we allow $\mu$-letters overlapping with the previous $0^m$ in the string $x$ [1, 8, 13]. (explicit formulae are not known.)

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.

# Contents

1. Explicit formulae for the joint distributions of nonoverlapping words.
2. Universal explicit formulae for a family of distributions of runs.
3. Total variational distance.
4. Algorithm and complexity.

# Main theorem 1: Distributions of nonoverlapping words

## Theorem (Takahashi [21, 20])

Let $X_1^n := X_1 X_2 \cdots X_n$ be finite valued i.i.d. random variables. Let $w_1, \ldots, w_l$ be a set of nonoverlapping words. Let $P(w_i)$ be the probability of $w_i$ (e.g. $P(10) = P(1)P(0)$) for $i = 1, \ldots, l$. Let

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

$$B(k_1, \ldots, k_l) = P(\mathbf{N}(w_1, \ldots, w_l; X_1^n) = (k_1, \ldots, k_l)),$$

$$F_A(z_1, \ldots, z_l) = \sum_{k_1, \ldots, k_l} A(k_1, \ldots, k_l) z^{k_1} \cdots z^{k_l}, \text{ and}$$

$$F_B(z_1, \ldots, z_l) = \sum_{k_1, \ldots, k_l} B(k_1, \ldots, k_l) z^{k_1} \cdots z^{k_l}.$$

# Main theorem 1: Distributions of nonoverlapping words

Then

$$A(k_1,\dots,k_l)=\sum B(t_1,\dots,t_l)\binom{t_1}{k_1}\cdots\binom{t_l}{k_l},$$

$$F_A(z_1,z_2,\dots,z_l)=F_B(z_1+1,z_2+1,\dots,z_l+1),\ \text{and}$$

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

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

Here $N(w_1,\dots,w_l;X_1^n)=(s_1,\dots,s_l)\Leftrightarrow w_i$ appear $s_i$ times in $X_1^n$ for all $i$.

# Notations

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

Example:
$\mathbf{N}(100, 1000; 10010001) = (2, 1)$ and $\mathbf{N}'(100, 1000; 10010001) = (1, 1)$.

# Main theorem 2: Distributions of runs

## Theorem

Let $X_1, X_2, \ldots,$ be i.i.d. binary random variables from $P$. Let $w_1 \sqsubset w_2 \cdots \sqsubset w_l$ be an increasing sequence of non-overlapping words. Let

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

$$B(k_1, \ldots, k_l) := P(\mathbf{N}'(w_1, \ldots, w_l; X_1^n) = (k_1, k_2, \ldots, k_l)),$$

$$F_A(z_1, \ldots, k_l) := \sum_{\substack{k_1, \ldots, k_l: \\ \sum_i |w_i| k_i \leq n}} A(k_1, \ldots, k_l) z^{k_1} \cdots z^{k_l}, \text{ and}$$

$$F_B(z_1, \ldots, z_l) := \sum_{\substack{k_1, \ldots, k_l: \\ \sum_i |w_i| k_i \leq n}} B(k_1, \ldots, k_l) z^{k_1} \cdots z^{k_l}.$$

# Main theorem 2: Distributions of runs

## Theorem (Continue)

*Then*

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

$$P(C_{n,(w_1,\ldots,w_l)}(X_1^n) = t) = \sum_{\substack{r,k_1,\ldots,k_l: \\ \sum |w_i| k_i \leq n,\ r \leq \sum k_i \\ t = \sum i k_i - r}} (-1)^r \binom{n - \sum |w_i| k_i + \sum k_i}{k_1, \ldots, k_l} \times \binom{\sum k_i}{r} \prod P^{k_i}(w_i), \text{ where}$$

$$C_{n,(w_1,\ldots,w_l)}(x) := t \text{ if } \sum i k_i = t \text{ and } \mathbf{N}'(w_1, \ldots, w_l; x_1^n) = (k_1, k_2, \ldots, k_l).$$

# Notation

Let $m_1 < \ldots < m_l$. For $x \in \{0,1\}^n$, let

$$D_{n,(m_1,\ldots,m_l)}(x) := t \Leftrightarrow$$
$$\sum ik_i = t \text{ and } \mathbf{N}'(10^{m_1}\ldots, 10^{m_l}; 1x) = (k_1, k_2, \ldots, k_l),$$

where $1x$ is the concatenation of 1 and $x$.

$C_n$ dose not count run that start from the head of sample $x$ but $D_n$ does.

Example: $C_{n,100}(00100) = 1$ and $D_{n,2}(00100) = 2$.

## Corollary

Let $X_1, X_2, \ldots, X_n$ be i.i.d. binary random variables. Let $m_1 < \ldots < m_l$ and $w_i = 10^{m_i}$ for $1 \leq i \leq l$. For all $t \geq 0$,

$$P(D_{n,(m_1,\ldots,m_l)}(X_1^n) = t)$$
$$= (P(C_{n+1,(w_1,\ldots,w_l)}(X_1^{n+1}) = t) - P(0)P(C_{n,(w_1,\ldots,w_l)}(X_1^n) = t))/P(1).$$

Proof) Observe that

$$\{0x \mid C_{n+1,(w_1,\ldots,w_l)}(0x) = t, |x| = n\} = \{0x \mid C_{n,(w_1,\ldots,w_l)}(x) = t, |x| = n\},$$
$$\{1x \mid C_{n+1,(w_1,\ldots,w_l)}(1x) = t, |x| = n\} = \{1x \mid D_{n,(w_1,\ldots,w_l)}(x) = t, |x| = n\}.$$

We have

$$P(C_{n+1,(w_1,\ldots,w_l)}(X_1^{n+1}) = t)$$
$$= P(C_{n+1,(w_1,\ldots,w_l)}(0X_1^n) = t) + P(C_{n+1,(w_1,\ldots,w_l)}(1X_1^n) = t)$$
$$= P(0)P(C_{n,(w_1,\ldots,w_l)}(X_1^n) = t) + P(1)P(D_{n,(w_1,\ldots,w_l)}(X_1^n) = t).$$

# Main Theorem2: Distributions of runs

## Theorem

*Let $X_1, X_2, \ldots,$ be i.i.d. binary random variables from $P(X_i = 1) = q$ and $P(X_i = 0) = p$ for all $i$, where $p + q = 1$.*

*(i)* $$P(\bar{E}_{n,m}(X_1^n) = t) = \sum_{\substack{k_1,k_2: \\ (m+1)k_1+(m+2)k_2 \le n, \\ t \le k_1+k_2}} (-1)^{k_1-t} \binom{n - mk_1 - (m+1)k_2}{k_1, k_2} \times \binom{k_1 + k_2}{t} q^{k_1+k_2} p^{k_1 m + k_2(m+1)},$$ *and*

$$P(E_{n,m}(X_1^n) = t) = (P(\bar{E}_{n+1,m}(X_1^{n+1}) = t) - pP(\bar{E}_{n,m}(X_1^n) = t))/q,$$

*where $\bar{E}_{n,m}(x)$, the number of $10^m$ of size exactly $m+1$ in $x$.*

# Main Theorem 2

## Theorem (Continue)

$$\textit{(ii)}\ P(\bar{G}_{n,m}(X_1^n)=t)=\sum_{t\le k\le\lfloor\frac{n}{m+1}\rfloor}(-1)^{k-t}\binom{n-mk}{t,k-t}q^k p^{km},\ \textit{and}$$

$$P(G_{n,m}(X_1^n)=t)=(P(\bar{G}_{n+1,m}(X_1^{n+1})=t)-pP(\bar{G}_{n,m}(X_1^n)=t))/q,$$

*where* $\bar{G}_{n,m}(x)$, *the number of* $10^h$ *for* $h\ge m$ *in* $x$.

$$\textit{(iii)}\ P(T_m>n)=P(L_n<m)=\sum_{0\le k\le\lfloor\frac{n+1}{m+1}\rfloor}(-1)^k\binom{n+1-mk}{k}q^{k-1}p^{km}$$
$$-\sum_{0\le k\le\lfloor\frac{n}{m+1}\rfloor}(-1)^k\binom{n-mk}{k}q^{k-1}p^{km+1}.$$

# Main Theorem 3

## Theorem (Continue)

*(iv)* $P(N_{n,m,\mu}(X_1^n) = t)$

$= (P(C_{n+1,(w_1,\ldots,w_l)}(X_1^{n+1} = t) - pP(C_{n,(w_1,\ldots,w_l)}(X_1^n) = t))/q$

*for all* $0 \leq \mu \leq m-1$, *where* $m_i = mi - \mu(i-1)$ *and* $w_i = 10^{m_i}$ $\forall i$ *and* $T$ *is the largest integer such that* $mT - \mu(T-1) \leq n$.

Remark: if $\mu = m-1$, $N_{n,m,\mu} = M_{n,m}$ and if $\mu = 0$, $N_{n,m,\mu} = N_{n,m}$.

# Total variational distance

## Proposition

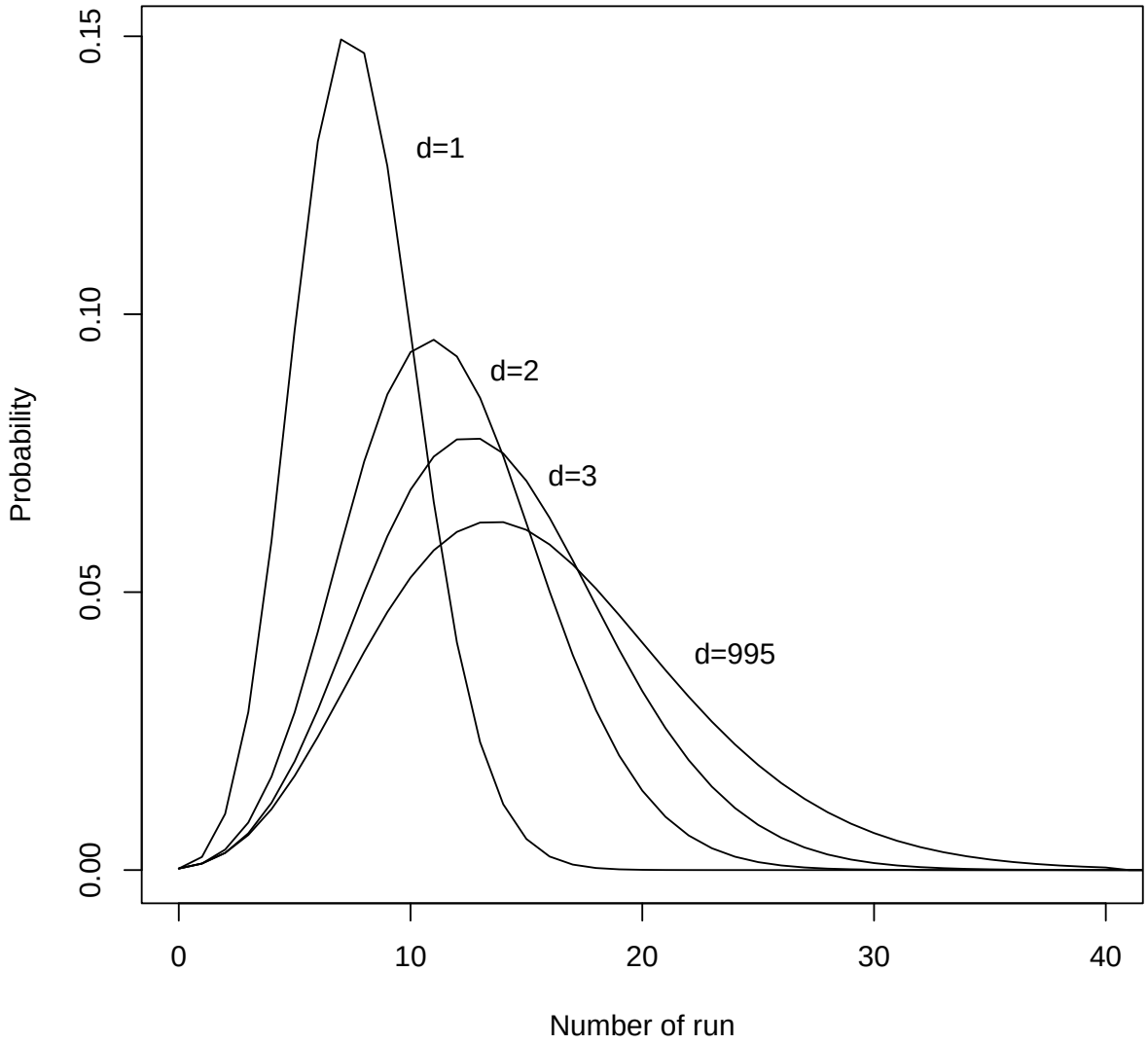
Let $X_1, \ldots, X_n$ be i.i.d. binary random variables and $P(0) = p$. Assume that $d < l$. Then for all $t$,

$$\begin{aligned} &|P(D_{n,(m_1,\ldots,m_d)}(X_1^n) = t) - P(D_{n,(m_1,\ldots,m_l)}(X_1^n) = t)| \\ &\leq 2(n + 1 - m_{d+1})p^{d+1}. \end{aligned}$$

Dist : total variational distance.

$$\begin{aligned} &\mathrm{Dist}(D_{n,d}, D_{n,l}) := \\ &\sup_{A \subseteq [0,n]} |P(D_{n,(m_1,\ldots,m_d)}(X_1^n) \in A) - P(D_{n,(m_1,\ldots,m_l)}(X_1^n) \in A)| \\ &\leq n^2 p^{d+1}. \end{aligned}$$

Graph of $P(D_{n,(m_1,\ldots,m_d)})$ for $d = 1, 2, 3, 995$, $m_1 = 5, m_2 = 6, \ldots$, sample size 1000, P(0)=0.5.



# Algorithm and Complexity for Computing $P(C_n = t)$

Let $\mathbb{Z}_0 = \{0, 1, 2, \ldots\}$ and

$$G_t(n, l, |w_1|, \ldots, |w_l|) :=$$

$$\{(r, k_1, \ldots, k_l) \in \mathbb{Z}_0^{l+1} \mid \sum_{1 \le i \le l} k_i |w_i| \le n,\ r \le \sum_{i=1}^{l} k_i,\ \sum_{i=1}^{l} i k_i - r = t\}.$$

$$P(C_{n,(w_1,\ldots,w_l)}(X_1^n) = t) = \sum_{(r,k_1,\ldots,k_l) \in G_t} (-1)^r \binom{n - \sum |w_i| k_i + \sum k_i}{k_1, \ldots, k_l} \times \binom{\sum k_i}{r} \prod P^{k_i}(w_i).$$

$$G(n, l, |w_1|, \ldots, |w_l|) :=$$

$$\{(r, k_1, \ldots, k_l) \in \mathbb{Z}_0^{l+1} \mid \sum_{1 \le i \le l} k_i |w_i| \le n,\ r \le \sum_{i=1}^{l} k_i,\ 0 \le \sum_{i=1}^{l} i k_i - r \le t\}.$$

# Algorithm and Complexity for computing $P(C_n = h),\ h = 0, \ldots, t$ (Application of Bucket sort algorithm)

**Algorithm**

1. Initialize $P(C_n = h) = 0$ for all $h = 0, \ldots, t$.
2. Enumerate all nonnegative $(r, k_1, \ldots, k_l) \in G(n, l, |w_1|, \ldots, |w_l|)$.
   For each vector $(r, k_1, \ldots, k_l) \in G(n, l, |w_1|, \ldots, |w_l|)$, set

$$P(C_{n,(w_1,\ldots,w_l)}(X_1^n) = h) := P(C_{n,(w_1,\ldots,w_l)}(X_1^n) = h) + f(r, k_1, \ldots, k_l), \text{ where}$$

$$h = \sum_{i=1}^{l} i k_i - r \text{ and } f(r, k_1, \ldots, k_l) = (-1)^r \binom{n - \sum |w_i| k_i + \sum k_i}{k_1, \ldots, k_l} \times \binom{\sum k_i}{r} \prod P^{k_i}(w_i).$$

3. Output $P(C_n = h)$ for all $h = 0, \ldots, t$.

## Lemma

*Assume that* $1 \leq |w_1| \leq \cdots \leq |w_l|$. *Then*

$$|\{(k_1, \ldots, k_l) \in \mathbb{Z}_0^l \mid \sum_{1 \leq i \leq l} k_i |w_i| \leq n\}| \leq \frac{(n + \sum_i |w_i|)^h}{h! \prod_i |w_i|}.$$

## Theorem

*Let $n$ be the sample size. For given $n, l, t$, and $w_1, \ldots, w_l$,*
*1.*

$$|G(n, l, |w_1|, \ldots, |w_l|)| \leq (t+1)(t + \frac{l(l-1)}{2})^{l-1}((l-1)!)^{-2}(\frac{n}{|w_1|} + 1),$$

*2. Fix $l$ and $t$. Then $|G(n, l, |w_1|, \ldots, |w_l|)| = O(n)$,*

*3. Fix $l$. Assume that $t = O(n^{\frac{\alpha}{l}})$ and $\alpha > 0$. Then*

$$|G(n, l, |w_1|, \ldots, |w_l|)| = O(n^{\alpha+1}), \text{ and}$$

*4. Assume $r > 0$. Let $l$ be the least integer such that $l \geq -\frac{2}{\log p}\log n - \frac{r}{\log p} - 1$ and $t \sim \frac{l^2}{2}$. Then $\mathrm{Dist}(C_{n,l}, C_{n,m}) \leq 2^{-r}$ for $l < m$ and*

$$|G(n, l, |w_1|, \ldots, |w_l|)| = O((\log n)n^{-\frac{4}{\log p}+1}).$$

# Concluding Remark

Let sample size $n$.

Computational complexity of Markov imbedding method is $O(n^{2+\alpha} \log n)$ for some $0 \leq \alpha < 1$.

The only practical computational complexity for large sample size is $O(n)$ or $O(n \log n)$.

Computational complexity of our algorithm is $O(n)$ for fixed dimension of parameter and limited value of random variable, and our algorithm still work for large sample size.

# Reference I


[1] S. Aki and K. Hirano.
Numbers of success-runs of specified length until certain stopping time rules and generalized binomial distributions of order k.
*Ann. Inst. Statist. Math.*, 52(4):767–777, 2000.

[2] S. Aki, H. Kuboki, and K. Hirano.
On discrete distributions of order k.
*Ann. Inst. Statist. Math.*, 36:431–440, 1984.

[3] D. L. Antzoulakos and S. Chadjiconstantindis.
Distributions of numbers of success runs of fixed length in Markov dependent trials.
*Ann. Inst. Statist. Math.*, 53(3):599–619, 2001.

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

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

# Reference II


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

[7] A. P. Godbole.
The exact and asymptotic distribution of overlapping success runs.
*Comm. Statist. Theory Methods*, 21:953–967, 1992.

[8] S. Han and S. Aki.
A unified approach to binomial-type distributions of order k.
*Commun. Statist. Theor. Meth.*, 29:1929–1943, 2000.

[9] 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.

[10] M. V. Koutras and V. A. Alexandrou.
Non-parametric randomness tests based on success runs of fixed length.
*Statist. Probab. Lett.*, 32:393–404, 1997.

# Reference III


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

[12] 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.

[13] F. S. Makri and Z. M. Psillakis.
On l-overlapping runs of ones of length k in sequences of independent binary random vairables.
*Commun. Statist. Theor. Meth.*, 44:3865–3884, 2015.

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

[15] A. N. Philippou, C. Georghiou, and G. N. Philippou.
A generalized geometric distribution and some of its properties.
*Statist. Probab. Letters*, 1:171–175, 1983.

# Reference IV


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

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

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

[19] V. R. R.Uppuluri and S. A. Patil.
Waiting times and generalized Fibonacci sequences.
*Fibonacci Quart.*, 21:242–249, 1983.

[20] H. Takahashi.
The explicit formulae for the distributions of nonoverlapping words and its applications to statistical tests for pseudo random numbers, May 2021.
Arxiv 2105.05172.

# Reference V

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