An easy criterion for randomness

(Sankhya A 77-1 (2015), pp.126-152, corrected after published) *†

Teturo KAMAE and Yu-Mei XUE

Abstract

We propose a single valued criterion for randomness of a binary sequence $x_1x_2\cdots x_n \in \{0,1\}^n$ defined by

$$\Sigma^n(x_1x_2\cdots x_n) = \sum_{\xi\in\{0,1\}^+} |x_1x_2\cdots x_n|_\xi^2,$$

where $\{0,1\}^+ = \cup_{k=1}^\infty \{0,1\}^k$ is the set of nonempty finite sequences over $\{0,1\}$ and for $\xi \in \{0,1\}^k$,

$$|x_1x_2\cdots x_n|_\xi = \#\{i;\ 1 \le i \le n-k+1,\ x_ix_{i+1}\cdots x_{i+k-1} = \xi\}.$$

We prove that

$$\lim_{n\to\infty} n^{-2}\Sigma^n(X_1X_2\cdots X_n) = 3/2$$

holds with probability 1 if $X_1X_2\cdots X_n$ is an i.i.d. process with $P(X_i = 0) = P(X_i = 1) = 1/2$. Moreover, if a sample path $x_1x_2\cdots$ satisfies this almost all condition, then it is a normal number in the sense of E. Borel, but this converse is not true. We also propose a method to generate infinite sequences $x_1x_2\cdots$ satisfying this almost all condition, which are found out to be reasonable pseudorandom numbers from the point view of the block frequencies.

1 Introduction

We are interested in finding a function for finite binary sequences which measures quantitatively how random are they. The Kolmogorov-Chaitin complexity [5] is theoretically the best one for this purpose, but it has two shortcomings. First is that it is not a computable function, second is that it has an ambiguity up to adding an arbitrary constant.

*MSC2000 subject classifications: 65C10 (primary), 11K45 (secondary)
†Key words; randomness criterion, pseudorandom number, normal number

In this paper, we propose an easy criterion for the randomness. Namely, for $n = 1, 2, \cdots$, define a function $\Sigma^n : \{0,1\}^n \to \mathbb{R}$ by

$$\Sigma^n(x_1x_2\cdots x_n) = \sum_{\xi\in\{0,1\}^+} |x_1x_2\cdots x_n|_\xi^2,$$

where $\{0,1\}^+ = \sum_{k=1}^{\infty}\{0,1\}^k$ and for $\xi \in \{0,1\}^k$,

$$|x_1x_2\cdots x_n|_\xi = \#\{i;\ 1 \le i \le n-k+1,\ x_ix_{i+1}\cdots x_{i+k-1} = \xi\}.$$

Of course, $|x_1x_2\cdots x_n|_\xi = 0$ for $\xi \in \{0,1\}^k$ with $k > n$.

Since the function $f(x) = x^2$ is convex, the value $\sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2$ for any $k = 1, 2, \cdots$ is smaller if the values $|x_1x_2\cdots x_n|_\xi$ for $\xi \in \{0,1\}^k$ are less deviated as a whole from the mean value $(n-k+1)/2^k$. That is, $\Sigma^n(x_1x_2\cdots x_n)$ is smaller if the block frequencies in the sequence $x_1x_2\cdots x_n$ is more uniform, or in some restricted sense, it is more random. Thus, $\Sigma^n(x_1x_2\cdots x_n)$ represents the degree how uniformly distributed and random in some restricted sense the sequence $x_1x_2\cdots x_n$ is. The smallest possible value of $\Sigma^n(x_1x_2\cdots x_n)$ for a fixed n is attained by $x = x_1x_2\cdots x_n$ satisfying that $||x|_\xi - |x|_\eta| \le 1$ for any $\xi, \eta \in \{0,1\}^+$ with $|\xi| = |\eta|$ if it exists. We call such x an *equi-distributed* word. For an infifitely many special n, we can construct equi-distributed words using De Bruijin sequences, but we don't know whether it exists for any n or not.

We prove that

Theorem 1. (1) *For any $x = x_1x_2\cdots \in \{0,1\}^\infty$ and $n = 1, 2, \cdots$, we have*

$$\begin{aligned}\Sigma^n(x) &\ge n^2 - 2n + 3 - 3\cdot 2^{-n} \\ &\quad + \sum_{k=1}^{n} 2^k\{(n-k+1)2^{-k}\}(1 - \{(n-k+1)2^{-k}\}) \\ &\ge (3/2)n^2 - n\log_2 n - (7/2)n + (5/2).\end{aligned}$$

where $\{t\}$ is the fractional part of a real number t. The equality in the first inequlity is attained if and only if x is an equi-distributed word.

(2) *Let $X_1X_2\cdots$ be a sequence of independent and identically distrubuted random variables with $P(X_i = 0) = P(X_i = 1) = 1/2$ and let $\Sigma^n := \Sigma^n(X_1X_2\cdots X_n)$ $(n = 1, 2, \cdots)$ be a sequence of random variables. Then,*

$$\mathbb{E}[\Sigma^n] = (3/2)n^2 - (5/2)n + 4 - 4\cdot 2^{-n}\ ,\ \mathbb{V}[\Sigma^n] = 12n^2 + O(n)$$

for $n = 1, 2, \cdots$. Hence,

$$\lim_{n\to\infty} (1/n^2)\Sigma^n(X_1X_2\cdots X_n) = 3/2$$

holds with probability 1.

(3) *If $\lim_{n\to\infty} n^{-2}\Sigma^n(x_1x_2\cdots x_n) = 3/2$, then $x_1x_2\cdots \in \{0,1\}^\infty$ is a normal number in the sense of E. Borel. This converse is not true.*

For the other extreme, it is easy to see that

$$\Sigma^n(x_1x_2\cdots x_n) \le (1/3)n^3 + (1/2)n^2 + (1/6)n$$

with the equality attained if and only if $x_1x_2\cdots x_n$ is a constant word. It is proved that

Theorem 2. ([6]) *$x_1x_2\cdots \in \{0,1\}^\infty$ is eventually periodic if and only if $\lim_{n\to\infty}(1/n^3)\Sigma(x_1x_2\cdots x_n)$ exists and take a posisitive value. Moreover, this limit is $1/(3k)$ if k is the minimum period of $x_1x_2\cdots$.*

We also propose an algorithm to generate a binary sequence $x_1x_2\cdots$ starting from an arbitrary finite sequence which satisfies

$$\lim_{n\to\infty}(1/n^2)\Sigma^n(x_1x_2\cdots x_n) = 3/2. \tag{1.1}$$

These sequences are shown to behave very well as pseudorandom numbers as long as the uniformity of block frequencies are concerned. In fact, they are accepted as random numbers by the χ^2-test at any of reasonable levels.

For $x_1x_2\cdots \in \{0,1\}^\infty$, denote $\Sigma_k^n = \sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2$. As for the behavior of Σ_k^n as $n\to\infty$, we consider typical 3 regions of k:
Bernoulli part, where $k = O(1)$,
Poisson part, where $k = O(\log n)$,
Long tail part, where $k = O(n)$.

In the Bernoulli part, the law of large numbers holds. The notion of normal numbers is related only to this part.

In the Poisson part, the Poisson law of small numbers holds. Yuval Peres and Benjamin Weiss [7] discussed this part. They considered $x_1x_2\cdots \in \{0,1\}^\infty$ such that

$$\lim_{k\to\infty} 2^{-k}\#\{\xi\in\{0,1\}^k;\ |x_1x_2\cdots x_{2^k+k-1}|_\xi = m\} = \frac{e^{-1}}{m!} \quad (m = 0,1,2,\cdots) \tag{1.2}$$

holds. They proved that if $X_1X_2\cdots$ is a sequence of independent and identically distributed random variables with $P(X_i = 0) = P(X_i = 1) = 1/2$, then the sample paths satisfy (1.2) with probability 1. Moreover, the sequences satisfying (1.2) are always normal numbers, but this converse is not true.

Our random numbers satisfying (1.1) are required to behave properly in any of these parts. In fact, it is proved in that (1.1) implies

$$\lim_{K\to\infty}\limsup_{n\to\infty}(1/n^2)\sum_{1\le k<K}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2 = 1, \tag{1.3}$$

$$\lim_{K\to\infty}\limsup_{n\to\infty}(1/n^2)\sum_{(1/K)\log n<k<K\log n}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2 = 0, \tag{1.4}$$

and

$$\lim_{K\to\infty}\limsup_{n\to\infty}(1/n^2)\sum_{(1/K)n<k\le n}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2 = 1/2. \tag{1.5}$$

Moreover, it is proved in Corollary 2 that (1.1) is equivalent with (1.3), (1.5) and the following stronger version of (1.4)

$$\lim_{K\to\infty} \limsup_{n\to\infty}(1/n^2) \sum_{K<k<n/K} \sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2 = 0 \qquad (1.4')$$

It is proved in Corollary 3 that (1.3) holds if and only if $x_1x_2\cdots \in \{0,1\}^\infty$ is a normal number. If $x_1x_2\cdots \in \{0,1\}^\infty$ has a repitition of order $O(n)$ like a normal number in Example 1, then (1.5) does not hold. Hence, our random number is more strict than just a normal number. So far, the implications between (1.1) and (1.2) are not known.

For general informations on the probability theory and statistics, refer [2] and [3].

2 Preliminary lemmas

Throughout this paper, $X_1X_2\cdots$ denotes a sequence of independent and identically distributed random variables with $P(X_i = 0) = P(X_i = 1) = 1/2$. Also, n is an arbitrary positive integer.

Definition 1. Let $I = \{i, i+1, \cdots, i+l-1\} \subset \{1, 2, \cdots, n\}$, we call I an *interval* of $\{1, 2, \cdots, n\}$. We denote $|I| = l$. For $x = x_1x_2\cdots x_n \in \{0,1\}^n$, we denote $x[I] = x_ix_{i+1}\cdots x_{i+l-1} \in \{0,1\}^l$.

Definition 2. Let $\Lambda_n = \{\{i,j\};\ i,j \in \{1,2,\cdots,n\}\}$, where $\{i,j\}$ is a one-point set if $i = j$ and a two-point set if $i \neq j$. For a nonempty subset A of Λ_n, let $G(A)$ be the graph with the vertex set $V(A) := \cup_{\{i,j\}\in A}\{i,j\}$ and the edge set A. Note that a one-point set in A is a loop in $G(A)$. For any intervals I, I' of $\{1,2,\cdots,n\}$ with $|I| = |I'|$, define

$$A(I,I') = \{\{i+j, i'+j\};\ j = 0, 1, \cdots, |I|-1\},$$

where i, i' are the first elements of I and I', respectively. Define

$$\tau(A) := \#V(A) - \text{the number of connected components of } G(A).$$

Lemma 1. *For any nonempty subset $A \subset \Lambda_n$, we have*

$$P(X_i = X_j \textit{ for any } \{i,j\} \in A) = 2^{-\tau(A)}.$$

Proof Let $S_1, \cdots, S_k$ be the decomposition of $V(A)$ into the connected components of $G(A)$. Then, the event that $X_i = X_j$ for any $\{i,j\} \in A$ coincides with the event that the process $X_1X_2\cdots$ is constant at each S_i for any $i = 1, \cdots, k$. Since the event that the process $X_1X_2\cdots$ is constant at S_i has probability $2^{-\#S_i+1}$ and these events for $i = 1, \cdots, k$ are independent, we have

$$P(X_i = X_j \text{ for any } \{i,j\} \in A) = 2^{-\#S_1+1} \cdots 2^{-\#S_k+1} = 2^{-\tau(A)}.$$

□

Lemma 2. *For any intervals I, I' of $\{1, 2, \cdots, n\}$ with $\#I = \#I' = k \geq 1$, we have*

$$\tau(A(I, I')) = \begin{cases} k & I \neq I' \\ 0 & I = I' \end{cases}.$$

Proof If $I = I'$, then clearly $\tau(A(I, I')) = 0$. Assume that $I \neq I'$. Then, there exists $d \neq 0$ such that $I = \{i, i+1, \cdots, i+l\}$ and $I' = \{i+d, i+d+1, \cdots, i+l+d\}$. Then, it holds that $V(A(I, I')) = I \cup I'$ and that $i, j \in V(A(I, I'))$ are in a same connected component of $G(A(I, I'))$ if and only if $i \equiv j \ (mod\ d)$. Therefore, there are exactly $\#(I \setminus I')$ number of connected components of $G(A(I, I'))$. Hence,

$$\tau(A(I, I')) = \#(I \cup I') - \#(I \setminus I') = \#I = k.$$

□

Corollary 1. *For any intervals I, I' of $\{1, 2, \cdots, n\}$ with $\#I = \#I' = k \geq 1$, we have*

$$\mathbb{E}[1_{X[I]=X[I']}] = P(X[I] = X[I']) = \begin{cases} 1 & \text{if } I = I' \\ 2^{-k} & \text{if } I \neq I' \end{cases}.$$

Lemma 3. *Let A be a nonempty subset of Λ_n. Then for any $\{u, v\} \in \Lambda_n$ with $u \neq v$, we have*

$$\tau(A \cup \{u, v\}) = \begin{cases} \tau(A) & \text{if } \{u, v\} \subset V(A) \text{ and } u, v \text{ are in the} \\ & \text{same connected component of } G(A) \\ \tau(A) + 1 & \text{else} \end{cases}.$$

Proof There are 4 cases.

Case 1 $\{u, v\} \cap V(A) = \emptyset$: In this case, the number of elements in $V(A \cup \{u, v\})$ increases by 2 from $V(A)$, while the number of connected components in $G(A \cup \{u, v\})$ increases by 1 from $G(A)$. Thus, we have $\tau(A \cup \{(u, v)\}) = \tau(A) + 1$.

Case 2 $\#(\{u, v\} \cap V(A)) = 1$: In this case, the number of elements in $V(A \cup \{(u, v)\})$ increases by 1 from $V(A)$, while the number of connected components in $G(A \cup \{u, v\})$ is unchanged from $G(A)$. Thus, we have $\tau(A \cup \{u, v\}) = \tau(A) + 1$.

Case 3 $\{u, v\} \subset V(A)$ and u, v belong to distinct connected components of $G(A)$: In this case, the number of elements in $V(A \cup \{u, v\})$ is unchanged from $V(A)$, while the number of connected components in $G(A \cup \{u, v\})$ decreases by 1 from $G(A)$. Hence, we have $\tau(A \cup \{u, v\}) = \tau(A) + 1$.

Case 4 $\{u, v\} \in A$: In this case, both of the number of elements in $V(A \cup \{u, v\})$ and the number of connected components in $G(A \cup \{(u, v)\})$ are unchanged. Hence, we have $\tau(A \cup \{u, v\}) = \tau(A)$. □

Definition 3. Let a quadruple of intervals (I, I', J, J') of $\{1, 2, \cdots, n\}$ be such that $|I| = |I'| \geq 1,\ |J| = |J'| \geq 1$. Let the first elements of I, I', J, J' be i, i', j, j', respctively. It is said to be *coupled* if either
(1) $I \cap J \neq \emptyset$, $i - j = i' - j'$ and $(I \cup J) \cap (I' \cup J') = \emptyset$
or
(2) $I \cap J' \neq \emptyset$, $i - j' = i' - j$ and $(I \cup J') \cap (I' \cup J) = \emptyset$.

Definition 4. A quadruple of intervals (I, I', J, J') of $\{1, 2, \cdots, n\}$ is said to be *connected* if there exists a permutation I_1, I_2, I_3, I_4 of I, I', J, J' such that $I_1 \cap I_2 \neq \emptyset$, $I_2 \cap I_3 \neq \emptyset$ and $I_3 \cap I_4 \neq \emptyset$.

Lemma 4. *Let I, I', J, J' be intervals of $\{1, 2, \cdots, n\}$ with $\#I = \#I' = k \geq 1,\ I \neq I',\ \#J = \#J' = l \geq 1$ and $J \neq J'$ such that (I, I', J, J') is neither coupled nor connected. Then, we have $\tau(A(I, I') \cup A(J, J')) = k + l$.*

Proof Under the assumption, the following 4 cases are possible. That is,
(1) $(I \cup I') \cap (J \cup J') = \emptyset$,
(2) one of I, I', J, J' is disjoint from the other 3 intervals
(3) $I \cap J \neq \emptyset$, $I' \cap J' \neq \emptyset$, $i - j \neq i' - j'$ and $(I \cup J) \cap (I' \cup J') = \emptyset$,
or
(4) $I \cap J' \neq \emptyset$, $I' \cap J \neq \emptyset$, $i - j' \neq i' - j$ and $(I \cup J') \cap (I' \cup J) = \emptyset$,
where i, i', j, j' are the first elements of I, I', J, J', respctively.

In the case (1), starting from $A(I, I')$, add $\{j+h, j'+h\}$ for $h = 0, 1, \cdots, l-1$ successively to get $A(I, I') \cup A(J, J')$, where $J = \{j, j+1, \cdots, j+l-1\}$ and $J' = \{j', j'+1, \cdots, j'+l-1\}$. Then, τ-value increases by 1 each times since we always have Case 1 or Case 2 in the proof of Lemma 3. Since $\tau(A(I, I') = k$, we have $\tau(A(I, I') \cup A(J, J')) = k + l$. In (2), we may assume without loss of generality that I is disjoint from $I' \cup J \cup J'$. Then, by the same argument as above starting from $A(J, J')$, we have $\tau(A(I, I') \cup A(J, J')) = k + l$. In (3) or (4), by the same argument as above with possibly Case 1, Case 2 or Case 3, we get $\tau(A(I, I') \cup A(J, J')) = k + l$. □

Lemma 5. *Let I, I', J, J' be intervals of $\{1, 2, \cdots, n\}$ with $\#I = \#I' = k \geq 1,\ \#J = \#J' = l \geq 1$. Let $\mathbb{C}\mathrm{ov}(1_{X[I]=X[I']}, 1_{X[J]=X[J']})$ be the covariance between $1_{X[I]=X[I']}$ and $1_{X[J]=X[J']}$. Then,*

$$0 \leq \mathbb{C}\mathrm{ov}(1_{X[I]=X[I']}, 1_{X[J]=X[J']}) \leq 2^{-k \vee l}$$

holds in general. Moreover,

$$\mathbb{C}\mathrm{ov}(1_{X[I]=X[I']}, 1_{X[J]=X[J']}) = 0$$

holds if either $I = I'$, $J = J'$, or the quadruple (I, I', J, J') is not coupled and not connected at the same time.

Proof Since by Lemma 1,

$$\begin{aligned}&\mathbb{C}\mathrm{ov}(1_{X[I]=X[I']}, 1_{X[J]=X[J']})\\&= \mathbb{E}[1_{X[I]=X[I']}1_{X[J]=X[J']}] - \mathbb{E}[1_{X[I]=X[I']}]\mathbb{E}[1_{X[J]=X[J']}]\\&= 2^{-\tau(A(I,I')\cup A(J,J'))} - 2^{-k-l},\end{aligned}\tag{2.1}$$

and $k \vee l \leq \tau(A(I,I') \cup A(J,J')) \leq k+l$ by Lemma 3, we have

$$0 \leq \mathbb{C}\mathrm{ov}(1_{X[I]=X[I']}, 1_{X[J]=X[J']}) \leq 2^{-k\vee l}.$$

If $I = I'$, then $1_{X[I]=X[I']} \equiv 1$ and hence, $\mathbb{C}\mathrm{ov}(1_{X[I]=X[I']}, 1_{X[J]=X[J']}) = 0$. Similarly in the case $J = J'$. Let $I \neq I'$, $J \neq J'$ and the quadruple (I, I', J, J') is nither coupled nor connected. Then by Lemma 4 and (2.1), we have $\mathbb{C}\mathrm{ov}(1_{X[I]=X[I']}, 1_{X[J]=X[J']}) = 0$. □

Lemma 6. *Let I, I', J, J' be intervals of $\{1, 2, \cdots, n\}$ with $\#I = \#I' = k \geq 1$, $I \neq I'$, $\#J = \#J' = l \geq 1$, $J \neq J'$ and the quadruple (I, I', J, J') is coupled. Then,*

$$\mathbb{C}\mathrm{ov}(1_{X[I]=X[I']}, 1_{X[J]=X[J']}) = 2^{-k-l}(2^{|I\cap J|\vee|I\cap J'|} - 1)$$

holds.

Proof Assume without loss of generality that (1) of Definition 3 holds with

$$\begin{aligned}I &= \{i, i+1, \cdots, i+k-1\},\ I' = \{i', i'+1, \cdots, i'+k-1\},\\J &= \{j, j+1, \cdots, j+l-1\},\ J' = \{j', j'+1, \cdots, j'+l-1\}.\end{aligned}$$

Starting from $A(I, I')$, add $\{j+h, j'+h\}$ for $h = 0, 1, \cdots, l-1$ successively to get $A(I, I') \cup A(J, J')$. Then, τ-value increases by 1 if $j + h \notin I$, while stays at the same value if $j + h \in I$ since $j + h = i + p$, $j' + h = i' + p$ and $\{i + p, i' + p\} \in A(I, I')$ by Lemma 3. Hence,

$$\tau(A(I,I') \cup A(J,J')) = \tau(A(I,I') + l - |I \cap J| = k + l - |I \cap J|.$$

Thus,

$$\begin{aligned}&\mathbb{C}\mathrm{ov}(1_{X[I]=X[I']}, 1_{X[J]=X[J']}) = 2^{-\tau(A(I,I')\cup A(J,J'))} - 2^{-k}2^{-l}\\&= 2^{-k-l}(2^{|I\cap J|} - 1) = 2^{-k-l}(2^{|I\cap J|\vee|I\cap J'|} - 1).\end{aligned}$$

The prrof is same in the other case (2) of Definition 3. □

3 Probablistic results

Let $\Sigma^n := \Sigma^n(X_1 X_2 \cdots X_n)$ $(n = 1, 2, \cdots)$ be the random variables. Since $|X_1 X_2 \cdots X_n|_\xi$ is increasing in n, Σ^n is increasing in n.

Lemma 7. *We have* $\mathbb{E}[\Sigma^n] = (3/2)n^2 - (5/2)n + 4 - 4 \cdot 2^{-n}$.

Proof Using Corollary 1, we have

$$\begin{aligned}
\mathbb{E}[\Sigma^n] &= \sum_{k=1}^{n} \sum_{\xi \in \{0,1\}^k} \mathbb{E}[|X_1 X_2 \cdots X_n|_\xi^2] \\
&= \sum_{k=1}^{n} \sum_{\xi \in \{0,1\}^k} \mathbb{E}\left[\left(\sum_{i=1}^{n-k+1} 1_\xi(X_i X_{i+1} \cdots X_{i+k-1})\right)^2\right]
\end{aligned}$$

$$\begin{aligned}
&= \sum_{k=1}^{n} \sum_{\xi \in \{0,1\}^k} \mathbb{E}\left[\sum_{i,j=1}^{n-k+1} 1_\xi(X_i X_{i+1} \cdots X_{i+k-1}) 1_\xi(X_j X_{j+1} \cdots X_{j+k-1})\right] \\
&= \sum_{k=1}^{n} \sum_{i,j=1}^{n-k+1} \mathbb{E}\left[\sum_{\xi \in \{0,1\}^k} 1_\xi(X_i X_{i+1} \cdots X_{i+k-1}) 1_\xi(X_j X_{j+1} \cdots X_{j+k-1})\right] \\
&= \sum_{k=1}^{n} \sum_{i,j=1}^{n-k+1} \mathbb{E}\left[1_{X_i X_{i+1} \cdots X_{i+k-1} = X_j X_{j+1} \cdots X_{j+k-1}}\right] \\
&= \sum_{k=1}^{n} \left(\sum_{i=1}^{n-k+1} 1 + \sum_{\substack{i,j=1 \\ i \neq j}}^{n-k+1} 2^{-k}\right) \\
&= \sum_{k=1}^{n} (n - k + 1 + (n - k)(n - k + 1)2^{-k}) \\
&= (3/2)n^2 - (5/2)n + 4 - 4 \cdot 2^{-n}.
\end{aligned}$$

□

To estimate the variance $\mathbb{V}[\Sigma^n]$, we need the following lemma.

Lemma 8. *We have*

$$\mathbb{V}[\Sigma^n] = \sum_{\substack{I, I', J, J' : \text{intervals of } \{1,2,\cdots,n\} \\ \#I = \#I' \geq 1,\ \#J = \#J' \geq 1}} \mathbb{C}\mathrm{ov}(1_{X[I] = X[I']},\ 1_{X[J] = X[J']}).$$

Proof We have

$$\mathbb{E}[(\Sigma^n)^2] = \sum_{k,l=1}^{n} \sum_{\substack{\xi\in\{0,1\}^k \\ \eta\in\{0,1\}^l}} \mathbb{E}[|X_1X_2\cdots X_n|_\xi^2 \; |X_1X_2\cdots X_n|_\eta^2]$$

$$= \sum_{k,l=1}^{n} \sum_{\substack{\xi\in\{0,1\}^k \\ \eta\in\{0,1\}^l}} \mathbb{E}\left[\left(\sum_{i=1}^{n-k+1} 1_\xi(X_iX_{i+1}\cdots X_{i+k-1})\right)^2 \times \left(\sum_{i=1}^{n-l+1} 1_\eta(X_iX_{i+1}\cdots X_{i+l-1})\right)^2\right]$$

$$= \sum_{k,l=1}^{n} \sum_{\substack{\xi\in\{0,1\}^k \\ \eta\in\{0,1\}^l}} \mathbb{E}\left[\sum_{i,i'=1}^{n-k+1} 1_\xi(X_i\cdots X_{i+k-1})1_\xi(X_{i'}\cdots X_{i'+k-1}) \times \sum_{j,j'=1}^{n-l+1} 1_\eta(X_j\cdots X_{j+l-1})1_\eta(X_{j'}\cdots X_{j'+l-1})\right]$$

$$= \sum_{k,l=1}^{n} \sum_{i,i'=1}^{n-k+1} \sum_{j,j'=1}^{n-l+1} \mathbb{E}\left[\sum_{\xi\in\{0,1\}^k} 1_\xi(X_i\cdots X_{i+k-1})1_\xi(X_{i'}\cdots X_{i'+k-1}) \times \sum_{\eta\in\{0,1\}^l} 1_\eta(X_j\cdots X_{j+l-1})1_\eta(X_{j'}\cdots X_{j'+l-1})\right]$$

$$= \sum_{k,l=1}^{n} \sum_{i,i'=1}^{n-k+1} \sum_{j,j'=1}^{n-l+1} \mathbb{E}\left[1_{X_i\cdots X_{i+k-1}=X_{i'}\cdots X_{i'+k-1}} 1_{X_j\cdots X_{j+l-1}=X_{j'}\cdots X_{j'+l-1}}\right]$$

$$= \sum_{\substack{I,I',J,J':\textit{intervals of } \{1,2,\cdots,n\} \\ \#I=\#I'\geq 1,\ \#J=\#J'\geq 1}} \mathbb{E}[1_{X[I]=X[I']}1_{X[J]=X[J']}].$$

On the other hand,

$$\mathbb{E}[\Sigma^n]^2 = \left(\sum_{\substack{I,I':\textit{intervals of } \{1,2,\cdots,n\} \\ \#I=\#I'\geq 1}} \mathbb{E}[1_{X[I]=X[I']}]\right)^2$$

$$= \sum_{\substack{I,I',J,J':\textit{intervals of } \{1,2,\cdots,n\} \\ \#I=\#I'\geq 1,\ \#J=\#J'\geq 1}} \mathbb{E}[1_{X[I]=X[I']}]\mathbb{E}[1_{X[J]=X[J']}].$$

Hence, we have

$$\begin{aligned}
\mathbb{V}[\Sigma^n] &= \mathbb{E}[(\Sigma^n)^2] - \mathbb{E}[\Sigma^n]^2 \\
&= \sum_{\substack{I,I',J,J': \textit{inteval of } \{1,2,\cdots,n\} \\ \#I=\#I'\geq 1,\ \#J=\#J'\geq 1}} \Big(\mathbb{E}[1_{X[I]=X[I']}1_{X[J]=X[J']}] \\
&\qquad\qquad -\mathbb{E}[1_{X[I]=X[I']}]\mathbb{E}[1_{X[J]=X[J']}]\Big) \\
&= \sum_{\substack{I,I',J,J': \textit{intervals of } \{1,2,\cdots,n\} \\ \#I=\#I'\geq 1,\ \#J=\#J'\geq 1}} \mathbb{C}\mathrm{ov}(1_{X[I]=X[I']},\ 1_{X[J]=X[J']}).
\end{aligned}$$

□

Lemma 9. (suggested by [8]) *We have* $\mathbb{V}[\Sigma^n] = 12n^2 + O(n)$.

Proof By Lemmas 5 and 8, we have

$$\begin{aligned}
\mathbb{V}[\Sigma^n] &= \sum_{\substack{I,I',J,J': \textit{intervals of } \{1,2,\cdots,n\} \\ \#I=\#I'\geq 1,\ \#J=\#J'\geq 1}} \mathbb{C}\mathrm{ov}(1_{X[I]=X[I']},\ 1_{X[J]=X[J']}) \\
&= \sum_{\substack{I,I',J,J': \textit{intervals of } \{1,2,\cdots,n\} \\ \#I=\#I'\geq 1,\ \#J=\#J'\geq 1 \\ (I,I',J,J') \textit{ is coupled}}} \mathbb{C}\mathrm{ov}(1_{X[I]=X[I']},\ 1_{X[J]=X[J']}) \\
&\quad + \sum_{\substack{I,I',J,J': \textit{intervals of } \{1,2,\cdots,n\} \\ \#I=\#I'\geq 1,\ \#J=\#J'\geq 1 \\ I\neq I',J\neq J',\ (I,I',J,J') \textit{ is connected}}} \mathbb{C}\mathrm{ov}(1_{X[I]=X[I']},\ 1_{X[J]=X[J']}) \\
&=: H_1 + H_2
\end{aligned}$$

We evaluate $H_2 \geq 0$ from above using Lemma 5.

$$\begin{aligned}
H_2 &\leq \sum_{k,l=1}^{n} 2^{-k\vee l}\#\{(I,I',J,J'); \\
&\qquad |I| = |I'| = k,\ |J| = |J'| = l,\ (I,I',J,J') \text{ is connected}\} \\
&\leq 2\sum_{k=1}^{n}\sum_{l=1}^{k} 2^{-k}\#\{(I,I',J,J'); \\
&\qquad |I| = |I'| = k,\ |J| = |J'| = l,\ (I,I',J,J') \text{ is connected}\} \\
&\leq 2\sum_{k=1}^{n}\sum_{l=1}^{k} 2^{-k}\#\{(I,I',J,J');\ I = \{i,i+1,\cdots,i+k-1\} \subset \{1,2,\cdots,n\}, \\
&|I'| = k,\ |J| = |J'| = l,\ I'\cup J\cup J' \subset \{i-3k+3, i-3k+4,\cdots,i+4k-4\}\} \\
&\leq 2\sum_{k=1}^{n}\sum_{l=1}^{k} 2^{-k}\#\{I \subset \{1,2,\cdots,n\};\ |I| = k\} \times (7k-6)^3 \\
&\leq 2\sum_{k=1}^{n} k2^{-k}n(7k-6)^3 \leq 2n\sum_{k=1}^{\infty} 7^3k^42^{-k} = O(n).
\end{aligned}$$

On the other hand, by Lemma 6

$$\begin{aligned}
H_1 &= 2\sum_{k,l=1}^{n} \sum_{\substack{I,I',J,J': \textit{intervals of } \{1,2,\cdots,n\} \\ \#I=\#I'=k,\ \#J=\#J'=l \\ (I,I',J,J') \textit{ is coupled},\ I\cap J\neq\emptyset}} 2^{-k-l}(2^{|I\cap J|}-1) \\
&= 2\sum_{k,l=1}^{n} \sum_{\substack{I,I'J,J': \textit{intervals of } \{1,2,\cdots,n\} \\ \#I=\#I'=k,\ \#J=\#J'=j \\ (I,I',J,J') \textit{ is coupled},\ |I\cap J|=|I'\cap J'|=d\geq 1}} 2^{-k-l}(2^{d}-1) \\
&= 4\sum_{\substack{k,l=1 \\ k>l}}^{n} n^2 2^{-k-l}(2\sum_{d=1}^{l-1}(2^d-1)+(k-l+1)(2^l-1)) \\
&\qquad +2\sum_{k=1}^{n} n^2 2^{-2k}(2\sum_{d=1}^{k-1}(2^d-1)+(2^k-1))+O(n) \\
&= 12n^2+O(n).
\end{aligned}$$

□

Proof of (2) of Theorem 1: By Lemmas 6 and 8, we already have

$$\mathbb{E}[\Sigma^n] = (3/2)n^2 - (5/2)n + 4 - 4\cdot 2^{-n}, \tag{3.1}$$
$$\text{and } \mathbb{V}[\Sigma^n] = 12n^2 + O(n).$$

By Chebyshev's inequality, we have

$$P(|\Sigma^n - \mathbb{E}[\Sigma^n]| \geq \epsilon_n) \leq (12n^2 + O(n))/\epsilon_n^2$$

for any $\epsilon_n > 0$. Hence with $\epsilon_n = n^{7/4}$, we have

$$P(|(1/n^2)\Sigma^n - (1/n^2)\mathbb{E}[\Sigma^n]| \geq n^{-1/4}) \leq 12n^{-3/2} + O(n^{-5/2})$$

for any $n = 1, 2, \cdots$. Hence,

$$\lim_{n\to\infty} |(1/n^2)\Sigma^n - (1/n^2)\mathbb{E}[\Sigma^n]| = 0$$

with probability 1. Thus by (3.1), we have

$$\lim_{n\to\infty} (1/n^2)\Sigma^n = 3/2$$

with probability 1. □

4 Combinatorial results

Definition 5. A binary word $x_1x_2\cdots x_n \in \{0,1\}^n$ $(n = 1, 2, \cdots)$ is called an *equi-distributed* word if $||x_1x_2\cdots x_n|_\xi - |x_1x_2\cdots x_n|_\eta| \leq 1$ for any $\xi, \eta \in \{0,1\}^+$ with $|\xi| = |\eta|$.

Lemma 10. *For any $x_1x_2\cdots x_n \in \{0,1\}^n$, we have*

$$\Sigma^n(x_1x_2\cdots x_n) \geq n^2 - 2n + 3 - 3\cdot 2^{-n} + \sum_{k=1}^{n} 2^k\{(n-k+1)2^{-k}\}(1 - \{(n-k+1)2^{-k}\}),$$

where $\{x\}$ is the fractional part of a real number x. Moreover, the equality holds if and only if $x_1x_2\cdots x_n$ is an equi-distributed word.

Proof For any $k = 1, 2, \cdots, n$, $\sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2$ is smaller if

$$\sum_{\xi\in\{0,1\}^k} \Big(|x_1x_2\cdots x_n|_\xi - (n-k+1)/2^k\Big)^2$$

is smaller. Therefore, a lower bound of it is taken when

$$|x_1x_2\cdots x_n|_\xi \text{ is either } \lfloor (n-k+1)/2^k \rfloor \text{ or } \lceil (n-k+1)/2^k \rceil$$

for any $\xi \in \{0,1\}^k$, which holds if and only if $x_1x_2\cdots x_n$ is an equi-distributed word. This implies that $|x_1x_2\cdots x_n|_\xi = \lfloor (n-k+1)/2^k \rfloor$ for $(1 - \{(n-k+1)/2^k\})2^k$ number of $\xi \in \{0,1\}^k$ and $|x_1x_2\cdots x_n|_\xi = \lfloor (n-k+1)/2^k \rfloor + 1$ for $\{(n-k+1)/2^k\}2^k$ number of $\xi \in \{0,1\}^k$.

Hence, we have

$$\begin{aligned}
&\sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2 \\
&\geq \lfloor (n-k+1)/2^k \rfloor^2 (1 - \{(n-k+1)/2^k\})2^k \\
&\qquad + (\lfloor (n-k+1)/2^k \rfloor + 1)^2 \{(n-k+1)/2^k\}2^k \\
&= \lfloor (n-k+1)/2^k \rfloor^2 2^k + 2\lfloor (n-k+1)/2^k \rfloor \{(n-k+1)/2^k\}2^k \\
&\qquad + \{(n-k+1)/2^k\}2^k \\
&= (\lfloor (n-k+1)/2^k \rfloor + \{(n-k+1)/2^k\})^2 2^k \\
&\qquad + \{(n-k+1)/2^k\}(1 - \{(n-k+1)/2^k\})2^k \\
&= (n-k+1)^2/2^k + \{(n-k+1)/2^k\}(1 - \{(n-k+1)/2^k\})2^k.
\end{aligned}$$

Since

$$\sum_{k=1}^{n} (n-k+1)^2/2^k = n^2 - 2n + 3 - 3\cdot 2^{-n},$$

we complete the proof. □

Remark 1. The following words are equi-distributed:

0, 01, 010, 0110, 00110, 001101, 0011010, 01001110
011100010, 0101110001, 01011100010, 010111000110

Let $x_1x_2\cdots x_{n-1}$ be an equi-distributd word of length $n-1$. Take a De Bruijin word (see [1]) of period 2^n starting with $x_1x_2\cdots x_{n-1}$, that is, a periodic word of period 2^n:

$$(x_1x_2\cdots x_{n-1}x_n\cdots x_{2^n})^\infty$$

such that each word in $\{0,1\}^n$ appears just once at a position with indices from 1 to 2^n. Hence, each word in $\{0,1\}^n$ appears just once in

$$x = x_1x_2\cdots x_{n-1}x_n\cdots x_{2^n}x_1x_2\cdots x_{n-1}.$$

We prove that x is an equi-distributed word of length 2^n+n-1. For any ξ with $|\xi|\ge n$, it holds that $|x|_\xi \le 1$, which implies that for any $\xi,\eta\in\{0,1\}^+$ with $|\xi|=|\eta|\ge n$, we have $||x|_\xi - |x|_\eta| \le 1$. On the other hand, let $\xi\in\{0,1\}^+$ be $|\xi|=k<n$. Then, it holds that

$$\#\{i;\ 1\le i\le 2^n,\ x_ix_{i+1}\cdots x_{i+k-1}=\xi\} = 2^{n-k}.$$

Therefore,

$$\#\{i;\ 1\le i\le 2^n+n-1,\ x_ix_{i+1}\cdots x_{i+k-1}=\xi\} = 2^{n-k} + |x_1x_2\cdots x_{n-1}|_\xi,$$

which implies that for any $\xi,\eta\in\{0,1\}^+$ with $|\xi|=|\eta|<n$, we have $||x|_\xi - |x|_\eta| \le 1$ since $x_1x_2\cdots x_{n-1}$ is equi-distributed. Thus, x is equi-distributed.

In this way, we can construct an equi-distributed word of length 2^n+n-1 if there exists an equi-distributed word of length $n-1$. But we don't know whether for any n, an equi-distributed word of length n exists on not.

Proof of (1) of Theorem 1: Let us estimate

$$S := \sum_{k=1}^{n} 2^k\{(n-k+1)2^{-k}\}(1-\{(n-k+1)2^{-k}\}).$$

If $n<2^k$, then $\{(n-k+1)2^{-k}\} = (n-k+1)2^{-k}$, and hence,

$$\begin{aligned}&2^k\{(n-k+1)2^{-k}\}(1-\{(n-k+1)2^{-k}\})\\&=(n-k+1)(1-(n-k+1)2^{-k}).\end{aligned}$$

Therefore,

$$
\begin{aligned}
S &\geq \sum_{k=\lfloor \log_2 n \rfloor+1}^{n} 2^k\{(n-k+1)2^{-k}\}(1-\{(n-k+1)2^{-k}\}) \\
&= \sum_{k=\lfloor \log_2 n \rfloor+1}^{n} (n-k+1)(1-(n-k+1)2^{-k}) \\
&= (1/2)(n-\lfloor \log_2 n \rfloor)(n-\lfloor \log_2 n \rfloor+1) - \sum_{k=\lfloor \log_2 n \rfloor+1}^{n} (n-k+1)^2 2^{-k} \\
&\geq (1/2)(n-\log_2 n)(n-\log_2 n+1) - \sum_{k=1}^{n-\lfloor \log_2 n \rfloor} (n-\lfloor \log_2 n \rfloor-k+1)^2 2^{-\lfloor \log_2 n \rfloor-k} \\
&\geq (1/2)(n-\log_2 n)(n-\log_2 n+1) - 2^{-\lfloor \log_2 n \rfloor} \sum_{k=1}^{n} (n-k+1)^2 2^{-k}
\end{aligned}
$$

$$
\begin{aligned}
&\geq (1/2)(n-\log_2 n)(n-\log_2 n+1) - 2n^{-1} \sum_{k=1}^{n} (n-k+1)^2 2^{-k} \\
&= (1/2)(n-\log_2 n)(n-\log_2 n+1) - 2n + 4 - (6/n)(1-2^{-n}) \\
&= (1/2)n^2 - n\log_2 n - (3/2)n + (1/2)(\log_2 n)(\log_2 n - 1) \\
&\qquad\qquad +4 - (6/n)(1-2^{-n}).
\end{aligned}
$$

Then by Lemma 10, we have

$$\Sigma^n(x_1x_2\cdots x_n) \geq (3/2)n^2 - n\log_2 n - (7/2)n + C(n)$$

with

$$C(n) = (1/2)(\log_2 n)(\log_2 n - 1) + 7 - (6/n)(1-2^{-n}) - 3\cdot 2^{-n},$$

Since $C(n) \geq 5/2$, we have

$$\Sigma^n(x_1x_2\cdots x_n) \geq (3/2)n^2 - n\log_2 n - (7/2)n + (5/2).$$

□

Lemma 11. *For any $x_1x_2\cdots \in \{0,1\}^\infty$, we have*
(1)

$$\lim_{K\to\infty} \liminf_{n\to\infty} (1/n^2) \sum_{k=1}^{K} \sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2 \geq 1,$$

(2)

$$\lim_{K\to\infty} \liminf_{n\to\infty} (1/n^2) \sum_{k=\lfloor n/K \rfloor}^{n} \sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2 \geq 1/2.$$

Proof (1) For $k = 1, 2, \cdots$, we have

$$\begin{aligned}
&\sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2 \\
&= \sum_{\xi\in\{0,1\}^k} \Big((|x_1x_2\cdots x_n|_\xi - (n-k+1)2^{-k})^2 + ((n-k+1)2^{-k})^2 \Big) \\
&\geq \sum_{\xi\in\{0,1\}^k} ((n-k+1)2^{-k})^2 = ((n-k+1)2^{-k})^2 \; 2^k
\end{aligned}$$

Hence,

$$\begin{aligned}
&\liminf_{n\to\infty} \; (1/n^2) \sum_{k=1}^{K} \sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2 \\
&\geq \lim_{n\to\infty} \; (1/n^2) \sum_{k=1}^{K} ((n-k+1)2^{-k})^2 \; 2^k \\
&= \sum_{k=1}^{K} 2^{-k} = 1 - 2^{-K}
\end{aligned}$$

Letting $K \to \infty$, we have (1).

(2) There exists n_0 depending only on K such that $k \geq \lfloor n/K \rfloor$ implies $(n-k+1)/2^k < 1$ for any $n \geq n_0$. Assume that $n \geq n_0$. Then, a lower bound of $B_k := \sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2$ is attained if $|x_1x_2\cdots x_n|_\xi$ is either 0 or 1 for any $\xi \in \{0,1\}^k$. In this case, $B_k = n - k + 1$. Therefore,

$$\sum_{k=\lfloor n/K \rfloor}^{n} B_k \geq \sum_{k=\lfloor n/K \rfloor}^{n} (n-k+1) = (1/2)(n - \lfloor n/K \rfloor + 1)(n - \lfloor n/K \rfloor + 2).$$

Hence,

$$\lim_{K\to\infty} \liminf_{n\to\infty} \; (1/n^2) \sum_{k=\lfloor n/K \rfloor}^{n} \sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2 \geq 1/2.$$

□

Corollary 2. *The property* (1.1) *for* $x_1x_2\cdots \in \{0,1\}^\infty$ *is equivalent with the properties* (1.3)*,* (1.5) *and* (1.4′)*.*

Proof Assume (1.1) for $x_1x_2\cdots \in \{0,1\}^\infty$. Then by Lemma 11, we have

$$\begin{aligned}
3/2 \leq \lim_{K\to\infty} \Bigg(&\liminf_{n\to\infty} \; (1/n^2) \sum_{k=1}^{K} \sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2 \\
&+ \liminf_{n\to\infty} \; (1/n^2) \sum_{k=\lfloor n/K \rfloor}^{n} \sum_{\xi\in\{0,1\}^k} |x_1x_2\cdots x_n|_\xi^2 \Bigg)
\end{aligned}$$

$$\leq \lim_{K\to\infty} \left(\liminf_{n\to\infty} \ (1/n^2) \sum_{k=1}^{K} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2 \right.$$

$$+ \limsup_{n\to\infty} \ (1/n^2) \sum_{k=K+1}^{\lfloor n/K \rfloor -1} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2$$

$$\left. + \liminf_{n\to\infty} \ (1/n^2) \sum_{k=\lfloor n/K \rfloor}^{n} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2 \right)$$

$$\leq \limsup_{n\to\infty} \ (1/n^2) \sum_{k=1}^{n} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2 = 3/2.$$

Hence, we have

$$\lim_{K\to\infty} \limsup_{n\to\infty} \ (1/n^2) \sum_{k=K+1}^{\lfloor n/K \rfloor -1} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2 = 0,$$

which implies (1.4′). In the same way, we have

$$1 = \lim_{K\to\infty} \liminf_{n\to\infty} \ (1/n^2) \sum_{k=1}^{K} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2$$

$$= \lim_{K\to\infty} \limsup_{n\to\infty} \ (1/n^2) \sum_{k=1}^{K} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2 \ ,$$

since

$$3/2 \leq \lim_{K\to\infty} \left(\liminf_{n\to\infty} \ (1/n^2) \sum_{k=1}^{K} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2 \right.$$

$$\left. + \liminf_{n\to\infty} \ (1/n^2) \sum_{k=\lfloor n/K \rfloor}^{n} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2 \right)$$

$$\leq \lim_{K\to\infty} \left(\limsup_{n\to\infty} \ (1/n^2) \sum_{k=1}^{K} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2 \right.$$

$$\left. + \liminf_{n\to\infty} \ (1/n^2) \sum_{k=\lfloor n/K \rfloor}^{n} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2 \right)$$

$$\leq \limsup_{n\to\infty} \ (1/n^2) \sum_{k=1}^{n} \sum_{\xi \in \{0,1\}^k} |x_1 x_2 \cdots x_n|_\xi^2 = 3/2.$$

Thus, we have (1.3). In the same way, we have (1.5).

Conversely, assume that (1.3), (1.4′) and (1.5). Then by (1) of Theorem 1, we have

$$
\begin{aligned}
3/2 &\le \liminf_{n\to\infty}\ (1/n^2)\sum_{k=1}^{n}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2 \\
&\le \limsup_{n\to\infty}\ (1/n^2)\sum_{k=1}^{n}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2 \\
&\le \lim_{K\to\infty}\Bigg(\limsup_{n\to\infty}\ (1/n^2)\sum_{k=1}^{K}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2 \\
&\qquad + \limsup_{n\to\infty}\ (1/n^2)\sum_{k=K+1}^{\lfloor n/K\rfloor-1}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2 \\
&\qquad + \limsup_{n\to\infty}\ (1/n^2)\sum_{k=\lfloor n/K\rfloor}^{n}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2\Bigg) = 3/2.
\end{aligned}
$$

Thus, we have (1.1). □

Lemma 12. *For $x_1x_2\cdots\in\{0,1\}^\infty$, if there exist a positive integer h and $\eta\in\{0,1\}^h$ such that*

$$
\limsup_{n\to\infty}\left(\frac{1}{n-h+1}\ |x_1x_2\cdots x_n|_\eta - 2^{-h}\right)^2 = \delta > 0.
$$

Then we have

$$
\lim_{K\to\infty}\limsup_{n\to\infty}(1/n^2)\sum_{k=1}^{K}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2 \ge 1+\delta.
$$

Proof We have

$$
\begin{aligned}
&\lim_{K\to\infty}\limsup_{n\to\infty}(1/n^2)\sum_{k=1}^{K}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2 \\
&= \lim_{K\to\infty}\limsup_{n\to\infty}(1/n^2)\sum_{k=1}^{K}\sum_{\xi\in\{0,1\}^k} \\
&\qquad\Big((|x_1x_2\cdots x_n|_\xi-(n-k+1)2^{-k})^2+((n-k+1)2^{-k})^2\Big)
\end{aligned}
$$

$$\geq \limsup_{n\to\infty}(1/n^2)\left(|x_1x_2\cdots x_n|_\eta - (n-h+1)2^{-h}\right)^2$$

$$+ \lim_{K\to\infty}\lim_{n\to\infty}(1/n^2)\sum_{k=1}^{K}\sum_{\xi\in\{0,1\}^k}((n-k+1)2^{-k})^2$$

$$= \delta + \sum_{k=1}^{\infty} 2^{-k} = 1+\delta.$$

□

Proof of (3) of Theorem 1: By Corollary 2 and Lemma 12, if

$$\lim_{n\to\infty}(1/n^2)\Sigma^n(x_1x_2\cdots x_n) = 3/2,$$

then

$$\lim_{n\to\infty}\frac{1}{n-h+1}\ |x_1x_2\cdots x_n|_\eta = 2^{-h}$$

for any positive integer h and $\eta \in \{0,1\}^h$. Hence, $x_1x_2\cdots \in \{0,1\}^\infty$ is a normal number. This converse is not true (Example 1). □

Corollary 3. *A sequence $x_1x_2\cdots \in \{0,1\}^\infty$ is a normal number if and only if it satisfies* (1.3).

Proof The "if" part follows from Lemma 12. To prove the "only if" part, assume that $x_1x_2\cdots \in \{0,1\}^\infty$ is a normal number. Then since

$$\limsup_{n\to\infty}(1/n^2)\left(|x_1x_2\cdots x_n|_\xi - (n-h+1)2^{-k}\right)^2 = 0$$

for any $k=1,2,\cdots$ and $\xi \in \{0,1\}^k$, by (1) of Lemma 11, we have

$$1 \leq \lim_{K\to\infty}\liminf_{n\to\infty}(1/n^2)\sum_{k=1}^{K}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2$$

$$\leq \lim_{K\to\infty}\limsup_{n\to\infty}(1/n^2)\sum_{k=1}^{K}\sum_{\xi\in\{0,1\}^k}|x_1x_2\cdots x_n|_\xi^2$$

$$= \lim_{K\to\infty}\limsup_{n\to\infty}(1/n^2)\sum_{k=1}^{K}\sum_{\xi\in\{0,1\}^k}$$

$$\left((|x_1x_2\cdots x_n|_\xi - (n-k+1)2^{-k})^2 + ((n-k+1)2^{-k})^2\right)$$

$$\leq \lim_{K\to\infty}\sum_{k=1}^{K}\sum_{\xi\in\{0,1\}^k}\limsup_{n\to\infty}(1/n^2)\left(|x_1x_2\cdots x_n|_\xi - (n-h+1)2^{-h}\right)^2$$

$$+ \lim_{K\to\infty}\lim_{n\to\infty}(1/n^2)\sum_{k=1}^{K}\sum_{\xi\in\{0,1\}^k}((n-k+1)2^{-k})^2$$

$$= \sum_{k=1}^{\infty} 2^{-k} = 1.$$

Thus, we have (1.3). □

Example 1. Let $x = x_1x_2\cdots\{0,1\}^\infty$ be a normal number. Let $x|_n = x_1x_2\cdots x_n$ and $1 \le m_1 \le m_2 \le \cdots$ be an arbitrary sequence of integers such that $\lim_{n\to\infty} m_n = \infty$. Define $y \in \{0,1\}^\infty$ by the concatenation

$$y = (x|_{m_1})^{l_1}(x|_{m_2})^{l_2}\cdots$$

with positive integers $l_1, l_2, \cdots$. Then, y is a normal number. Let $\varphi : \{1,2,\cdots\} \to (0,1/2)$ be any decreasing function such that $\lim_{n\to\infty}\varphi(n) = 0$. We choose $1 \le m_1 \le m_2 \le \cdots$ so that $m_n \le \varphi(n)^{-1/2}$ $(n = 1,2,\cdots)$. Then, it is possible to take l_n $(n = 1,2,\cdots)$ so that $m_n^3 l_n^3 > N^3\varphi(n)^{1/2}$ with $N = \sum_{i=1}^n m_i l_i$. Then, $m_n^2 l_n^3 > N^3\varphi(n)^{1/2}/m_n \ge N^3\varphi(n)$ holds. Hence,

$$\liminf_{N\to\infty} \Sigma^N(x_1x_2\cdots x_N)/(N^3\varphi(N)) \ge \liminf_{N\to\infty}(1/48)m_n^2 l_n^3/(N^3\varphi(n)) \ge 1/48$$

holds since it is known in [6] that

$$\Sigma^N(x_1x_2\cdots x_N) \ge (1/48)m_n^2 l_n^3.$$

That is, normal number can take the value $\Sigma^N(x_1x_2\cdots x_N)$ arbitrary close to N^3 in the ratio. This fact is suggested by Bo Tan [8] for the first time.

5 Random number genarator

Definition 6. For $x_1x_2\cdots x_n \in \{0,1\}^n$, let

$$\mathrm{Suf}(x_1x_2\cdots x_n) = \{x_ix_{i+1}\cdots x_n;\ i = 1,2,\cdots,n\}$$

be the set of suffixes of $x_1x_2\cdots x_n$. Define

$$D(x_1x_2\cdots x_n) = \#\{(i,j) \in \{1,2,\cdots,n\}^2;\ i \le j,\\ x_ix_{i+1}\cdots x_j \in \mathrm{Suf}(x_1x_2\cdots x_n)\}.$$

For convenience, define $D(\emptyset) = 0$ for the empty sequence $\emptyset$.

Lemma 13. *For any $x_1x_2\cdots x_n \in \{0,1\}^n$, it holds that*

$$\Sigma^n(x_1x_2\cdots x_n) - \Sigma^{n-1}(x_1x_2\cdots x_{n-1}) = 2D(x_1x_2\cdots x_n) - n$$

Proof Note that for any $\xi \in \{0,1\}^k$ with $k = 1,2,\cdots,n$, we have

$$|x_1x_2\cdots x_n|_\xi = \begin{cases} |x_1x_2\cdots x_{n-1}|_\xi + 1 & \text{if } \xi \in \mathrm{Suf}(x_1x_2\cdots x_n) \\ |x_1x_2\cdots x_{n-1}|_\xi & \text{otherwise.} \end{cases}$$

Therefore,

$$|x_1x_2\cdots x_n|_\xi^2 - |x_1x_2\cdots x_{n-1}|_\xi^2 = (2|x_1x_2\cdots x_{n-1}|_\xi + 1)1_{\xi\in \mathrm{Suf}(x_1x_2\cdots x_n)},$$

and hence,

$$\begin{aligned}
&\Sigma^n(x_1x_2\cdots x_n) - \Sigma^{n-1}(x_1x_2\cdots x_{n-1})\\
&= \sum_{\xi\in \mathrm{Suf}(x_1x_2\cdots x_n)} (2|x_1x_2\cdots x_{n-1}|_\xi + 1)\\
&= \sum_{\xi\in \mathrm{Suf}(x_1x_2\cdots x_n)} (2(|x_1x_2\cdots x_n|_\xi - 1) + 1)\\
&= 2D(x_1x_2\cdots x_n) - n.
\end{aligned}$$

□

Lemma 14. *For any $x_1x_2\cdots x_{n-1} \in \{0,1\}^{n-1}$, it holds that*

$$D(x_1x_2\cdots x_{n-1}0) + D(x_1x_2\cdots x_{n-1}1) = D(x_1x_2\cdots x_{n-1}) + 2n$$

Proof For any $(i,j) \in \{1,2,\cdots,n-2\}^2$ with $i \le j$ such that

$$x_ix_{i+1}\cdots x_j \in \mathrm{Suf}(x_1x_2\cdots x_{n-1}),$$

either

$$x_ix_{i+1}\cdots x_jx_{j+1} \in \mathrm{Suf}(x_1x_2\cdots x_{n-1}0)$$

or

$$x_ix_{i+1}\cdots x_jx_{j+1} \in \mathrm{Suf}(x_1x_2\cdots x_{n-1}1)$$

holds depending on $x_{j+1} = 0$ or 1.

Hence, we have

$$\begin{aligned}
&\#\{(i,j) \in \{1,2,\cdots,n-2\}^2;\ i \le j,\ x_ix_{i+1}\cdots x_j \in \mathrm{Suf}(x_1x_2\cdots x_{n-1})\}\\
&= \#\{(i,j) \in \{1,2,\cdots,n-1\}^2;\ i < j,\ x_ix_{i+1}\cdots x_j \in \mathrm{Suf}(x_1x_2\cdots x_{n-1}0)\}\\
&+\#\{(i,j) \in \{1,2,\cdots,n-1\}^2;\ i < j,\ x_ix_{i+1}\cdots x_j \in \mathrm{Suf}(x_1x_2\cdots x_{n-1}1)\}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&D(x_1x_2\cdots x_{n-1}0) + D(x_1x_2\cdots x_{n-1}1)\\
&= n + |x_1x_2\cdots x_{n-1}|_0\\
&+\#\{(i,j) \in \{1,2,\cdots,n-1\}^2;\ i < j,\ x_ix_{i+1}\cdots x_j \in \mathrm{Suf}(x_1x_2\cdots x_{n-1}0)\}\\
&+n + |x_1x_2\cdots x_{n-1}|_1\\
&+\#\{(i,j) \in \{1,2,\cdots,n-1\}^2;\ i < j,\ x_ix_{i+1}\cdots x_j \in \mathrm{Suf}(x_1x_2\cdots x_{n-1}1)\}\\
&= 2n + (n-1)\\
&+\#\{(i,j) \in \{1,2,\cdots,n-2\}^2;\ i \le j,\ x_ix_{i+1}\cdots x_j \in \mathrm{Suf}(x_1x_2\cdots x_{n-1})\}\\
&= 2n + D(x_1x_2\cdots x_{n-1}).
\end{aligned}$$

□

Lemma 15. *For any $m = 0, 1, 2, \cdots$ and $\zeta \in \{0,1\}^m$, there exists $x_1x_2\cdots \in \{0,1\}^\infty$ such that $D(\zeta x_1x_2\cdots x_n) \le 2^{-n}D(\zeta) + 2m + 2n$ holds for any $n = 1, 2, \cdots$.*

Proof We choose x_i for $i = 1, 2, \cdots$ inductively. By Lemma 14, either

$$D(\zeta x_1 \cdots x_{i-1}0) \le (1/2)D(\zeta x_1 \cdots x_{i-1}) + m + i$$

or

$$D(\zeta x_1 \cdots x_{i-1}1) \le (1/2)D(\zeta x_1 \cdots x_{i-1}) + m + i$$

holds. We choos x_i such that

$$D(\zeta x_1 \cdots x_{i-1}x_i) \le (1/2)D(\zeta x_1 \cdots x_{i-1}) + m + i.$$

Multiplying this inequality by 2^{i-n} and adding them for $i = 1, 2, \cdots, n$, we get

$$D(\zeta x_1 \cdots x_n) \le 2^{-n}D(\zeta)+m(1+2^{-1}+\cdots+2^{-n+1})+n+(n-1)2^{-1}+\cdots+2^{1-n}.$$

Thus, we have

$$D(\zeta x_1 \cdots x_n) \le 2^{-n}D(\zeta)+2m+2n-2+(n+1)2^{1-n} \le 2^{-n}D(\zeta)+2m+2n.$$

□

Theorem 3. *For any $m = 0, 1, 2, \cdots$ and $\zeta \in \{0,1\}^m$, there exists $x_1x_2\cdots \in \{0,1\}^\infty$ such that*

$$\Sigma^{m+n}(\zeta x_1x_2\cdots x_n) \le \Sigma^m(\zeta) + 2D(\zeta) + 3mn + (3/2)n(n+1)$$

for any $n = 1, 2, \cdots$. Hence,

$$\lim_{n\to\infty}(1/(m+n)^2)\Sigma^{m+n}(\zeta x_1x_2\cdots x_n) = 3/2.$$

Moreover, $x_1x_2\cdots \in \{0,1\}^\infty$ as this is obtained inductively by choosing x_i so that

$$D(\zeta x_1x_2\cdots x_{i-1}x_i) \le D(\zeta x_1x_2\cdots x_{i-1}\overline{x_i}), \tag{5.1}$$

where we denote $\overline{0} = 1$ and $\overline{1} = 0$.

Proof We take $x_1x_2\cdots \in \{0,1\}^\infty$ satisfying the condition in Lemma 15. Then by Lemma 13, we have

$$\begin{aligned}&\Sigma^{m+i}(\zeta x_1x_2\cdots x_i) - \Sigma^{m+i-1}(\zeta x_1x_2\cdots x_{i-1})\\ &\le 2(2^{-i}D(\zeta) + 2m + 2i) - (m+i) = 2^{-i+1}D(\zeta) + 3m + 3i.\end{aligned}$$

Adding these inequalities for $i = 1, 2, \cdots, n$, we have

$$\begin{aligned}&\Sigma^{m+n}(\zeta x_1x_2\cdots x_n) - \Sigma^m(\zeta)\\ &\le (1 + 2^{-1} + \cdots + 2^{-n+1})D(\zeta) + 3mn + 3(1 + 2 + \cdots + n)\\ &\le 2D(\zeta) + 3mn + (3/2)n(n+1).\end{aligned}$$

Thus, we have

$$\Sigma^{m+n}(\zeta x_1 x_2 \cdots x_n) \le \Sigma^m(\zeta) + 2D(\zeta) + 3mn + (3/2)n(n+1)$$

for any $n = 1, 2, \cdots$.

The rest parts of Theorem 2 follows from (1) of Theorem 1 and (5.1). □

6 Experimental results

We consider the following algorithm to generate random numbers in our sense.

Algorithm: Starting with any $\zeta \in \{0,1\}^k$ $(k = 0, 1, 2, \cdots)$, we construct $x_1 x_2 \cdots \in \{0,1\}^\infty$ inductively by defining x_n so that

$$x_n = \begin{cases} 0 & \text{if } D(\zeta x_1 x_2 \cdots x_{n-1} 0) \le D(\zeta x_1 x_2 \cdots x_{n-1} 1) \\ 1 & \text{if } D(\zeta x_1 x_2 \cdots x_{n-1} 0) > D(\zeta x_1 x_2 \cdots x_{n-1} 1). \end{cases}$$

From the sequence $\zeta x_1 x_2 \cdots$, we remove the initial nonrandom part and take $x_{3k+1} x_{3k+2} \cdots$ as our random number. By Theorem 2, we have

$$\lim_{n\to\infty} (1/n^2)\Sigma^n(x_{3k+1} x_{3k+2} \cdots x_{3k+n}) = 3/2.$$

Example 2. Let

$$x_1 x_2 \cdots = 0100110111000101000111101001011000011101100100 \cdots$$

be the sequence constructed by the above algorithm starting with $\zeta = \emptyset$ (empty sequence). Take the first 50000 digits of it and divide it into 10000 number of blocks of length 5 like

(01001)(10111)(00010)(10001)(11101)(00101)(10000)(11101)(10010)(0 $\cdots$

Then, the frequencies of blocks of length 5 in this list are obtained as

00000	00001	00010	00100	01000	10000	00011	00101	00110
318	329	304	324	306	305	311	291	291

01001	01010	01100	10001	10010	10100	11000	00111	01011
342	322	306	326	302	317	301	323	322

01101	01110	10011	10101	10110	11001	11010	11100	01111
306	315	302	317	297	301	307	339	303

10111	11011	11101	11110	11111	Total
320	313	295	323	322	10000

Then,

$$T = \frac{(318 - 10000 \cdot 2^{-5})^2}{10000 \cdot 2^{-5}} + \frac{(329 - 10000 \cdot 2^{-5})^2}{10000 \cdot 2^{-5}} + \cdots + \frac{(322 - 10000 \cdot 2^{-5})^2}{10000 \cdot 2^{-5}} = 16.3776.$$

If T is χ^2_{31}−distributed, then $P(T > 16.3776) = 0.985$. Hence, the randomness hypothesis for $x_1 x_2 \cdots x_{50000}$ is accepted at any of reasonable levels.

It is uniformly distributed so well that the random walk path W_t associated to it as $W_t = \sum_{i \le t}(1 - 2x_i)$ looks like staying in a bounded region as $t \to \infty$ (see Figure 1), while it is expected to diffuse in the order of $\sqrt{t}$ if $x_1 x_2 \cdots$ is random. Thus, our sequence $x_1 x_2 \cdots$ is an excellen pseudoramdom number as long as the uniformity of block frequencies is concerned. But since it is too much uniform to be expected from a random sequence, if adequate dispersion caused by randomness is concerned, it is not a good pseudo-ramdom number.

Figure 1: Random walk $W_t = \sum_{i \le t}(1 - 2x_i)$ ($t \le 50000$)

Example 3. Take 2 different types of binary sequences ζ and ζ' with an identical length k. Construct the binary sequences $x = x_{3k+1} x_{3k+2} \cdots$ and $y = y_{3k+1} y_{3k+2} \cdots$ by the above algorithm with starter ζ and ζ', respectively. We expect that x and y are independent pseudorandom numbers. In fact, let

$$\zeta = 0110100110010110100101100110100 1\cdots 01$$

be the first 100 digits of the Thue-Morse sequence [4] and

$$\zeta' = 0100101001001010010100100101001 0\cdots 00$$

be the first 100 digits of the Fibonacci sequence [4] and construct x and y as above. Take the first 9699 digits of them and devide them into 3233 number of blocks of length 3. Consider the 3233 number of the combinations

$$\begin{pmatrix} x_{301}x_{302}x_{303} \\ y_{301}y_{302}y_{303} \end{pmatrix}, \begin{pmatrix} x_{304}x_{305}x_{306} \\ y_{304}y_{305}y_{306} \end{pmatrix}, \cdots, \begin{pmatrix} x_{9697}x_{9698}x_{9699} \\ y_{9697}y_{9698}y_{9699} \end{pmatrix}.$$

The frequencies of $\begin{pmatrix} 000 \\ 000 \end{pmatrix}, \begin{pmatrix} 000 \\ 001 \end{pmatrix}, \cdots, \begin{pmatrix} 111 \\ 111 \end{pmatrix}$ in this list are 53, 54, $\cdots$, 56. Thus, we obtain

$$T = \frac{(53 - 3233 \cdot 2^{-6})^2}{3233 \cdot 2^{-6}} + \frac{(54 - 3233 \cdot 2^{-6})^2}{3233 \cdot 2^{-6}} + \cdots + \frac{(56 - 3233 \cdot 2^{-6})^2}{3233 \cdot 2^{-6}} = 47.5889.$$

If T is χ^2_{63}−distributed, then $P(T > 47.5889) = 0.925$. Therefore, the hypothesis that x and y are independent random numbers is accepted at any of reasonable levels.

Acknowledgement: The authors thank Professor Benjamin Weiss (Hebrew University of Jerusalem) for useful suggestions. The authors also thank Professor Bo Tan (Huazhong University of Science and Technology) for useful long discussions with one of the authors which improve the paper a lot. In fact, Lemma 9 and Example 1 are suggested by him.

References

[1] De Bruijn, N. G., "A combinatorial problem", Indagationes Mathematicae 8 (1946), pp.461-467.

[2] I. Guttman and S. S. Wilks, Introductory Engineering Statistics, John Wiley & Sons, New York, 1965.

[3] W. Feller, An Introduction to Probability Theory and Its Applications, Vol. 2, John Wiley & Sons, New York, 1971.

[4] V. Berthe, S. Ferenczi, C. Mauduit and A. Siegel (ed.), Substitutions in dynamics, arithmetics and combinatorics, Lecture Notes in Mathematics 1794, Springer-Verlag, 2002.

[5] M. Li and P. Vitányi, An introduction to Kolmogorov complexity and Its applications. Springer, New York, third edition, 2008.

[6] Teturo Kamae and Dong Han Kim, A characterization of eventually periodicity (preprint, available at Kamae's homepage below).

[7] Yuval Peres and Benjamin Weiss, Private commnications (2004).

[8] Bo Tan, Private communications (2014).

Teturo Kamae
Advanced Mathematics Institute, Osaka City University
Osaka, 558-8585 Japan
(e-mail) kamae@apost.plala.or.jp
(home page) http://www14.plala.or.jp/kamae

Yu-Mei Xue
School of Mathematics and System Sciences & LMIB, BeiHang University
Beijing 100191, PR China
(e-mail) yxue@buaa.edu.cn