Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Departamento de Computación

Secuencias completamente equidistribuidas basadas en secuencias de Ford

Tesis de Licenciatura en Ciencias de la Computación

Emilio Guido Almansi

Directora: Verónica Becher

Buenos Aires, 2019

CONTENTS

1. PRELIMINARIES

1.1 Random Sequences

In engineering, computer science, and other branches of science we are often required to simulate random processes. Simulations usually involve the generation of non-random, deterministic sequences of numbers which approximate a sequence of independent, random samples from a given probability distribution. While a deterministic sequence can never be random in the sense of not following any patterns or being entirely unpredictable, nothing prevents one from identifying properties common to all random sequences and constructing pseudo-random sequences satisfying these properties.

Significant work has been done in this direction. In a paper by Franklin [2], the notion of equidistribution is discussed and presented as a first requirement for randomness. Franklin proves that the sequence of fractional parts $\{\alpha\}, \{\alpha^2\}, \{\alpha^3\}, ...$ is completely equidistributed for *almost all* $\alpha > 1$. However, no specific value of α is provided that verifies this property.

The first concrete construction yielding a completely equidistributed sequence is due to Knuth [3], and based on De Bruijn sequences of increasing order and alphabet size. In this work, we provide a similar albeit simpler construction and prove that the sequence it yields is also completely equidistributed.

Since a real computer with finite word-length and finite memory can only produce numbers of limited precision and sequences that will ultimately be periodic, we ignore this limitation by considering a computational model with infinite memory and with infinite word-length, where real numbers can be stored and computed with perfect precision.

1.2 A Note on Notation

Since notation on sequences can differ across the literature, we state the conventions that we will be using throughout this work.

When discussing a sequence $X = x_1, x_2, \ldots$, the expression X_i will denote the i-th element of the sequence, with the first element having an index of 1.

The expression $X_{1:n}$ will denote a prefix of length n from sequence X. Namely, $X_{1:n} = x_1, x_2, \ldots, x_n$.

Given two sequences X and Y, the expression $\langle X; Y \rangle$ will denote the sequence obtained

by concatenating Y after X, when this operation is well-defined. We also extend this notation to more than two input sequences; for example, if $A = 1, 2, 3$, $B = 3$, and $C = 4, 5$, then $\langle A; B; C\rangle = 1, 2, 3, 3, 4, 5$.

1.3 Complete Equidistribution

In order to define complete equidistribution, we will first need to define a notion of "probability" for deterministic sequences. Let $P = p_1, p_2, ...$ be an infinite sequence of predicates and σ a function such that $\sigma(p_i) = 1$ if p_i is *true*, and $\sigma(p_i) = 0$ otherwise. We define:

$$Pr(P) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sigma(p_i)$$

when this limit exists.

For example, given a sequence $x_1, x_2, \ldots$, we can define:

$$Pr(x_i < x_{i+1}) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sigma(x_i < x_{i+1})$$

if the limit exists. Note that $Pr(x_i < x_{i+1})$ denotes $Pr(P)$ where $P = (x_i < x_{i+1})_{i=1}^{\infty}$.

We say that an infinite sequence $X = x_1, x_2, \ldots$ of real numbers is *equidistributed in the unit interval* if the probability of finding x_i in any subinterval is proportional to the length of the subinterval. Formally, if for every interval $I = [u, v) \subseteq [0, 1)$ the following is true:

$$Pr(x_i \in I) = |I| = v - u$$

Analogously, we say that an infinite sequence $\bar{X} = \bar{x}_1, \bar{x}_2, \ldots$ of k-dimensional vectors of real numbers is *equidistributed in the unit cube* if for every set $I = [u_1, v_1) \times [u_2, v_2) \times \cdots \times [u_k, v_k) \subseteq [0, 1)^k$ the following is true:

$$Pr(\bar{x}_i \in I) = |I| = \prod_{d=1}^{k} v_d - u_d$$

For every positive integer k we define the *windows sequence of X of order k*, denoted $W_k(X)$, as the sequence of k-dimensional vectors containing every possible window (or

contiguous subsequence) of X from left to right:

$$\begin{aligned} W_k(X) &= (x_1, x_2, \ldots, x_k), (x_2, x_3, \ldots, x_{k+1}), \ldots \\ &= \big((x_i, x_{i+1}, \ldots, x_{i+k-1})\big)_{i=1}^{\infty} \end{aligned} \tag{1.1}$$

We say that X is *k-distributed* in the unit interval if its windows sequence of order k is equidistributed in the unit cube. Note that, as per the definition above, $W_k(X)$ includes windows with superposition. If one instead considers windows without superposition, then the derived notion of k-distribution is not equivalent. See Theorem 19 in [2] for an example of a sequence which is 2-distributed in the former sense, but not in the latter.

If X is k-distributed for every positive integer k, then we additionally say that X is *completely equidistributed*. In [2], Franklin proves that if X is completely equidistributed, then it also satisfies many other statistical properties common to all random sequences. For example, for every fixed k the lag k autocorrelation of X is zero, and the probability of k consecutive terms having any specific relative order is $1/k!$.

For convenience, we extend the notion of windows sequences to finite and cyclic sequences. In the case of a finite sequence $Y = y_1, y_2, \ldots, y_l$ of length l, we define $W_k(Y)$ similarly to how we did in 1.1:

$$\begin{aligned} W_k(Y) &= (y_1, y_2, \ldots, y_k), \ldots, (y_{l-k+1}, y_{l-k+2}, \ldots, y_l) \\ &= \big((y_i, y_{i+1}, \ldots, y_{i+k-1})\big)_{i=1}^{max(0,l-k+1)} \end{aligned}$$

Note that the windows sequence of Y of order k will be empty if $l < k$, having $max(0, l-k+1)$ terms in general.

In order to avoid ambiguity with the previous definition, for a cyclic sequence $Z = z_1, z_2, \ldots, z_l$ of size l we denote its windows sequence of order k as $W_k^c(Z)$ instead. In this case, the derived sequence has exactly l terms:

$$\begin{aligned} W_k^c(Z) &= (z_1, z_2, \ldots, z_k), \ldots, (z_l, z_1, \ldots, z_{k-1}) \\ &= \big((z_i, z_{i+1}, \ldots, z_{i+k-1})\big)_{i=1}^{l} \end{aligned}$$

where the indices are taken modulo l.

1.4 De Bruijn Sequences and Ford Sequences

For the constructions presented in chapter 2, we will use Ford sequences as a basic primitive, which are themselves a specific type of De Bruijn sequences. Having been originally defined over a binary alphabet (see [1]) and later generalized to larger alphabets, their

definition varies slightly accross the literature. Throughout this work, we will use these terms according to the following definitions.

Definition. A b-ary De Bruijn sequence of order k is any sequence of length b^k which, when viewed as a cycle, contains every possible b-ary sequence of length k exactly once as a contiguous subsequence.

Example. Listed next are two distinct binary De Bruijn sequences of order 3:

$$0, 0, 0, 1, 0, 1, 1, 1;$$
$$0, 0, 0, 1, 1, 1, 0, 1.$$

Note that all possible binary sequences of length 3 appear exactly once as a contiguous subsequence in each example. This includes those instances such as $1, 0, 0$ which wrap around the right-hand end of the sequences.

Example. Next, we list two distinct 4-ary De Bruijn sequences of order 2:

$$0, 0, 1, 0, 2, 0, 3, 1, 1, 2, 1, 3, 2, 2, 3, 3;$$
$$0, 0, 1, 1, 2, 1, 3, 1, 0, 2, 2, 3, 2, 0, 3, 3.$$

In general, there may exist multiple different b-ary De Bruijn sequences of order k. The lexicographically smallest one is often also called a Ford sequence. More precisely:

Definition. A b-ary Ford sequence of order k, denoted $F^{(b,k)}$, is the lexicographically smallest b-ary De Bruijn sequence of order k.

Example. The first sequences in each of the examples above are, respectively, the binary Ford sequence of order 3, and the 4-ary Ford sequence of order 2.

2. COMPLETELY EQUIDISTRIBUTED SEQUENCES BASED ON FORD SEQUENCES

2.1 Knuth's Sequence

In order to define Knuth's sequence, denoted as K, we must first construct the following building blocks based on Ford sequences. Given a natural number n, we define:

i) an A sequence of order n, denoted $A^{(n)}$, as the finite sequence of rational numbers obtained from dividing by 2^n each of the terms in a 2^n-ary Ford sequence of order n:

$$A^{(n)} = \frac{f_1}{2^n}, \frac{f_2}{2^n}, \ldots, \frac{f_{2^{n^2}}}{2^n} = \left(\frac{f_i}{2^n}\right)_{i=1}^{2^{n^2}}$$
$$\text{where } F^{(2^n,n)} = f_1, \ldots, f_{2^{n^2}}$$

and, ii) a B sequence of order n, denoted $B^{(n)}$, as $n2^{2n}$ consecutive copies of $A^{(n)}$:

$$B^{(n)} = \left\langle \underbrace{A^{(n)}; A^{(n)}; \ldots; A^{(n)}}_{n2^{2n} \text{ times}} \right\rangle$$

By construction, the size of $A^{(n)}$ is $|A^{(n)}| = |F^{(2^n,n)}| = 2^{n^2}$, and the size of $B^{(n)}$ is $|B^{(n)}| = n2^{2n}|A^{(n)}| = n2^{2n}2^{n^2}$. Note as well that, for any given n, all terms in $A^{(n)}$ and in $B^{(n)}$ are numbers in the set $\left\{0, \frac{1}{2^n}, \frac{2}{2^n}, \ldots, \frac{2^n-1}{2^n}\right\} \subset [0, 1)$.

For example, when $n = 2$:

$$F^{(4,2)} = 0, 0, 1, 0, 2, 0, 3, 1, 1, 2, 1, 3, 2, 2, 3, 3$$
$$A^{(2)} = \frac{0}{4}, \frac{0}{4}, \frac{1}{4}, \frac{0}{4}, \frac{2}{4}, \frac{0}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{2}{4}, \frac{1}{4}, \frac{3}{4}, \frac{2}{4}, \frac{2}{4}, \frac{3}{4}, \frac{3}{4}$$
$$B^{(2)} = \left\langle \underbrace{A^{(2)}; \ldots; A^{(2)}}_{32 \text{ times}} \right\rangle = \underbrace{\frac{0}{4}, \frac{0}{4}, \ldots, \frac{3}{4}, \frac{3}{4}}_{A^{(2)}}, \ldots, \underbrace{\frac{0}{4}, \frac{0}{4}, \ldots, \frac{3}{4}, \frac{3}{4}}_{A^{(2)}}$$

and $|A^{(2)}| = 16$, $|B^{(2)}| = 512$.

We now define Knuth's sequence, denoted as K, as the infinite sequence of real numbers resulting from the concatenation of all possible B sequences in increasing order:

$$K = \left\langle B^{(1)}; B^{(2)}; B^{(3)}; \dots \right\rangle$$

Theorem (Knuth 1965, [3], page 268)**.** *Sequence K is completely equidistributed.*

Two choices in the construction yielding K may seem arbitrary at first glance, but play an important role in the proof of the theorem stated above. Namely, that the number of repetitions of each A sequence within a B sequence is $n2^{2n}$, and the fact that the alphabet sizes of the composing Ford sequences grow exponentially as 2^n.

On account of the first choice, one can adapt Knuth's proof in a rather straightforward manner to show that a sufficient condition for the complete equidistribution of K is for the number of repetitions to grow asymptotically faster than 2^{2n}. A proof of this fact is omitted here since it follows from Knuth's work, together with the technique that will be used in the following section to obtain an analogous result.

In regard to the second choice, the use of successive powers of 2 as alphabet sizes allows one to reason more easily in terms of bits about the rational numbers comprised in sequence K. In particular, Knuth uses this device to derive properties of the distribution of the m most-significant bits of the terms in a 2^n-ary Ford sequence, where $m \leq n$.

As we will see in the next section, the second choice can be relaxed to using linearly increasing alphabet sizes while preserving the property of complete equidistribution, and at the same time allowing a much more reduced number of repetitions.

2.2 Linearly Increasing Alphabet Sizes

We now provide a variant of Knuth's sequence based on Ford sequences with linearly increasing alphabet sizes. Within the current section, consider $t : \mathbb{N} \mapsto \mathbb{N}$ to be an arbitrary but fixed function. Later, we will study which conditions t must satisfy in order for the generated sequence to be completely equidistributed. Similar to the previous section, we define for any given natural number n:

i) a C sequence of order n, denoted $C^{(n)}$, as the finite sequence of rational numbers obtained from dividing by n each of the terms in an n-ary Ford sequence of order n:

$$C^{(n)} = \frac{f_1}{n}, \frac{f_2}{n}, \dots, \frac{f_{n^n}}{n} = \left(\frac{f_i}{n}\right)_{i=1}^{n^n}$$
$$\text{where } F^{(n,n)} = f_1, \dots, f_{n^n}$$

and, ii) a D sequence of order n, denoted $D^{(n)}$, as $t(n)$ consecutive copies of $C^{(n)}$:

$$D^{(n)} = \left\langle \underbrace{C^{(n)}; C^{(n)}; \ldots; C^{(n)}}_{t(n) \text{ times}} \right\rangle$$

Note again that the size of $C^{(n)}$ is $|C^{(n)}| = |F^{(n,n)}| = n^n$, and the size of $D^{(n)}$ is $|D^{(n)}| = t(n)|C^{(n)}| = t(n)n^n$. In this case, for any given n all terms in $C^{(n)}$ and in $D^{(n)}$ are numbers in the set $\left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\right\} \subset [0, 1)$.

The key difference between the way C sequences are constructed when compared to A sequences from the previous section is that, as the order of the sequence grows, the alphabet size for the underlying Ford sequence grows linearly $(1, 2, 3, 4, \ldots)$ rather than exponentially $(2, 4, 8, 16, \ldots)$.

For example, when $n = 3$ and t is equal to the identity function:

$$F^{(3,3)} = 0,0,0,1,0,0,2,0,1,1,0,1,2,0,2,1,0,2,2,1,1,1,2,1,2,2,2$$

$$C^{(3)} = \frac{0}{3}, \frac{0}{3}, \frac{0}{3}, \frac{1}{3}, \frac{0}{3}, \frac{0}{3}, \frac{2}{3}, \frac{0}{3}, \frac{1}{3}, \frac{1}{3}, \frac{0}{3}, \frac{1}{3}, \frac{2}{3}, \frac{0}{3}, \frac{2}{3}, \frac{1}{3}, \frac{0}{3}, \frac{2}{3}, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{2}{3}, \frac{1}{3}, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}$$

$$D^{(3)} = \left\langle C^{(3)}; C^{(3)}; C^{(3)} \right\rangle = \underbrace{\frac{0}{3}, \frac{0}{3}, \ldots, \frac{2}{3}, \frac{2}{3}}_{C^{(3)}}, \underbrace{\frac{0}{3}, \frac{0}{3}, \ldots, \frac{2}{3}, \frac{2}{3}}_{C^{(3)}}, \underbrace{\frac{0}{3}, \frac{0}{3}, \ldots, \frac{2}{3}, \frac{2}{3}}_{C^{(3)}}$$

and $|C^{(3)}| = 27$, $|D^{(3)}| = 81$.

We now define sequence L as the infinite sequence of real numbers resulting from the concatenation of all possible D sequences in increasing order:

$$L = \left\langle D^{(1)}; D^{(2)}; D^{(3)}; \ldots \right\rangle$$

Theorem 1. *If t is non-decreasing and $\lim_{n\to\infty} n/t(n) = 0$, then sequence L is completely equidistributed.*

Example. If $t(n) = n^2$, then:

$$L = \left\langle C^{(1)}; \underbrace{C^{(2)}; \ldots; C^{(2)}}_{4 \text{ copies}}; \underbrace{C^{(3)}; \ldots; C^{(3)}}_{9 \text{ copies}}; \ldots \right\rangle$$

and L is completely equidistributed.

2.2.1 Proof of Theorem 1

In order to present our proof of Theorem 1, we must first establish some preliminary definitions.

Consider a prefix of L of length N, denoted $L_{1:N}$. It is always possible to find numbers $p, q, r \in \mathbb{N}$ such that:

$$L_{1:N} = \left\langle D^{(1)}; \ldots; D^{(r-1)}; \underbrace{C^{(r)}; \ldots; C^{(r)}}_{q \text{ times}}; C^{(r)}_{1:p} \right\rangle$$

where $0 \leq q < t(r)$ and $1 \leq p \leq r^r$. Here, r is the order of the rightmost, possibly incomplete D sequence present in $L_{1:N}$. The number q is the amount of complete C sequences of order r appearing before the rightmost, possibly incomplete C sequence, while p is the amount of terms present in said sequence. Note that the values of p, q and r are uniquely determined by the value of N.

By considering the length of the sequence on each side of the previous equation, we obtain a functional relationship between N, p, q, and r:

$$\begin{aligned} N &= \sum_{s=1}^{r-1} |D(s)| + q|C(r)| + p \\ &= \sum_{s=1}^{r-1} t(s)s^s + qr^r + p \end{aligned} \tag{2.1}$$

Let k be a positive integer and $I = [u_1, v_1) \times [u_2, v_2) \times \cdots \times [u_k, v_k)$ a set such that $I \subseteq [0, 1)^k$, where both k and I have arbitrary but fixed values. Let N range freely over the natural numbers, and the quantity ν_N denote the number of windows of L of size k starting at indices $i = 1 \ldots N$ that belong to the set I:

$$\nu_N = \sum_{i=1}^{N} \sigma\Big(\big(W_k(L)\big)_i \in I\Big)$$

We can now express the probability, in the sense defined in chapter 1, of any given window of L of size k belonging to the set I as:

$$Pr\Big(\big(W_k(L)\big)_i \in I\Big) = \lim_{N \to \infty} \frac{\nu_N}{N}$$

Consider sufficiently large values of N such that $k < r$. This is always possible since r is an unbounded, non-decreasing function of N. We can decompose $L_{1:N}$ into four

consecutive sections; namely, sequences $S^{(1)}$, $S^{(2)}$, $S^{(3)}$ and $S^{(4)}$:

$$\begin{aligned} L_{1:N} &= \left\langle S^{(1)}; S^{(2)}; S^{(3)}; S^{(4)} \right\rangle, \qquad \text{where} \\ S^{(1)} &= \left\langle D^{(1)}; D^{(2)}; \ldots; D^{(k-1)} \right\rangle \\ S^{(2)} &= \left\langle D^{(k)}; D^{(k+1)}; \ldots; D^{(r-1)} \right\rangle \\ S^{(3)} &= \left\langle \underbrace{C^{(r)}; \ldots; C^{(r)}}_{q \text{ times}} \right\rangle \\ S^{(4)} &= C^{(r)}_{1:p} \end{aligned} \tag{2.2}$$

Note that $S^{(1)}$ and $S^{(3)}$ could potentially be empty in the cases when $k = 1$ or $q = 0$, respectively.

We denote the cummulative sums of the sizes of the sequences defined above as $n_0 = 0$, and $n_j = n_{j-1} + |S^{(j)}|$ for $j = 1, 2, 3, 4$. Now, we can similarly decompose ν_N into five parts:

$$\begin{aligned} \nu_N &= \nu_N^{(1)} + \nu_N^{(2)} + \nu_N^{(3)} + \nu_N^{(4)} + \varepsilon_b, \qquad \text{where} \\ \nu_N^{(j)} &= \sum_{i=1+n_{j-1}}^{n_j-k+1} \sigma\Big(\big(W_k(L)\big)_i \in I\Big) \qquad j = 1, 2, 3, 4 \\ &= \sum_{i=1}^{|S^{(j)}|-k+1} \sigma\Big(\big(W_k(S^{(j)})\big)_i \in I\Big) \end{aligned} \tag{2.3}$$

for some $\varepsilon_b \leq 3(k-1)$.

For each $j = 1, 2, 3, 4$, the quantity $\nu_N^{(j)}$ accounts for windows contained entirely within the sequence $S^{(j)}$, and ε_b accounts for all windows crossing any of the three borders between the four sections. This is enough to account for all possible windows, since any given window will either be entirely contained in some section, or it will start at a given section and end at a subsequent one, thereby crossing a border.

Before obtaining more precise expressions for these quantities, we must first state the following three technical propositions.

Proposition 2. *If $n \in \mathbb{N}$ and $x, y \in \mathbb{R}$ such that $[x, y) \subseteq [0, n)$, then the number of integers from the set $\{0, 1, \ldots, n-1\}$ contained in $[x, y)$ is equal to $y - x + \varepsilon$ for some $\varepsilon \in (-1, 1)$.*

Proposition 3. *Given two sequences $a_1, a_2, \ldots, a_k$ and $b_1, b_2, \ldots, b_k$ of real numbers, the product of their element-by-element sums can be expanded in the following way:*

$$\prod_{d=1}^{k} a_d + b_d = \prod_{d=1}^{k} a_d + \sum_{j=1}^{2^k-1} \left[\prod_{d=1}^{k} \begin{cases} a_d & \left\lfloor \frac{j}{2^{d-1}} \right\rfloor \text{ is even} \\ b_d & \text{otherwise} \end{cases} \right]$$

Proposition 4. *Given $n \in \mathbb{N}$:*

$$\sum_{i=1}^{n} i^{i-1} \leq 2n^{n-1}$$

Proofs of propositions 2, 3, and 4 are given in Appendix A.

We will now obtain an expression for the number of windows of a C sequence which are contained in the set I. This will be useful for evaluating ν_N later on.

Lemma 5. *Given $n \in \mathbb{N}$ such that $k \leq n$, if we consider the sequence $C^{(n)}$ as a cyclic sequence, then for some $\varepsilon \in (-1, 1)$:*

$$\sum_{i=1}^{n^n} \sigma\Big(\big(W_k^c(C^{(n)})\big)_i \in I\Big) = n^n|I| + n^{n-1}(2^k - 1)\varepsilon$$

Proof. The expression on the left-hand side counts the number of windows of size k in $C^{(n)}$ that are contained in I. First, note that any given window is contained in the set I if and only if the following is true:

$$\big(W_k^c(C^{(n)})\big)_i \in I \iff \begin{array}{rcl} u_1 \leq & C_i^{(n)} & < v_1 \\ & \vdots & \\ u_k \leq & C_{i+k-1}^{(n)} & < v_k \end{array}$$

where $i = 1 \ldots n^n$ and indices are taken modulo n^n.

Since all terms in $C^{(n)}$ are numbers in the set $\left\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\right\}$, we multiply both sides of each inequality by n, allowing us to reason about integers belonging to a Ford sequence instead of rational numbers. We obtain the following:

$$\big(W_k^c(C^{(n)})\big)_i \in I \iff \begin{array}{rcl} nu_1 \leq & F_i^{(n,n)} & < nv_1 \\ & \vdots & \\ nu_k \leq & F_{i+k-1}^{(n,n)} & < nv_k \end{array}$$

As per Proposition 2, for each inequality above with $d = 1 \ldots k$ there are exactly $nv_d - nu_d + \varepsilon_d$ possible solutions in the set $\{0, 1, \ldots, n-1\}$ for some value $\varepsilon_d \in (-1, 1)$. This yields a total of $\prod_{d=1}^{k}[n(v_d - u_d) + \varepsilon_d]$ possible solutions to the system of inequalities. Each solution, when seen as an n-ary sequence of length k, appears exactly n^{n-k} times in $F^{(n,n)}$. This is true because there are n^{n-k} ways of extending an n-ary sequence of length k to one of length n and, by construction, each of these appears exactly once in $F^{(n,n)}$ when viewed as a cycle. Since i ranges exactly once over each possible window of $F^{(n,n)}$,

then:

$$\begin{aligned}\sum_{i=1}^{n^n} \sigma\Big(\left(W_k^c(C^{(n)})\right)_i \in I\Big) &= n^{n-k} \prod_{d=1}^{k} [n(v_d - u_d) + \varepsilon_d] \\ &= n^n \prod_{d=1}^{k} [(v_d - u_d) + \varepsilon_d/n]\end{aligned} \tag{2.4}$$

Using Proposition 3, we can expand this into the following:

$$\begin{aligned}n^n \prod_{d=1}^{k} [(v_d - u_d) + \varepsilon_d/n] = n^n \prod_{d=1}^{k} (v_d - u_d) \\ + n^n \sum_{j=1}^{2^k-1} \left[\prod_{d=1}^{k} \begin{cases} (v_d - u_d) & \left\lfloor \frac{j}{2^{d-1}} \right\rfloor \text{ is even} \\ \varepsilon_d/n & \text{otherwise} \end{cases} \right]\end{aligned} \tag{2.5}$$

If we define ε'_j for $j = 1 \dots 2^k - 1$ as:

$$\varepsilon'_j/n = \prod_{d=1}^{k} \begin{cases} (v_d - u_d) & \left\lfloor \frac{j}{2^{d-1}} \right\rfloor \text{ is even} \\ \varepsilon_d/n & \text{otherwise} \end{cases}$$

then, for each j, the value $\varepsilon'_j \in (-1, 1)$. This is true because the product on the right-hand side is composed of terms $(v_d - u_d) \in (-1, 1)$ and $\varepsilon_d/n \in (-1/n, 1/n)$ and, since $j > 0$, there is always at least one term of the second kind. Given that $|I| = \prod_{d=1}^{k}(v_d - u_d)$, we can further simplify equation 2.5 to get:

$$n^n \prod_{d=1}^{k} [(v_d - u_d) + \varepsilon_d/n] = n^n |I| + n^n \sum_{j=1}^{2^k-1} \varepsilon'_j/n \tag{2.6}$$

Finally, since $-(2^k - 1) < \sum_{j=1}^{2^k-1} \varepsilon'_j < (2^k - 1)$, there exists some $\varepsilon \in (-1, 1)$ such that:

$$\sum_{j=1}^{2^k-1} \varepsilon'_j = (2^k - 1)\varepsilon$$

and hence, by 2.4 and 2.6:

$$\sum_{i=1}^{n^n} \sigma\Big(\left(W_k^c(C^{(n)})\right)_i \in I\Big) = n^n |I| + n^{n-1}(2^k - 1)\varepsilon$$

□

Proof of Theorem 1. We will now obtain an expression for ν_N/N and compute its limit when $N \to \infty$. Recall the following definitions:

$$\begin{aligned}
\nu_N^{(2)} &= \sum_{i=1}^{|S^{(2)}|-k+1} \sigma\Big(\big(W_k(S^{(2)})\big)_i \in I\Big) \\
\nu_N^{(3)} &= \sum_{i=1}^{|S^{(3)}|-k+1} \sigma\Big(\big(W_k(S^{(3)})\big)_i \in I\Big) \\
S^{(2)} &= \left\langle D^{(k)}; D^{(k+1)}; \ldots; D^{(r-1)} \right\rangle \\
S^{(3)} &= \left\langle \underbrace{C^{(r)}; \ldots; C^{(r)}}_{q \text{ times}} \right\rangle
\end{aligned}$$

Note that the sequences $S^{(2)}$ and $S^{(3)}$ are entirely composed of complete C sequences of increasing orders which are larger than or equal to k. Moreover, with the exception of the last, rightmost instance in each of $S^{(2)}$ and $S^{(3)}$, every single C sequence is immediately succeeded by another C sequence of the same or the following order, including those which are part of a D sequence. Additionally, any window starting at the right-hand end of a C sequence will necessarily finish within the first $k-1$ elements of the following C sequence, all of which are guaranteed to be 0.

Therefore, the amount of windows of size k contained in I ranging over $S^{(2)}$ and $S^{(3)}$ is equal to the sum over each composing C sequence *viewed as a cycle*, with an error of at most $k-1$ due to the fact that we are counting only windows entirely contained within each sequence:

$$\begin{aligned}
\nu_N^{(2)} &= \sum_{s=k}^{r-1} \underbrace{\left[t(s) \sum_{i=1}^{s^s} \sigma\Big(\big(W_k^c(C^{(s)})\big)_i \in I\Big) \right]}_{C \text{ sequences contained in } D^{(s)}} + \varepsilon_{\nu_N^{(2)}} \\
\nu_N^{(3)} &= q \sum_{i=1}^{r^r} \sigma\Big(\big(W_k^c(C^{(r)})\big)_i \in I\Big) + \varepsilon_{\nu_N^{(3)}}
\end{aligned}$$

for some values $\varepsilon_{\nu_N^{(2)}} \leq k-1$, and $\varepsilon_{\nu_N^{(3)}} \leq k-1$.

From Lemma 5:

$$\begin{aligned}
\nu_N^{(2)} &= \sum_{s=k}^{r-1} \Big[t(s) \left(s^s |I| + s^{s-1}(2^k - 1)\varepsilon_s \right) \Big] + \varepsilon_{\nu_N^{(2)}} \\
\nu_N^{(3)} &= q \left(r^r |I| + r^{r-1}(2^k - 1)\varepsilon_r \right) + \varepsilon_{\nu_N^{(3)}}
\end{aligned}$$

for some values of $\varepsilon_i \in (-1, 1)$, $i = k \ldots r$.

Substituting back into ν_N from equation 2.3:

$$\begin{aligned}\nu_N = {} & \nu_N^{(1)} \\ & + \sum_{s=k}^{r-1} \left[t(s) \left(s^s |I| + s^{s-1}(2^k - 1)\varepsilon_s \right) \right] + \varepsilon_{\nu_N^{(2)}} \\ & + q \left(r^r |I| + r^{r-1}(2^k - 1)\varepsilon_r \right) + \varepsilon_{\nu_N^{(3)}} \\ & + \nu_N^{(4)} + \varepsilon_b\end{aligned}$$

and factoring out terms multiplied by $|I|$, we get:

$$\begin{aligned}\nu_N = {} & |I| \left[\sum_{s=k}^{r-1} t(s) s^s + q r^r \right] \\ & + \nu_N^{(1)} \\ & + \sum_{s=k}^{r-1} \left[t(s) s^{s-1}(2^k - 1)\varepsilon_s \right] + \varepsilon_{\nu_N^{(2)}} \\ & + q r^{r-1}(2^k - 1)\varepsilon_r + \varepsilon_{\nu_N^{(3)}} \\ & + \nu_N^{(4)} + \varepsilon_b\end{aligned}$$

We now rewrite the first term using the relationship between p, r, q, and N from equation 2.1:

$$\begin{aligned}\nu_N = {} & |I| \left[N - \sum_{s=1}^{k-1} t(s) s^s - p \right] \\ & + \nu_N^{(1)} \\ & + \sum_{s=k}^{r-1} \left[t(s) s^{s-1}(2^k - 1)\varepsilon_s \right] + \varepsilon_{\nu_N^{(2)}} \\ & + q r^{r-1}(2^k - 1)\varepsilon_r + \varepsilon_{\nu_N^{(3)}} \\ & + \nu_N^{(4)} + \varepsilon_b\end{aligned}$$

and after dividing both sides by N and rearranging terms we obtain:

$$\begin{aligned}\frac{\nu_N}{N} - |I| = \frac{p}{N}\left[\frac{\nu_N^{(4)}}{p} - |I|\right] \\ + \frac{2^k - 1}{N}\left[\sum_{s=k}^{r-1} t(s)s^{s-1}\varepsilon_s + qr^{r-1}\varepsilon_r\right] \\ + \frac{1}{N}\left[\nu_N^{(1)} - |I|\sum_{s=1}^{k-1} t(s)s^s + \varepsilon_{\nu_N^{(2)}} + \varepsilon_{\nu_N^{(3)}} + \varepsilon_b\right]\end{aligned}$$

Taking limits on both sides as $N \to \infty$, the third term on the right-hand side approaches 0 since the contents of the brackets are dependent on k and bounded as a function of N. In regard to the second term, using the fact that t is non-decreasing together with Proposition 4 we can see that:

$$\frac{\sum_{s=k}^{r-1} t(s)s^{s-1}\varepsilon_s}{N} \leq \frac{t(r-1)\sum_{s=1}^{r-1} s^{s-1}}{t(r-1)(r-1)^{(r-1)}} \leq \frac{2(r-1)^{(r-2)}}{(r-1)^{(r-1)}} = \frac{2}{r-1}, \text{ and}$$
$$\frac{qr^{r-1}\varepsilon_r}{N} \leq \frac{qr^{r-1}}{qr^r} = \frac{1}{r}.$$

Since r is an unbounded, non-decreasing function of N, this term approaches 0 as well.

Finally, consider the first term on the right-hand side and note that $\left[\frac{\nu_N^{(4)}}{p} - |I|\right] \in [-1, 1]$, since both $\frac{\nu_N^{(4)}}{p}, |I| \in [0, 1]$. Moreover, since $p \leq r^r$ and using the identity:

$$\frac{(x+1)^{(x+1)}}{x^x} = (x+1)\left(1 + \frac{1}{x}\right)^x$$

for $x = r - 1$, we can see that for large values of r:

$$\frac{p}{N} \leq \frac{r^r}{t(r-1)(r-1)^{(r-1)}} = \frac{r}{t(r-1)}\left(1 + \frac{1}{r-1}\right)^{r-1} \leq \frac{r}{t(r-1)}e$$

which by hypothesis also approaches 0 as $N \to \infty$. Hence,

$$\lim_{N\to\infty} \frac{\nu_N}{N} = |I|$$

and, since k and I were chosen arbitrarily, sequence L is completely equidistributed and the proof of Theorem 1 is complete.

□

APPENDIX

Appendix A

Proposition 2. If $n \in \mathbb{N}$ and $x, y \in \mathbb{R}$ such that $[x, y) \subseteq [0, n)$, then the number of integers from the set $\{0, 1, \ldots, n-1\}$ contained in $[x, y)$ is equal to $y - x + \varepsilon$ for some $\varepsilon \in (-1, 1)$.

Proof. Since $0 \leq y$, there's exactly $\lceil y \rceil = y + \varepsilon_y$ non-negative integers in the set $[0, y)$ for some $\varepsilon_y \in [0, 1)$. Similarly for x, there's exactly $\lceil x \rceil = x + \varepsilon_x$ non-negative integers in the set $[0, x)$ for some $\varepsilon_x \in [0, 1)$. The difference between these two quantities is equal to the number of non-negative integers contained in the set $[x, y)$, which is $y - x + (\varepsilon_y - \varepsilon_x)$. Observing that $(\varepsilon_y - \varepsilon_x) \in (-1, 1)$, and that all non-negative integers between x and y belong to the set $\{0, 1, \ldots, n-1\}$, the proof is complete. □

Proposition 3. Given two sequences $a_1, a_2, \ldots, a_k$ and $b_1, b_2, \ldots, b_k$ of real numbers, the product of their element-by-element sums can be expanded in the following way:

$$\prod_{d=1}^{k} a_d + b_d = \prod_{d=1}^{k} a_d + \sum_{j=1}^{2^k-1} \left[\prod_{d=1}^{k} \begin{cases} a_d & \left\lfloor \frac{j}{2^{d-1}} \right\rfloor \text{ is even} \\ b_d & \text{otherwise} \end{cases} \right]$$

Proof. // TODO □

Proposition 4. Given $n \in \mathbb{N}$:

$$\sum_{i=1}^{n} i^{i-1} \leq 2n^{n-1}$$

Proof. By induction on n. The property holds for $n = 1$ and $n = 2$:

$$\sum_{i=1}^{1} i^{i-1} \leq 2, \; \sum_{i=1}^{2} i^{i-1} \leq 4$$

and the inductive step holds for $n \geq 2$:

$$\sum_{i=1}^{n+1} i^{i-1} = \underbrace{\sum_{i=1}^{n} i^{i-1}}_{\leq 2n^{n-1} \text{ by I. H.}} + (n+1)^n \leq nn^{n-1} + (n+1)^n \leq 2(n+1)^n.$$

Therefore, the property holds for all $n \in \mathbb{N}$. □

BIBLIOGRAPHY

[1] N. G. DE BRUIJN. A combinatorial problem. *Proc. Koninklijke Nederlandse Academie van Wetenschappen*, 49:758–764, 1946.

[2] Joel N. Franklin. Deterministic simulation of random processes. *Mathematics of Computation*, 17(81):28–59, 1963.

[3] Donald E. Knuth. Construction of a random sequence. *BIT Numerical Mathematics*, 5(4):246–250, Dec 1965.