Journal of Statistical Physics, Vol. 63, Nos. 5/6, 1991

The Hierarchy of Correlations in Random Binary Sequences

Aaldert Compagner[1]

Received December 18, 1990

The meaning of randomness is studied for the simple case of binary sequences. Ensemble theory is used, together with correlation coefficients of any order. Conservation laws for the total amount of correlation are obtained. They imply that true randomness is an ensemble property and can never be achieved in a single sequence. The relation with entropy is discussed for different ensembles. Well-tempered pseudorandom sequences turn out to be suitable sources of random numbers, and practical recipes to generate them for use in large-scale Monte Carlo simulations are found.

KEY WORDS: Randomness; correlation coefficients; entropy; random-number generation; Monte Carlo calculations.

1. INTRODUCTION

Randomness is a notion that often occurs in physics. Random systems appear as limiting cases at infinite temperature, or also as quenched states at low temperatures. In ergodic theory and in the description of irreversible phenomena randomness plays a central role, be it in the Ehrenfests' urn model or in the dynamic theory of chaotic systems. Perhaps the simplest and most specific example is found in the field of Monte Carlo simulations, where random numbers are used to implement transition probabilities. Much effort has been spent to find the deterministic methods for their generation that are essential for efficiency and general control; see the review given by Knuth[1] and many recent papers.[2-6]

In mathematics, randomness is an important notion in the foundation of probability theory, and in algorithmic theory it is related to basic notions like complexity and incomputability. Many mathematicians and

[1] Laboratory of Applied Physics, Delft, The Netherlands.

0022-4715/91/0600-0883$06.50/0

even philosophers, such as von Mises, Kolmogorov, Popper, Martin-Löf, and Chaitin, have discussed the subject. A survey of the mathematical debate was given by van Lambalgen.[7]

In spite of all this attention, a general and operational notion of randomness has not emerged, not even for the simple case of a random binary sequence. This is unsatisfactory, also for practical reasons. At present, several methods to generate random numbers exist that are suitable for many purposes, but for future large-scale Monte Carlo calculations improved methods have to be found, capable of generating random bits of high quality at GHz rates. Their development is hampered by the confusion still surrounding randomness, which is due to the conflict between stochastic and deterministic descriptions and, in mathematics, between methods based on existence proofs and constructive methods.

A straightforward discussion of randomness for the simple case of binary sequences may diminish the confusion by providing a concrete example. Such a discussion is based here on ensemble theory and on the hierarchy of correlation coefficients that was introduced in an earlier paper.[8] Reliable recipes for random-number generation at GHz rates will serve as a practical objective. The present paper contains the technical part of the discussion; a general report, concentrating on the conceptual aspects and giving explicit examples, will be published elsewhere.[9]

2. THE HIERARCHY OF CORRELATION COEFFICIENTS

Consider a binary sequence $\{a_i\}$ of N bits $a_i = 0, 1$ with $i = 1,..., N$. It is equivalent to a parity sequence $\{b_i\}$, where the parities are given by $b_i = (-1)^{a_i}$. The sequences are identified by the decimal representation

$$j = \sum_{i=1}^{N} a_i \cdot 2^{N-i} \tag{1}$$

An ensemble of sequences of N bits is defined by assigning a statistical weight p_j to sequence j for $j = 0,..., 2^N - 1$. The weights obey

$$0 \leqslant p_j \leqslant 1, \qquad \sum_{j=0}^{2^N-1} p_j = 1 \tag{2}$$

An increasing series of indices arbitrarily selected from the range 1 to N is indicated by the set

$$I(q, s) = \{i_1,..., i_q\} \tag{3}$$

where the number of elements q is the order of the set and $s = i_q - i_1 + 1 \geqslant q$ is its size. Equation (3) is not meant to imply that q and s together define

the set of fixed indices completely. The total number of sets is 2^N, including the empty set $\varnothing \equiv I(0, 0)$, and there are $N!/[q!\,(N-q)!]$ sets of order q. For each set a product of parities $b_i^{(j)}$ of sequence j is given by

$$P(I, j) = \prod_{i \in I} b_i^{(j)} \tag{4}$$

This product appears as the general term in the following expansion over all sets I, including the empty set $\varnothing$ with $P(\varnothing, j) = 1$:

$$\prod_{i=1}^{N} (1 + b_i^{(j)}) = \sum_{I} P(I, j) = 2^N \delta_{j,0} \tag{5}$$

where Kronecker's symbol is used to express that this quantity always vanishes unless $b_i^{(j)} = 1$ for all i, which happens only for $j = 0$.

A correlation coefficient C_I of order q and size s is now defined as the expectation value of $P(I, j)$ over the ensemble of sequences,

$$C_{I(q,s)} \equiv \sum_{j=0}^{2^N - 1} P(I, j)\, p_j \tag{6}$$

For the empty set $I = \varnothing$, not measuring a true correlation, $C_\varnothing = 1$ holds always. All values of C_I lie on the segment $(-1, 1)$. When they are summed over all sets I, the expectation value of the quantity given in Eq. (5) appears. The mean value of the correlation coefficients therefore obeys

$$\langle C_I \rangle \equiv 2^{-N} \sum_{I} C_I = p_0 \tag{7}$$

where p_0 is the weight of the sequence for which $a_i = 0$ with $b_i = 1$ for all i. Equation (7) is a weak conservation law for the total amount of correlation; weak, because the different correlation coefficients have different signs and cancel one another rather effectively.

Equation (6) gives the correlation coefficients in terms of the statistical weights, but also a reverse relation holds. One may write

$$p_j = 2^{-N} \sum_{k=0}^{2^N - 1} \left[\prod_{i=1}^{N} (1 + b_i^{(j)} b_i^{(k)}) \right] p_k$$

since the product is equal to $2^N \delta_{j,k}$. Expanding this product and using Eqs. (4) and (6), one finds

$$p_j = 2^{-N} \sum_{k=0}^{2^N - 1} \sum_{I} \left[\prod_{i \in I} b_i^{(j)} b_i^{(k)} \right] p_k = 2^{-N} \sum_{I} P(I, j) C_I \tag{8}$$

Equations (6) and (8) describe a transformation between the 2^N weights and the 2^N correlation coefficients, without loss of information. A symmetric notation would result when the sets $I(q, s)$ were ordered by means of their binary code, with 1's only at the positions contained in I. This would reveal $P(I, j)$ as the $2^N \times 2^N$ Hadamard–Sylvester matrix it really is. Here, it suffices to note that the transformation is norm-conserving:

$$\begin{aligned}\langle C_I^2 \rangle &\equiv 2^{-N} \sum_I C_I^2 \\ &= 2^{-N} \sum_I \left[\sum_j P(I, j)\, p_j \right]^2 \\ &= 2^{-N} \sum_j \sum_k \left[\prod_{i=1}^{N} (1 + b_i^{(j)} b_i^{(k)}) \right] p_j p_k \\ &= \sum_{j=0}^{2^N - 1} p_j^2 \end{aligned} \tag{9}$$

This may be called a strong conservation law, since in the mean square correlation coefficient no cancellations occur.

3. SPECIAL ENSEMBLES

So far, the ensemble was completely general, but now three special cases will be considered. The gambling ensemble is defined by

$$p_j = 2^{-N} \qquad \text{for all } j \tag{10}$$

implying that the bits are independent. The correlation coefficients obey

$$C_I = 2^{-N} \sum_j P(I, j) = 0 \qquad \text{for} \quad I \neq \varnothing \tag{11}$$

because the sum over the sequences of parity products can be written as a product over I of parity sums per element, which are zero. In the gambling ensemble, $C_\varnothing$ is the only term that survives in Eqs. (7) and (9). It follows from Eq. (8) that if all true correlation coefficients are zero, the weights are those of the gambling ensemble. If randomness is the same as uncorrelatedness, it is a property of the gambling ensemble, not of a single sequence. The gambling ensemble is future-directed: it corresponds with the expected outcomes when a fair coin is going to be tossed N times. It is identical with the canonical ensemble for a chain of N Ising spins at infinite temperature.

At the other extreme is the singular ensemble, defined by

$$p_j = \delta_{j,k} \qquad \text{for all } j \tag{12}$$

where k is the decimal representation of the only sequence in the ensemble. All correlation coefficients are either -1 or 1, depending on the number of negative parities involved in $P(I, k)$; in a fixed sequence, everything is correlated and nothing is random. If $k \neq 0$, the negative and positive coefficients in Eq. (7) cancel one another completely. The mean square value of Eq. (9) is now equal to 1. The singular ensemble refers to a known sequence of past events, the actual outcomes of N tosses of a coin, fair or not; it corresponds to a quenched Ising chain at zero temperature.

However, the idea that a single sequence could not be called random in any sense is too severe; after a coin has been tossed many times, one should be able to decide whether it is reasonably fair or most likely unfair by scanning the sequence of outcomes. The wording reflects that a certain subjectivity is unavoidable. The decision can be based on the average behavior of C_I along the sequence, especially for sets $I(q, s)$ of small order and size; the smaller C_I^2 is for these sets, the fairer the coin appears to be.

A formal base is provided by the scanning ensemble, consisting of a sequence k and its $N-1$ cyclically translated version $k', k'',...$ that are found by iteration of

$$k' = 2k \bmod 2^N + (2k - 2k \bmod 2^N)/2^N$$

The nonzero weights for the scanning ensemble are

$$p_j = \frac{1}{N} \qquad \text{for} \quad j = k, k', k'',... \tag{13}$$

The translated versions are assumed to be different (if not, the period of the sequence is shorter than N and should be used instead). The all-zero sequence thus being absent from the scanning ensemble for any interesting value of N, the mean correlation coefficient of Eq. (7) vanishes for that ensemble. The mean square value of Eq. (9) is $1/N$, which is rather small, but which, as an average over 2^N positive quantities, is far from excluding the presence of sets with correlation coefficients close to ± 1.

The more of those coefficients happen to be equal to 0, the closer to ± 1 the remaining ones must be. This is the predicament that all methods for generating random sequences must face up to. Any neglect of the difference between the true randomness expected when a fair coin is going to be tossed N times and the approximate form of randomness found afterward when scanning the actual outcomes is an example of the more general confusion around irreversibility.

Averaging $P(I, k)$ for a fixed set I over the scanning ensemble is the same as averaging the parity product over a single sequence while keeping only the relative positions in I fixed (apart from boundary effects). The correlation coefficients for the scanning ensemble, the scanning coefficients for short, are indeed identical with the usual multispin correlation coefficients for a periodic chain of frozen Ising spins. For instance, the scanning coefficients C_I of second order, for $I = I(2, s)$, form together the pair-correlation function for two spins or partities at a varying relative distance s.

Still other ensembles could be envisaged, introducing a certain bias between positive and negative parities or between like and unlike more or less neighboring parities (which is what the Ising model is all about). Another generalization, to sequences of numbers modulo m, can be obtained by using in the correlation coefficients the complex mth roots of unity instead of the parities. Neither generalization is needed here.

To summarize: if randomness implies the vanishing of all true correlation coefficients, it is a property of the gambling ensemble. A single sequence dealt with by means of the singular ensemble can never be called random. Some or even many scanning coefficients can be small, but only at the detriment of others. Any definition of randomness that differs from uncorrelatedness has to face the question what deviations from zero are tolerated for which correlation coefficients.

4. ENTROPY, DEGREES OF FREEDOM, AND PSEUDORANDOMNESS

The measures $\langle C_I \rangle$ and $\langle C_I^2 \rangle$ for the total amount of correlation are adversely related to the entropy S, which is the usual ensemble measure for randomness defined by

$$S \equiv -\sum_{j=0}^{2^N-1} p_j \ln p_j \tag{14}$$

The values of S, $\langle C_I \rangle$, and $\langle C_I^2 \rangle$ for the different ensembles are given in Table I, together with the number $n = S/\ln 2$ of binary degrees of freedom, which for simplicity will be assumed to be integer. The maximum value $S = N \ln 2$ is indeed obtained for the gambling ensemble, which contains no information on any sequence. In the singular ensemble the entropy is 0, corresponding to complete information about the sequence. The identification of randomness with uncorrelatedness, arbitrariness, disorder, and complete lack of information appears to be fully justified and is indeed straightforward.

Table I. Ensemble Properties[a]

Ensemble	General	Gambling	Singular	Scanning
Weight	p_I	2^{-N}	$\delta_{j,k}$	$1/N$
Number of sequences	2^N	2^N	1	N
$\langle C_I \rangle$	p_0	2^{-N}	$\delta_{0,k}$	0
$\langle C_I^2 \rangle$	$\sum p_i^2$	2^{-N}	1	$1/N$
Entropy S	$-\sum p_j \ln p_j$	$N \ln 2$	0	$\ln N$
n	$S/\ln 2$	N	0	$\ln N/\ln 2$

[a] The values 2^{-N} in the gambling ensemble for the mean and the mean square correlation coefficient are due to the empty set $I = \varnothing$ only. Note the large difference in entropy and degrees of freedom between the gambling and the scanning ensemble.

In the scanning ensemble, with entropy $S = \ln N$, only the absolute positions of the bits of the sequence are unknown, their relative positions being the same as in the sequence k on which the ensemble is based. In this respect there is no distinction between a rather haphazard sequence of period N and one in which a single bit differs from all the others. However, the effective use of the available $n = \ln N/\ln 2$ degrees of freedom differs. The most economical use of these few degrees of freedom results when all 2^n strings of n bits that are possible occur somewhere in the sequence, overlaps being allowed. A sequence for which this holds will be called pseudorandom.

A pseudorandom sequence is ergodic in the sense that during a period N precisely all strings of n bits are visited once. Being periodic, the sequence may even be called deterministic, though this does not imply that it obeys a simple law or can be constructed by a Turing machine program of less than N bits. Conversely, the absence of a law does not guarantee pseudorandomness and leaves the question open as to which scanning coefficients are nonzero in order to obey $\langle C_I^2 \rangle = N^{-1}$. For pseudorandom sequences a partial answer can be given: all scanning coefficients for sets $I(q, s)$ with $q \leqslant s \leqslant n$ are zero. The little amount of freedom that the scanning ensemble allows is used to accommodate the gambling ensemble for sequences of $n = \log_2 N$ bits within a pseudorandom sequence of N bits. For a more complete answer, sets of size larger than n must be considered.

5. WELL-TEMPERED MAXIMUM-LENGTH SEQUENCES

In a pseudorandom sequence of N bits, all sets of size s smaller than n are uncorrelated, but at the same time there must be a first correlated set, $I(q, n+1) = \{i, j, k, \ldots, i+n\}$, of size $s = n+1$ with a nonvanishing scanning

coefficient that depends only on the relative locations $j-i$, $k-i,\ldots, n$. Otherwise, there would be more degrees of freedom than n and the period would be longer than $N=2^n$. The size of this set is thus as large as can be, but its order may be rather small. Indeed, it is not difficult to find pseudorandom sequences that show a lot of structure due to low-order correlations of size $n+1$. To improve upon pseudorandomness one has to require that $I(q, n+1)$ is not just of large size (implying that n is sufficiently large), but also of rather high order. In addition, all other correlated sets of size not much larger than n must be of a similar high order. The requirements are met by well-tempered pseudorandom sequences, invented for the occasion. For these sequences, the nonzero scanning coefficients belong either to sets $I(q, s)$ with $q \gg 1$ when $s-n$ is small, or to sets with $s \gg n$ when q is small, or to irrelevant sets for which both q and $s-n$ are large. If such sequences exist, they exhaust the main possibilities to imitate randomness within a single sequence.

However, sheer existence is not enough; a practical though perhaps not unique method for their construction is also needed. Consider the special case of a pseudorandom sequence for which the first-correlated set $I(q, n+1)$ happens to have a scanning coefficient that is not just different from zero, but even equal to 1. This looks more awkward than it is, since the total amount of correlation is conserved. The completely correlated set acts as a simple law which upon iteration produces the whole sequence. When a string of n bits is given as a seed located at positions i to $i+n-1$, the modulo-2 sum of the bit at position i and the following $q-2$ bits of $I(q, n+1)$ is equal to the next bit of the sequence (located at $i+n$). The seed determines the whole sequence.

This is just a description of a shift-register sequence produced by a $(q-1)$-bit feedback rule; see Golomb.[10] For suitable choices of the feedback positions, the first $q-1$ positions of the set $I(q, n+1)$, a maximum-length sequence of $N=2^n-1$ bits is produced, which is a pseudorandom sequence from which a single 0 is lacking (a string of n zeros cannot act as seed, but all other strings occur). The difference between a maximum-length and a pseudorandom sequence is negligible. As a simple example, the 2-bit feedback rule that corresponds to the first-correlated set $I(3, 5)=\{i,\ i+1,\ i+4\}$ produces the maximum-length sequence $\{111100010011010\}$ of period $N=15$. Maximum-length sequences produced by 2-bit rules, with n up to ≈ 250, have often been used for random-number generation, with varying succes; indeed, these sequences are not well-tempered, because $q=3$ is far too small. Yet, maximum-length sequences are attractive random-number generators because of their simplicity, which also facilitates a study of the correlation properties.

What would be acceptable values for q and n? When the pseudo-

random sequence is used as a source of random numbers of 32 bits each, only $n/32$ of these numbers will be completely independent. To avoid correlations of order 10 or lower as long as possible, the rule of thumb $n^{1/2} < q < \frac{1}{2}n$ may be used. To be definite, the rather arbitrary conditions $n \gtrsim 3000$ and $q \gtrsim 100$ are suggested, which are too severe for most purposes, though not for large-scale Monte Carlo simulations where random bits at GHz rates are needed. Can one find maximum-length rules $I(q, n+1)$ obeying these conditions? Yes: when the elements of m maximum-length sequences generated by widely different 2-bit rules are added modulo 2, an almost pseudorandom sequence is obtained, with a period that is the product of the periods of the constituting sequences (if these are relative primes). The order of the first-correlated set of that sequence is $q = 3^m$, apart from chance cancellations. Suitable 2-bit rules for this recipe can be taken from a list provided by Zierler[11] for maximum-length sequences with Mersenne prime as periods, the largest one having the truly maximum length of $2^{9689} - 1$ bits. Explicit and, in principle, efficient recipes with $n \gtrsim 10^4$ and $q \gtrsim 100$ or even much larger result.[9]

6. THE BRANCHING PROCESS OF CORRELATED SETS

The question still remains whether the other correlated sets that are generated by a first-correlated set $I(q, n+1)$ with large q are also of high order, at least initially. A simplifying feature is that all sets taken from a maximum-length sequence of N bits are either completely correlated or uncorrelated, with a scanning coefficient 1 or $-1/N$, respectively (the $-1/N$ instead of 0 is due to the missing bit). The sets of order q but of any size consist of $A(q, N)$ uncorrelated sets and $B(q, N)$ correlated ones. It can be shown that

$$B(q, N) = \frac{1}{N+1}\binom{N}{q} + \frac{N}{N+1} \times \begin{cases} (-1)^{q/2}\binom{(N-1)/2}{q/2}, & q \text{ even} \\ (-1)^{(q+1)/2}\binom{(N-1)/2}{(q-1)/2}, & q \text{ odd} \end{cases} \tag{15}$$

holds exactly.[8] The total number of sets is

$$A(q, N) + B(q, N) = \binom{N}{q} \tag{16}$$

For large N, the term in Eq. (15) depending on q being even or odd is negligible. The asymptotic expressions are

$$A(q, N) \approx \frac{N}{N+1}\binom{N}{q}, \qquad B(q, N) \approx \frac{1}{N+1}\binom{N}{q} \tag{17}$$

in agreement with the conservation laws of Eqs. (7) and (9). The correlated sets contribute N times more to the mean square scanning coefficient than the uncorrelated sets, although they are outnumbered by a factor N.

To find expressions depending on s, consider all sets $I(q, s) = \{1,..., s\}$ that start at position 1, consisting of $F(q, s)$ uncorrelated sets and $G(q, s)$ correlated ones, all of order q and size s. Their sum $H(q, s)$ is equal to the number of possibilities to chose the $q-2$ elements of the set between the positions 1 and s:

$$H(q, s) = F(q, s) + G(q, s) = \binom{s-2}{q-2} \qquad \text{for} \quad s \geqslant q \geqslant 2 \tag{18}$$

with $H(0, 0) = H(1, 1) = 1$ as separate values. For $q > s$ one has $G(q, s) = 0$, and because of pseudorandomness one has $G(q, s) = 0$ for $s \leqslant n$. A general property of maximum-length sequences is that the pair correlation and the

Table II. The Number $G(q, s)$ of Correlated Sets $I(q, s)$ for the Maximum-Length Sequence Generated by the First-Correlated Set $I(3, 5) = \{i, i+4\}$, Indicated by the Box[a]

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
0	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
1	.	.	.	.	.	.	.	.	.	.	.	$q \rightarrow$	.	.	.	.	.
2	.	.	*	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1
6	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1
7	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	2
8	.	.	.	.	2	2	.	.	.	.	.	.	.	.	.	.	4
9	.	.	.	1	2	1	1	2	1	.	.	.	.	.	.	.	8
10	.	.	.	1	1	3	6	3	1	1	.	.	.	.	.	.	16
11	.	s	.	.	1	3	3	7	11	5	1	1	.	.	.	.	32
12	.	↓	.	.	.	4	9	10	15	16	7	2	1	.	.	.	64
13	.	.	.	.	1	4	9	19	31	31	19	9	4	1	.	.	128
14	.	.	.	.	1	5	12	28	54	62	44	28	17	5	.	.	256
15	.	.	.	.	1	6	16	40	87	116	96	72	55	22	.	.	512
	.	.	.	.	1	5	18	45	80	107	107	80	45	18	5	1	.

[a] To the right the sums over q are given. The asterisk is where the triangle of Pascal of Eq. (18) for the total number $H(q, s)$ of sets $I(q, s)$ starts. The lowest line is the approximation to $G(q, 15)$ of Eq. (23).

average parity vanish: $G(q, s) = 0$ for $q \leqslant 2$, except $G(0, 0) = 1$ for the empty set. As an example, Table II gives all nonzero elements of $G(q, s)$ found by numerical examination for the trivial case of the maximum-length sequence with $N = 15$ that was given explicitly above.

At $s = n + 1$ the first real contribution appears, due to the first-correlated set $I(q, n+1) = \{1, j, ..., n+1\}$. In the example of Table II, the corresponding value $G(q, n+1)$ is indicated by the box. At $s = n + 2$ a new correlated set

$$I(q', n+2) = I(q, n+1) \oplus TI(q, n+1) = (1 \oplus T)\, I(q, n+1) \tag{19}$$

appears, where T indicates an elementary translation of $I(q, n+1)$ to the set $\{2, j+1, ..., n+2\}$, and where $\oplus$ is an EXOR operation: elements common to $I(q, n+1)$ and $TI(q, n+1)$ cancel one another in the parity product and do not appear in $I(q', n+2)$. The new correlated set gives $G(q', n+2) = 1$. At $s = n + 3$, two new correlated sets appear,

$$\begin{aligned} I(q'', n+3) &= (1 \oplus T^2)\, I(q, n+1) \\ I(q''', n+3) &= (1 \oplus T \oplus T^2)\, I(q, n+1) \end{aligned} \tag{20}$$

At the next step, the contributing operations are $1 \oplus T^3$, $1 \oplus T \oplus T^3$, $1 \oplus T^2 \oplus T^3$, and $1 \oplus T \oplus T^2 \oplus T^3$, and so on. At step i, new correlated sets of size $n + i + 1$ are found by exor-ing the set $T^i I(q, n+1) = \{i+1, i+j, ..., i+n+1\}$ with each of the new correlated sets of the earlier steps. In this branching process, all correlated sets starting at position 1 will be generated. At every step after the first one the number of new correlated sets doubles:

$$\sum_{q=3}^{s} G(q, s) = \begin{cases} 1 & \text{for} \quad s = n+1 \\ 2^{s-n-2} = \dfrac{2^{s-2}}{N+1} & \text{for} \quad s \geqslant n+2 \end{cases} \tag{21}$$

The new correlated sets counted by $G(q, s)$ are called so, because they are of greater size than the ones obtained at earlier steps. A translated set, not starting at position 1, does not contribute to $G(q, s)$, but it does contribute to $B(q, N)$. Since a set of size s can be translated $N - s - 1$ times (to avoid boundary effects, all sets are taken from a single period), the contribution of all correlated sets of size s is $(N - s - 1)\, G(q, s)$. That is, the relation

$$\sum_{s=n+1}^{N} (N - s + 1)\, G(q, s) = B(q, N) \tag{22}$$

must hold. A similar relation connects $F(q, s)$ and $A(q, N)$.

Equations (16)–(18) suggest the following asymptotic expressions:

$$F(q, s) \approx \frac{N}{N+1}\binom{s-2}{q-2}, \qquad G(q, s) \approx \frac{1}{N+1}\binom{s-2}{q-2} \tag{23}$$

for $N > s \gtrsim q \gg 2$. It is indeed quite natural to expect an asymptotic proportionality between these quantities. Equation (23) agrees with the sum rules over q and over s in Eqs. (21) and (22). Even the agreement between the two bottom lines of Table II is already rather satisfactory; a similar agreement was found for a maximum-length sequence with $n = 5$ degrees of freedom, of period $N = 31$, where the numerical determination of $G(q, s)$ starts to be time-consuming. For larger values of n, further evidence for this asymptotic behavior has been obtained[8] for restricted values of q and $s - n$. The stochastic region of the branching process can be understood in terms of the probability with which cancellations occur in the EXOR operations of the process, and consists of normal distributions $G(q, s)$ with a width that behaves as $s^{-1/2}$.

The deterministic region bordering the stochastic one for small values of q or of $s - n$ is both more tricky and more important; there, the correlations are found that are most detrimental for a good imitation of randomness. For instance, it can be shown[8] that $G(q, s)$ has an adverse efect on the qth moment of the distribution for the number of bits equal to 1 in subsequences of length s. To study the detailed behavior of $G(q, s)$ in the deterministic region, however, nonpolynomial time algorithms are required because of Eq. (18). For $s > n \gtrsim 1000$, say, an awkward situation seems to result, but when n is large a strategic choice for the position of the first-correlated set, which dominates the whole process, is possible. In the case of Table II there is not much choice for the position of the box, but when n is large it can be moved to much higher q values by adopting a suitable first-correlated set of high order. From that initial condition, the branching process starts. If the structure of the first-correlated set is sufficiently irregular, chance cancellations that give rise to correlated sets of low order are very unlikely (by far the most correlated sets occur around $q \approx \frac{1}{2}s$) and will appear only when their size is very large compared with n. The high order of the remaining local structures implies also that the behavior of subsequences, essential when periods of the order of 2^{10000} are involved, will be uniform. A large number of irregular feedback positions means a strong tendency back to normality, or an efficient return to the equilibrium of the stochastic region.

7. DISCUSSION AND CONCLUSIONS

The condition of irregularity in the production rule is needed to prevent collective cancellations and is obeyed when the 2-bit rules that enter the recipe described above are chosen from Zierler's list,[11] but many other 2-bit rules will also lend themselves for suitable combinations. A peculiar parallel can be drawn between this condition and the notion of complexity of a sequence, defined (by Kolmogorov, Martin-Löf, and Chaitin) as the length of the shortest Turing-machine program that produces the sequence. In that approach, a sequence is called random when its complexity is not smaller than its own length. The condition that the production rule involves many feedback bits at irregular positions is in fact a requirement that the complexity of the sequence is not too small. The similarity is obvious, but so are the differences.

What about other methods to generate random numbers, like the linear-congruence method or the lagged-Fibonacci method? Much experience with these methods has been obtained and several reliable recipes do exist that are suitable for a multitude of purposes. They have been subjected to severe tests, but a discussion in terms of a hierarchy of correlation coefficients seems to be difficult because the simplifying property of complete correlation characteristic for maximum-length sequences does not exist. For other pseudorandom sequences, instead of the quantity $H(q, s)$ of Eq. (18) from which the number $G(q, s)$ of completely correlated sets can be separated, one should consider the quantity

$$H'(q, s) = \sum_{I(q,s)} C^2_{I(q,s)} \tag{24}$$

where the summation is over all sets of order q and size s. This distribution should not have unacceptable peaks in the deterministic region, but an inspection of $H'(q, s)$ is even more difficult than one of $G(q, s)$, and an escape by means of a suitable choice for the first-correlated set is impossible. The usual methods to test random-number generators are difficult to use when high bit rates and very long sequences are desired. Moreover, they depend only indirectly on the hierarchy of correlation coefficients, and it seems to me that they do not sufficiently take into account that an increase of randomness in one respect leads to a decrease in another. The improved bit-scrambling achieved in recent developments such as lagged-Fibonacci sequences with multiple lags and very long periods[3] will, however, probably lead to properties similar to those of a well-tempered pseudorandom sequence.

To conclude: the meaning of randomness can be clarified by the use of ensembles, especially the scanning ensemble, in combination with the hierarchy of correlation coefficients. Well-tempered pseudorandom sequences exhaust the main possibilities to imitate randomness in a single sequence, and well-tempered maximum-length sequences are reliable and efficient random-number generators. Pseudorandom binary sequences are a simple example of a deterministic process that causes chaos, and they deserve the attention of theoretical physicists.

ACKNOWLEDGMENTS

It is a great pleasure to dedicate this article to Jerry Percus on the occasion of his 65th birthday, in appreciation of his outstanding contributions to mathematical physics, and of many stimulating discussions on continuum limits, combinatorial problems, and random numbers.

The research reported here is partially supported by a NATO grant, under reference CRG 900661.

REFERENCES

1. D. E. Knuth, *The Art of Computer Programming*, Vol. 2 (Addison-Wesley, 1981).
2. L. Afflerbach, *J. Comp. Appl. Math.* **31**:3 (1990).
3. G. Marsaglia and A. Zaman, preprint (1990).
4. F. James, *Comp. Phys. Commun.* **60**:329 (1990).
5. B. D. Ripley, *J. Comp. Appl. Math.* **31**:153 (1990).
6. H. Niederreiter, *Ann. Oper. Res.*, to be published.
7. M. van Lambalgen, *J. Symbolic Logic* **52**:725 (1987); Ph.D. Thesis, University of Amsterdam (1987).
8. A. Compagner and A. Hoogland, *J. Comp. Phys.* **71**:391 (1987).
9. A. Compagner, *Am. J. Phys.*, to be published (1991).
10. S. W. Golomb, *Shift Register Sequences* (Holden-Day, 1967).
11. N. Zierler, *Inform. Control* **15**:67 (1969).