Nonlinear dynamical model for texts

EUROPHYSICS LETTERS 20 August 1996

Europhys. Lett., **35** (6), pp. 401-406 (1996)

Nonlinear dynamical model for texts

K. Rateitschak(*), W. Ebeling and J. Freund

Institute of Physics, Humboldt University - Berlin
Invalidenstraße 110, D-10099 Berlin, Germany

(received 14 February 1996; accepted in final form 1 July 1996)

PACS. 02.50−r – Probability theory, stochastic processes, and statistics.
PACS. 05.45+b – Theory and models of chaotic systems.
PACS. 89.70+c – Information science.

Abstract. – We introduce a symbol sequence generator producing sequences with long-range correlations. We analytically derive the scaling behaviour of block entropies. A similar scaling behaviour was suggested for natural information carrying symbol sequences, *e.g.* texts.

Introduction. – A symbol sequence generator is a set of deterministic or stochastic or both rules to construct symbol sequences. Important questions are the strength and the range of correlations between the symbols.

Our aim is to find a generator producing sequences similar to natural, evolutionary created sequences, like texts [1], [2]. Information processing is one characteristic of living creatures. Reproduction, mutations, and variety play a central role in the evolution of life [3]. Consequently, our generator contains deterministic and stochastic rules.

In order to analyze the syntax of the sequences, one can use statistical tools. Let $A_1, A_2, \ldots, A_n$ be a subsequence or word of n symbols from an alphabet of λ different symbols. The probability of finding this n-word in the whole sequence is denoted by $p(A_1, A_2, \ldots, A_n)$. The block entropy of words of length n is defined by

$$H_n := -\sum_{(A_1,\ldots,A_n)} p(A_1,\ldots,A_n)\log p(A_1,\ldots,A_n)\ . \tag{1}$$

The H_n can be interpreted as the mean uncertainty about the prediction of an n-word. The average uncertainty per symbol is $H(n) := H_n/n$. Another quantity is the conditional entropy defined by

$$h_n := H_{n+1} - H_n\ . \tag{2}$$

The h_n establish a measure of the mean uncertainty about the prediction of a symbol following n known symbols. McMillan and Khinchin have shown [4], [5] that

$$h := \lim_{n\to\infty} H(n) = \lim_{n\to\infty} h_n\ . \tag{3}$$

(*) E-mail: katja@summa.physik.hu-berlin.de.

The limit h is named entropy of the source.

If the conditional entropies h_n decay exponentially, one cannot essentially improve the prediction of the next symbol even for relatively small n. This means that there exist only short-range correlations between the symbols.

If the h_n show a subexponential decay or a power law decay, then there are long-range correlations between the symbols. Even for large n one can considerably improve the prediction.

Szépfalusy and Győrgi have shown that a special weak intermittent system leads to $h_n \sim 1/n^\alpha, \alpha > 2.5$ with $h > 0$ [6].

Self-similar sequences, which are deterministic and fully calculable sequences, have been studied by Grassberger [7], Gramss [8], and ourselves [1], [9]-[11]. They show a $h_n \sim 1/n^\alpha, \alpha = 1$ law with $h = 0$.

Numerical analyses of text and music sequences yield $h_n \sim 1/n^\alpha$ with $0 < h \ll 1$ and $\alpha = 0.5$ for texts and $0.5 \leq \alpha \leq 1$ for music [2], [12], [13]. The exact exponents are difficult to extract from the empirical material. DNA sequences are strongly different from text and music on the statistical level. They have been studied in [14], [15].

In this paper we introduce a symbol generator producing binary symbol sequences. We will show that a resulting sample sequence gives rise to a subexponential decay of conditional entropies $h_n \sim 1/n^\alpha, \alpha < 1$. For $\alpha = 0.5$ we analytically derive an upper and a lower bound for the block entropies. Since these bounds essentially obey the same scaling law, namely a square root behaviour, we conclude that the overall scaling of block entropies necessarily has to be the same. Finally, a numerical example will be provided to validate our analytical results.

Description of the generator. – Starting from a series of equidistributed n-words, *i.e.* we have $2^{n^{1-\alpha}}$ different n-words with equal probability $p(A_1, \ldots, A_n) = 1/2^{n^{1-\alpha}}$, obviously will result in a scaling law $H_n \sim n^{1-\alpha}$, respectively, $h_n \sim 1/n^\alpha$. However, there exists no simple construction method yielding a (binary) sequence with such a related series of word distributions.

Nevertheless, the above trivial connection motivates to try the following sequence generator. (For the sake of simplicity we restrict to $\alpha = 0.5$ in the second and third sections. The general case will be discussed in the conclusions.)

First choose a length $n_{k_0} := 2^{2k_0}$ with arbitrary $k_0 \geq 0$. Out of all generally possible $2^{n_{k_0}}$ words collect $2^{(2^{k_0})}$ different sequences in a sample set. To construct a sequence of length n_{k_0+1} one has to independently and randomly select two words from this sample set (note that both selected words may be identical). Now both such selected words are considered as building blocks. They are concatenated and, in the sequel, the resulting sequence of length $n_{k_0+1}/2$ is repeated yielding a sequence of length n_{k_0+1}.

In this manner $2^{(2^{k_0+1})}$ different such sequences of length n_{k_0+1} can be constructed. Below these elements will be considered as "main words". All of those, again, are collected in a new sample set and the generator proceeds iteratively as explained above, *i.e.* with twice random selection, concatenation and repetition.

For an illustration, take a look at fig. 1 which exemplifies this construction process starting with $k_0 = 0$, *i.e.* $n_0 = 1$. Upper and lower bounds to the H_{n_k}, respectively h_{n_k}, derived below are valid only for $k \geq k_0$.

Scaling behaviour for the block entropies. – In this section we show how to relate the probability distribution $p(A_1, \ldots, A_{n_k})$ of n_k-words to a sample sequence produced by the above-described generator.

In practice, to extract the probability distribution one simply should do overlapping word-counting. Here, however, we will employ analytical reasoning to estimate $p(A_1, \ldots, p_{A_{n_k}})$ and, in the sequel, to derive bounds for n_k-block entropies.

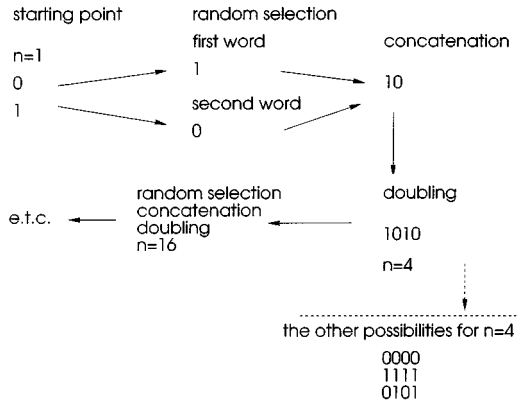

Fig. 1. – Sequence generator, $\alpha = 0.5$.

To get the probability of a word of length n_k, one has to count the sample: one of $2^{\sqrt{n}_k}$ main words follows every main word of length n_k. The sample consists of $2^{2\sqrt{n}_k}$ different words of length $2n_k$ with equal probability. We count overlapping and we have to count the words starting on the first n_k positions in a word of the sample.

In the following we omit the index k at n for reasons of clarity.

First we calculate an upper bound for H_n. We assume all occurring words in the sample to be different. We have $2^{\sqrt{n}}2^{\sqrt{n}}n$ words with probability $p = 1/(2^{\sqrt{n}}2^{\sqrt{n}}n)$. This yields

$$H_n = 2\sqrt{n} + \log n \; . \tag{4}$$

Instead of $h_n = H_{n+1} - H_n$, one can approximately write $h_n = \frac{\mathrm{d}}{\mathrm{d}n} H_n$:

$$h_n = \frac{1}{\sqrt{n}} + \frac{1}{n} \; . \tag{5}$$

For the entropy of the source we get

$$h = \lim_{n\to\infty} h_n = 0 \; . \tag{6}$$

The reason for $h = 0$ is that the portion of randomly selected symbols vanishes for $n_k \to \infty$. To be specific: the portion of randomly selected symbols is $\frac{1}{\sqrt{n}_k}$.

Now we calculate a lower bound for the H_n. Our ansatz becomes better for larger n. In the following we consider two groups of words that have a contribution of $\frac{2}{3}$ to the standard words.

The first group is composed by the words that start in the middle of a first main word and end in the middle of the next main word. Most of these words occur only once in the sample because outside the word one does not randomly select a symbol (the second half of a main word is the same as the first half). They are named minimum class words or mc-words. But if the main words are similar, then the word in the middle can also occur at other positions in the first main word and the frequency is greater than one.

A main word of length n consists of four main words of length $\frac{n}{4}$: A, B, C, and D are different main words of length $\frac{n}{4}$. The following compositions of main words contribute to the mc-words.

1st main word	AA	AA	AB	AB	AB	AB	AB,
2nd main word	BB	BC	BA	CA	BC	CD	CC.

It is easy to calculate the number of different words. For the first row we get A: $2^{\sqrt{\frac{n}{4}}}$ and B: $2^{\sqrt{\frac{n}{4}}}-1$ different words. This yields $2^{\sqrt{\frac{n}{4}}}(2^{\sqrt{\frac{n}{4}}}-1)$ different words. The whole table leads to $(2^{\sqrt{n}}-2^{\sqrt{\frac{n}{4}}})^2$ different words. Mc-words also occur on other positions in the first main word. The last

$$\frac{n}{8}+\frac{n}{32}+\frac{n}{128}+\ldots+2 = \sum_{j=2}^{\log_4 n}\frac{2n}{4^j}, \tag{7}$$

$$\frac{n}{8}+\frac{n}{32}+\frac{n}{128}+\ldots+2 = \frac{n}{6}-\frac{2}{3} \tag{8}$$

positions before and including the position $\frac{n}{2}$ in a main word are repetitions of earlier positions. If the mc-word finishes at one of these positions, then outside the word one does not randomly select a symbol and the word occurs only once. Positions after $\frac{n}{2}$ are repeated later in the first main word. This yields a contribution of $\frac{n}{6}+\frac{1}{3}$ positions.

In total we have $(2^{\sqrt{n}}-2^{\sqrt{\frac{n}{4}}})^2(\frac{n}{3}-\frac{1}{3}) \sim 2^{2\sqrt{n}}(\frac{n}{3}-\frac{1}{3})$ different words with probability $p=1/(2^{2\sqrt{n}}n)$. The contribution to the normalization is $\frac{1}{3}$ for $n\to\infty$.

The second group consists of words starting at positions around $\frac{n}{4}$. One randomly selects a symbol $\sqrt{n}$ times in the contribution of the first main word and one randomly selects a symbol $\sqrt{\frac{n}{4}}$ times in the contribution of the second main word. This yields $2^{\frac{3}{2}\sqrt{n}}$ different words at one position. The last

$$\frac{n}{8}+\frac{n}{32}+\frac{n}{128}+\ldots+2 = \sum_{j=2}^{\log_4 n}\frac{2n}{4^j}, \tag{9}$$

$$\frac{n}{8}+\frac{n}{32}+\frac{n}{128}+\ldots+2 = \frac{n}{6}-\frac{2}{3} \tag{10}$$

positions before and including the position $\frac{n}{4}$ are repetitions of earlier positions. If a word finishes at one of these positions, then it also belongs to this class. To the last expression one has to add 1 for the word at position $\frac{n}{4}+1$ and to multiply it by 2 because the arguments are also valid for the words starting around position $\frac{3}{4}n$ in the first main word.

In total we have $2^{\frac{3}{2}\sqrt{n}}(\frac{n}{3}+\frac{2}{3})$ different words with probability $1/(2^{\frac{3}{2}\sqrt{n}}n)$. The contribution to the normalization is $\frac{1}{3}$ for $n\to\infty$.

The entropy of these two contributions is

$$H_n \sim \frac{7}{6}\sqrt{n}+\frac{2}{3}\log n+\mathcal{O}(\frac{1}{\sqrt{n}}). \tag{11}$$

It is impossible that the gradient of the scaling law of H_n is smaller than the right side of eq. (11). Otherwise the part of the entropy of the words contributing to the last $\frac{1}{3}$ of the normalization has the form $-\sqrt{n}+f(n)$ and the gradient of the scaling law of $f(n)$ is smaller than $\sqrt{n}$, for instance, $f(n)\sim\log n$. The expression $-\sqrt{n}+f(n)$ would then become negative. This is forbidden because $-p\log p\geq 0$.

Figure 2 shows the rank-ordered word distributions for word lengths $n=16$ and $n=64$.

Figure 3 shows the block entropies H_n. One can see that the numerical data of the entropies H_n of our generated sequence are smaller than the upper bound (4) and greater than the lower bound (11).

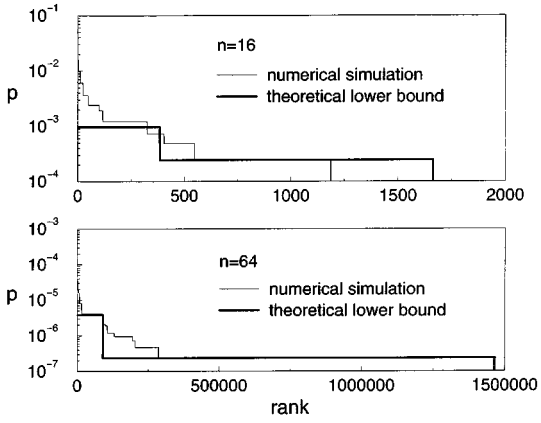

Fig. 2.

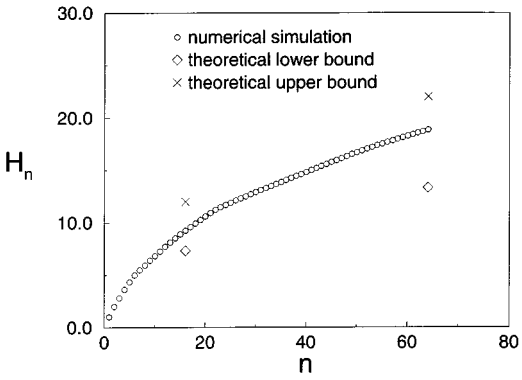

Fig. 3.

Fig. 2. – Rank-ordered word distributions.

Fig. 3. – Block entropies.

The lower and the upper bound for H_n yield a

$$H_n \sim a\sqrt{n} + b \log n \tag{12}$$

law (a and b are constants) and, consequently, eq. (12) follows for our numerical data.

For h_n we get

$$h_n \sim \frac{c}{\sqrt{n}} + \frac{d}{n} \tag{13}$$

(c and d are constants). This scaling behaviour indicates long-range correlations and was found for texts [2], [12], [13].

Conclusions and discussion of the results. – We have related an exceptional entropy scaling behaviour to a special sequence generator. There exists an interplay between random selection and repetition.

Referring to the number of different words the random selection causes the factor $2^{\sqrt{n}}$ and the repetition is responsible for the factor n. The numerator $2^{\sqrt{n}}$ of the probabilities also results from the random selection. Similar rules may play a role in the grammar of texts [16].

Finally, we extend our results to arbitrary α: The generator works for $0 < \alpha < 1$ and for $\frac{1}{1-\alpha}$ being an integer. Again, one starts by choosing a length $n_{k_0} := 2^{k_0/(1-\alpha)}$ with arbitrary $k_0 \geq 0$. In analogy, out of all generally possible $2^{n_{k_0}}$ words one collects $2^{(2^{k_0})}$ different ones in a sample set. To construct a sequence of length n_{k_0+1} one has to independently and randomly select two words from this sample set (both selected words may be identical). Finally, they are concatenated and, in the sequel, $2^{\alpha/(1-\alpha)}$ copies of the resulting sequence are produced and concatenated yielding a sequence of length n_{k_0+1}. For the estimation of n_k-word distributions and derivation of upper, respectively lower bounds analogous reasoning is straightforward.

***

Encouragement and support by I. Gross e and T. Pöschel are greatly acknowledged.

REFERENCES

[1] Ebeling W. and Nicolis G., *Chaos Solitons Fractals*, **2** (1992) 635.

[2] Ebeling W., Pöschel T. and Albrecht K. F., *Int. J. Bifurc. Chaos*, **5** (1995) 51.

[3] Ebeling W., Engel A. and Feistel R., *Physik der Evolutionsprozesse* (Akademie-Verlag, Berlin) 1990.

[4] Khinchin A. I., *Mathematical Foundations of Information Theory* (Dover Publ., New York, N.Y.) 1957.

[5] McMillan B., *Ann. Math. Stat.*, **24** (1953) 196.

[6] Szépfalusy P. and Győrgyi G., *Phys. Rev. A*, **33** (1986) 2852.

[7] Grassberger P., *Int. J. Theor. Phys.*, **25** (1986) 907.

[8] Gramss T., *Phys. Rev. E*, **50** (1994) 2616.

[9] Ebeling W., Freund J. and Rateitschak K., to be published in *Int. J. Bifurc. Chaos* (1996).

[10] Freund J., Ebeling W. and Rateitschak K., to be published in *Phys. Rev. E* (1996).

[11] Rateitschak K., Freund J. and Ebeling W., in *Entropy and Entropy Generation: Fundamentals and Applications*, edited by J. S. Shiner (Kluwer, Dordrecht) 1996.

[12] Ebeling W. and Pöschel T., *Europhys. Lett.*, **26** (1994) 241.

[13] Ebeling W., Neiman A. and Pöschel T., in *Coherent Approach to Fluctuations* (*Proceedings of Hayashibara Forum 95*), edited by M. Suzuki (World Scientific, Singapore) 1995.

[14] Herzel H. and Grosse I., *Physica A*, **216** (1995) 518.

[15] Herzel H., Ebeling W. and Schmitt A. O., *Phys. Rev. E*, **50** (1994) 5061.

[16] Chomsky N., *Syntactic Structures* (Mouton, Den Haag) 1957, 1962.