Decomposition of conditional probability for high-order symbolic Markov chains

S. S. Melnik and O. V. Usatenko

A. Ya. Usikov Institute for Radiophysics and Electronics Ukrainian Academy of Science, 12 Proskura Street, 61805 Kharkov, Ukraine

(Received 30 March 2017; revised manuscript received 9 June 2017; published 31 July 2017)

The main goal of this paper is to develop an estimate for the conditional probability function of random stationary ergodic symbolic sequences with elements belonging to a finite alphabet. We elaborate on a decomposition procedure for the conditional probability function of sequences considered to be high-order Markov chains. We represent the conditional probability function as the sum of multilinear memory function monomials of different orders (from zero up to the chain order). This allows us to introduce a family of Markov chain models and to construct artificial sequences via a method of successive iterations, taking into account at each step increasingly high correlations among random elements. At weak correlations, the memory functions are uniquely expressed in terms of the high-order symbolic correlation functions. The proposed method fills the gap between two approaches, namely the likelihood estimation and the additive Markov chains. The obtained results may have applications for sequential approximation of artificial neural network training.

DOI: 10.1103/PhysRevE.96.012158

I. INTRODUCTION

Natural sequences with a nontrivial information content have been the focus of a large number of studies in different fields of science for the past several decades. Such random sequences are a subject of study in the areas of algorithmic (Kolmogorov-Solomonoff-Chaitin) complexity, information theory, compressibility of digital data, the statistical inference problem, computability, data compression [1], natural language processing [2], and artificial intelligence [3]. In addition, many aspects of these sequences can be applied as a creative tool for designing devices and appliances with random components in their structure (e.g., different wave filters, diffraction gratings, artificial materials, antennas, converters, delay lines, etc. [4]).

Random sequences with a *finite state space* exist as natural sequences (e.g., DNA or natural language texts), or they arise as a result of coarse-grained mapping of the evolution of a chaotic dynamical system into a string of symbols [5–7]. The sequence items can also be phonemes, syllables, words, or DNA base pairs, depending on the application.

A standard method of understanding and describing the statistical properties of a given random symbolic sequence of data requires an estimation of the joint probability function of L words (subsequences of length L). Reliable estimations for the probability can be achieved only for small L because the number m^L (where m is the finite-alphabet length) of different words of length L has to be much less than the total number of words, $m^L \ll M - L$, in the whole sequence of length M. This is a crucial point, because the correlation lengths of natural sequences of interest are usually of the same order as the sequence length, whereas the last inequality can only be fulfilled for the maximal lengths of the words, $L_{\max} \lesssim 10$.

The main purpose and results of this paper can be described as follows: We elaborate on a decomposition procedure for the conditional probability function of sequences considered to be high-order Markov chains. We represent the conditional probability function as the sum of multilinear monomials of different orders (from zero up to the chain order). At weak correlations, the memory functions are uniquely expressed in terms of high-order symbolic correlation functions. Finally, we reveal a close connection between our analytical Markov chain approach and artificial neuron network models.

II. SYMBOLIC MARKOV CHAINS

Consider a semi-infinite random stationary ergodic sequence $\mathbb{S}$ of symbols (letters, characters) a_i,

$$\mathbb{S} = a_0, a_1, a_2, \dots \tag{1}$$

taken from the finite alphabet

$$\mathcal{A} = \{\alpha^1, \alpha^2, \dots, \alpha^m\},\ a_i \in \mathcal{A},\ i \in \mathbb{N}_+ = \{0,1,2,\dots\}. \tag{2}$$

We use the notation a_i to indicate a position i of the symbol a in the chain, and the unified notation α^k to stress the value of the symbol $a \in \mathcal{A}$. We also use the personified notation for the symbols a of the same alphabet, $\mathcal{A} = \{\alpha, \beta, \dots, \omega\}$.

We suppose that the symbolic sequence $\mathbb{S}$ is a *high-order Markov chain*. The sequence $\mathbb{S}$ is a Markov chain if it possesses the following property: the probability of symbol a_i to have a certain value $\alpha^k \in \mathcal{A}$ under the condition that *all* previous symbols are fixed depends only on N previous symbols,

$$\begin{aligned} &P(a_i = \alpha^k | \dots, a_{i-2}, a_{i-1}) \\ &\quad = P(a_i = \alpha^k | a_{i-N}, \dots, a_{i-2}, a_{i-1}). \end{aligned} \tag{3}$$

There are many other terms for such sequences. They are also referred to as follows: *categorical* [8], *higher-order* [9,10], and *multi-* or *N-step* [11–13] Markov chains. One of the most important and interesting applications of the symbolic sequences is the probabilistic language model, which specializes in predicting the next item in a sequence by means of N previous known symbols. Here the Markov chain is known as the *N-gram model.*

As a rule, the statistical properties of random sequences are determined using correlation functions. In contrast with numeric correlation functions,

$$\begin{aligned} &C(r_1, r_2, \dots, r_{k-1}) \\ &\quad = \overline{[a_0 - \bar{a})][a_{r_1} - \bar{a})] \cdots [a_{r_1+\dots+r_{k-1}} - \bar{a}]}, \end{aligned} \tag{4}$$

symbolic correlation functions of kth order are given by the following expression:

$$C_{\beta_1,\dots\beta_k}(r_1,r_2,\dots,r_{k-1}) = \overline{[\delta(a_0,\beta_1)-p_{\beta_1}]\cdots[\delta(a_{r_1+\dots+r_{k-1}},\beta_k)-p_{\beta_k}]}. \quad (5)$$

The overline indicates a statistical average over an ensemble of sequences. For numerical purposes, it can be replaced by the average along the chain for ergodic sequences, or by the arithmetic, Cesàro average. Note that in some sense, symbolic correlation function matrices are a more general construction than numeric correlation functions. They can describe in more detail even numeric sequences.

A. Likelihood estimation

If the sequence, the statistical properties of which we would like to analyze, is given, then the conditional probability function (CPF) of Nth order can be found using a standard method known as the likelihood estimation,

$$P(a_i=\alpha|a_{i-N},\dots,a_{i-1}) = \frac{P(a_{i-N},\dots,a_{i-1},\alpha)}{P(a_{i-N},\dots,a_{i-1})}, \quad (6)$$

where $P(a_{i-N},\dots,a_{i-1},\alpha)$ and $P(a_{i-N},\dots,a_{i-1})$ are the probabilities that the $(N+1)$-subsequence $a_{i-N},\dots,a_{i-1},\alpha$ and the N-subsequence $a_{i-N},\dots,a_{i-1}$ will occur, respectively. Hereafter, we often drop the superscript k from α^k to simplify the notations.

The conditional probability function completely determines *all the statistical properties* of the random chain and the method of its generation. Equation (6) shows that the CPF is determined if we know the probability that $(N+1)$ words will occur, with the words containing $(N+1)$ symbols without omissions among their indexes. Obviously, the average number of some word $a_1,\dots,a_L$ occurring in the whole sequence decreases exponentially with its length L. Let us evaluate the length $L_{\max}$ of a word that occurs on average one time. For a given length M of a weakly correlated sequence with fixed dimension m of the alphabet, this length, evidently, is equal to $L_{\max}\approx \ln M/\ln m$.

To make this evaluation more precise, we should take into account that the correlations decrease the number of *typical* words that one can encounter in the sequence, and this phenomenon increases the length $L_{\max}$. From the famous result of the information theory, known as the Shennon-McMillan-Breiman theorem [14], it follows that

$$L_{\max}\sim\frac{\log_2 M}{H}, \quad (7)$$

where H is the conditional entropy per letter of the sequence under the condition that all correlations are taken into account.

The words of length $L\ll L_{\max}$ are well represented in the sequence, thus one can use the statistical approach for these objects and calculate the probabilities of their occurrence in the chain. By contrast, the statistics of longer words, $L\gtrsim L_{\max}$, is insufficient, and the whole sequence for such words is no longer a probabilistic object. Many papers devoted to this topic call into question even the possibility of the existence of a "finite random sequence" [15,16].

Therefore, if the correlation length R_c of a sequence is less than $L_{\max}$, then the random sequence under consideration should be deemed quasiergodic because the words of length $L<R_c<L_{\max}$ provide statistically meaningful information for reconstructing the conditional probability function of the sequence.

We encounter a completely different situation when $L_{\max}<R_c$. In that case, the statistical properties of the studied sequence can be reconstructed only up to the length of order $L\ll L_{\max}$. Statistically important information on the property of the sequence in the interval $L_{\max}<L<R_c$ is inaccessible in the framework of the discussed likelihood estimation method.

For the sake of simplicity, let us fix our attention on the pair correlation function $C_2(r)$ only. If we know the statistics of $(N+1)$ words, we also know $C_2(r)$ for $r\leqslant N$. Nevertheless, at the same time, for the given sequence of length M, we can calculate $C_2(r)$ at r, which is of order M. For a weakly correlated sequence, the probability $P(a_i=\alpha,a_{i+r}=\beta)$ that the pair of letters α and β will be separated by a distance r is equal to $p_\alpha p_\beta$. This quantity determines the number of pairs in the hole sequence. The number is a decreasing function of r. As above, we can evaluate the distance $r_{\max}$ between the pair of letters that occurs on average one time. Thus, we have $r_{\max}\sim M-1/(p_\alpha p_\beta)$. It is clear that $r_{\max}$ can be much greater than $L_{\max}$.

For k-order correlation functions or k-words, the estimation is $r_{\max}^{(k)}\sim M-1/(p_{\alpha_1},\dots,p_{\alpha_k})$.

Let us note that in the frameworks of both methods, we cannot take into account the correlation functions of order higher than $L_{\max}$. This quantity determines both the maximal length of words, without or with the omission of symbols among them in the sequence (in mathematics, such sets are known as *cylinder* sets), and the maximal order of correlation functions, which can be used to describe the statistical properties of the sequence. In the method of likelihood estimation, this length limits the differences among the arguments of the correlation functions. In the general case, the differences among the arguments of the correlation functions are limited by $r_{\max}^{(k)}$. This information about the region $L_{\max}\lesssim L\ll \min(R_c,r_{\max})$ is presented for consideration by means of the memory functions, which are expressed through the correlation functions; see Eqs. (10), (25), and (26).

A method that allows us to use the information on the symbols separated by a distance $r\ll\min(R_c,r_{\max})$, not only in the narrower region with $r\ll L_{\max}$, is connected with the high-order additive Markov chains, a construction proposed in Refs. [17,18]. In this paper, we develop this method and introduce the family of multilinear high-order Markov chains.

B. Additive high-order Markov chains

For an *additive* high-order Markov chain, the conditional probability function takes on a specific, simplified, "linear form" with respect to the random variables a_i,

$$P_{\rm add} = P^{(1)}(a_i=\alpha|a_{i-N},\dots,a_{i-2},a_{i-1}) = p_\alpha+\sum_{r=1}^{N}\sum_{\beta\in\mathcal{A}}F_{\alpha\beta}(r)[\delta(a_{i-r},\beta)-p_\beta]. \quad (8)$$

Here p_β is the relative number of symbols β in the chain, or their probabilities of occurrence,

$$p_\beta = \overline{\delta(a_i,\beta)}. \tag{9}$$

The Kronecker delta $\delta(.,.)$ plays the role of the characteristic function of the random variable a_i, and it converts symbols to numbers. The additivity of the chain means that the "previous" symbols $a_{i-N},\dots,a_{i-2},a_{i-1} \equiv a_{i-N}^{i-1}$ exert an independent effect on the probability that the "final" generated symbol $a_i = \alpha$ will occur. The first term on the right-hand side of Eq. (8) is responsible for the correct reproduction of the statistical properties of uncorrelated sequences, while the second one takes into account and correctly reproduces the correlation properties of the chain up to second order. The higher-order correlation functions are not independent anymore. We cannot control them and reproduce correctly by means of the memory function $F(r)$.

There were two methods suggested for finding the *memory functions* $F_{\alpha\beta}(r)$ of a sequence with known pair correlation functions. The first one is a completely probabilistic straightforward calculation analogous to that presented in [17]. Its modification is used below while obtaining Eqs. (25) and (26). For any values of $\alpha,\beta \in \mathcal{A}$ and $r \geq 1$, the relationship between the correlation and memory functions was obtained,

$$C_{\alpha\beta}(r) = \sum_{r'=1}^{N} \sum_{\gamma\in\mathcal{A}} C_{\alpha\gamma}(r-r')F_{\beta\gamma}(r'). \tag{10}$$

Here the two-point (binary, pair) symbolic correlation function is the particular case of definition (5),

$$C_{\alpha\beta}(r) = \overline{[\delta(a_i,\alpha) - p_\alpha][\delta(a_{i+r},\beta) - p_\beta]}. \tag{11}$$

The second method [18] for deriving Eq. (10) is based on the minimization of the "distance" between the conditional probability function, containing the sought-after memory function, and the given sequence $\mathbb{S}$ of symbols with the known correlation functions,

$$\text{Dist} = \overline{\left[\delta(a_i,\alpha) - P\left(a_i = \alpha \middle| a_{i-N}^{i-1}\right)\right]^2}. \tag{12}$$

It is difficult to make a final conclusion about the relationship between these two methods and the cause of their coincidence. The principle of distance minimization may be considered as the ansatz, which finds its probabilistic confirmation. However, this second method also has an independent interest, both theoretical and practical. The first one explains that the memory function provides the minimization of some characteristics, e.g., the distance between the original and artificial sequences; the practical utility may consist in the use of numerical simulations for finding the unknown memory functions.

Let us note that in the considered case, the two-point quantities—the memory and correlation functions—are not obliged to satisfy the strong inequality $L \ll L_{\max}$. In other words, the additive Markov chain determined by Eq. (8) can describe and predict the two-point statistical properties of random sequences at distances longer than that based on Eq. (4). At the same time, the model of random sequences based on the likelihood estimation works better at short distances; see, e.g., the result of the DNA entropy estimation [19,20], where the discrepancy between the two theories is evident. The next step to improve the prediction quality of symbols in a random chain, which we present here, is based on the k-linear conditional probability function, which should close the above-mentioned gap and, probably, explain the very astonishing question of why binary correlations are so important and so long. The additive Markov chains are, in some sense, analogous to the chains described by autoregressive models [7,9,21].

III. DECOMPOSITION OF THE CONDITIONAL PROBABILITY FUNCTION

Equation (8) can be considered as an approximate model expression simplifying the general form of the conditional probability function. As a matter of fact, the conditional probability (6) of the symbolic sequence of random variables $a_i \in \mathcal{A}$ can be represented exactly as a *finite* polynomial series containing N Kronecker δ symbols: a specific decomposed form of the CPF, which expresses some "independence" of the random variables a and spatial coordinates i,

$$P(.|.) = P\left(a_i = \alpha \middle| a_{i-N}^{i-1}\right)$$

$$= \sum_{\beta_1,\dots,\beta_N\in\mathcal{A}} F_{\alpha;\beta_1,\dots,\beta_N}(1,\dots,N) \prod_{s=0}^{N} \delta\left(a_{i-r_s},\beta_s\right). \tag{13}$$

Here the arguments $1,\dots,N$ of the function $F_{\alpha;\beta_1,\dots,\beta_N}(1,\dots,N)$ indicate the distances between the final "generated" symbol $a_i = \alpha$ and symbols $a_{i-1},\dots,a_{i-N}$. It is clear that there is a one-to-one correspondence between $P(a_i = \alpha|a_{i-N}^{i-1})$ and the function $F_{\alpha;\beta_1,\dots,\beta_N}(1,\dots,N)$, which is referred to as the *generalized memory function*. The characteristic functions $\delta(a_{i-r_s},\beta_s)$ play the role of a basis, and the generalized memory functions are coordinates of the CPF. We hope the reader paid attention to the difference between α^k, $k \in (1,\dots,m)$ and α_s, where the subscript s, $s \in (1,\dots,N)$ enumerates different sets of letters.

Let us decouple the memory function $F_{\alpha;\beta_1,\dots,\beta_N}(1,\dots,N)$ and present it in the form of the sum of *memory functions* of *k*th *order*, $F^{(k)} = F_{\alpha;\beta_{i_1},\dots,\beta_{i_k}}(r_{i_1},\dots,r_{i_k})$,

$$F_{\alpha;\beta_1,\dots,\beta_N}(1,\dots,N) = \sum_{k=0}^{N} F_{\alpha;\beta_{i_1},\dots,\beta_{i_k}}(r_{i_1},\dots,r_{i_k}), \tag{14}$$

where all symbols r_i on the right-hand side (RHS) of Eq. (14) are different, ordered,

$$1 \leqslant r_{i_1} < r_{i_2} < \dots < r_{i_k} \leqslant N, \tag{15}$$

and contain all different subsets $\{r_{i_1},\dots,r_{i_k}\}$ picked out from the set $\{1,\dots,N\}$. The coordinates r_{i_k} of the memory function $F_{\alpha;\beta_{i_1},\dots,\beta_{i_k}}(r_{i_1},\dots,r_{i_k})$ indicate the positions of elements β_{i_k}, except the function $F_{\alpha;\beta_1,\dots,\beta_N}(1,\dots,N)$, for which all symbols $\beta_1,\dots,\beta_N$ have the prescribed coordinates $1,\dots,N$.

Inserting (14) in (13) and summing over the subset $\{r_{i_1},\dots,r_{i_k}\}$ variables satisfying Eq. (15), we get

$$P\left(a_i = \alpha \middle| a_{i-N}^{i-1}\right) = \sum_{k=0}^{N} Q^{(k)}\left(a_i = \alpha \middle| a_{i-N}^{i-1}\right) \tag{16}$$

with the following definitions of $Q^{(0)}$ and $Q^{(1)}$, see Eqs. (8):

$$Q^{(0)}\big(a_i = \alpha \big| a_{i-N}^{i-1}\big) = p_\alpha \tag{17}$$

and

$$Q^{(1)}(\alpha|.) = \sum_{\beta\in\mathcal{A}} \sum_{r=1}^{N} F_{\alpha\beta}(r)[\delta(a_{i-r},\beta) - p_\beta]. \tag{18}$$

The general term of kth order, referred to in multilinear algebra as a k-linear form, is

$$Q^{(k)}(.|.) = \sum_{\beta_1,\dots,\beta_k\in\mathcal{A}} \; \sum_{1\leqslant r_1<\cdots<r_k\leqslant N} F_{\alpha;\beta_1,\dots,\beta_k}(r_1,\dots,r_k)\left\{\prod_{s=1}^{k}\big[\delta\big(a_{i-r_s},\beta_s\big) - p_{\beta_s}\big] - C_{\beta_k,\dots,\beta_1}(r_k - r_{k-1},\dots,r_k - r_1)\right\}. \tag{19}$$

In Eq. (19), we have added the term $C_{\beta_1,\dots,\beta_k}$ to provide the equality $\overline{Q^{(k)}(.|.)} = 0$ for all $k = 1,2,\dots,N$. With this property, the ensemble-average value of the conditional probability function is always equal to p_α.

Definition (5) is correct for $r_i > 0$, $i = 1,\dots,k-1$. If some arguments of the correlation function in Eq. (5) are negative, one should interpret it in the following way, which is referred [22] to as "collating": we should order the arguments of the function according to definition (5). For example,

$$\begin{aligned} C_{\alpha\beta\gamma\delta}(2,2,-3) &= \overline{[\delta(a_0,\alpha) - p_\alpha][\delta(a_2,\beta) - p_\beta][\delta(a_4,\gamma) - p_\gamma][\delta(a_1,\delta) - p_\delta]} \\ &= \overline{[\delta(a_0,\alpha) - p_\alpha][\delta(a_1,\delta) - p_\delta][\delta(a_2,\beta) - p_\beta][\delta(a_4,\gamma) - p_\gamma]} = C_{\alpha\delta\beta\gamma}(1,1,2). \end{aligned} \tag{20}$$

We used this method to represent the correlation function in Eq. (19) in the collating form.

The function $C_{\alpha_1,\dots,\alpha_k}$ depends on k arguments r, but the Markov chain under consideration is supposed to be homogeneous. Then function $C_{\alpha_1,\dots,\alpha_k}$ depends on $k-1$ arguments, the differences between the indexes, $r_1,r_2,\dots,r_{k-1}$, of neighbor symbols. A trivial property of the function $C_{\alpha_1,\dots,\alpha_k}(r_1,\dots,r_{k-1})$ is

$$\sum_{\alpha_m\in\mathcal{A}} C_{\alpha_1,\dots,\alpha_k}(r_1,\dots,r_{k-1}) = 0, \quad 1 \leqslant m \leqslant k. \tag{21}$$

The last term $Q^{(N)}$ in Eq. (16) contains arguments $r_k = k$. For each fixed set of symbols $\alpha;\ \beta_1,\dots,\beta_N$ there is just one matrix constant $F_{\alpha;\beta_1,\dots,\beta_N}(1,\dots,N)$.

The *k-linear* conditional probability function $P^{(k)}(.|.)$ in the form of Eqs. (16) and (19) can reproduce correctly the correlations of the Markov chain up to $(k+1)$th order. For the value of $k = N$, the function $P^{(N)}(.|.)$ represents exactly the function $P(a_i = \alpha|a_{i-N}^{i-1})$ [Eq. (6)].

Thus, the conditional probability function $P(a_i = \alpha^k|a_{i-N}^{i-1})$ is presented as a decomposition of multilinear form into k-linear subspaces. Earlier, a similar idea was presented by Hosseinia, Leb, and Zideka [8], who proved rigorously that the conditional probability function can be written as a linear combination of the monomials of past process responses for the Markov chain; see also Besag's paper [23].

The utility of the decomposition procedure can be explained in the following way. First, the partial terms $Q^{(k)}$ of the CPF are mainly responsible for reproducing the correlation properties of $(k+1)$th order [see Eq. (33)], and second, they can be considered as an asymptotic successive approximation for the the CPF.

A. Bilinear CPF

The model of the additive high-order Markov chain is well studied [12]. In this section, we examine high-order Markov chains with a *bilinear* conditional probability function.

The right-hand side of Eq. (8) contains two first terms of the asymptotic expression of the exact form, Eq. (13). The next term $Q^{(2)}$ is

$$\begin{aligned} Q^{(2)}(.|.) &= Q^{(2)}\big(a_i = \alpha\big|a_{i-N}^{i-1}\big) \\ &= \sum_{\beta,\gamma\in\mathcal{A}} \; \sum_{1\leqslant r_1<r_2\leqslant N} F_{\alpha;\beta\gamma}(r_1,r_2) \\ &\quad \times\{[\delta(a_{i-r_1},\beta) - p_\beta][\delta(a_{i-r_2},\gamma) - p_\gamma] \\ &\quad - C_{\gamma\beta}(r_2 - r_1)\}. \end{aligned} \tag{22}$$

The conditional probability function, which contains the linear term $P_{\text{add}} = P^{(1)}(.|.)$ and the bilinear function $Q^{(2)}$ [see Eqs. (8) and (22)], defines the *bilinear Markov chain*,

$$P_{\text{bilin}}(.|.) = P^{(2)}(.|.) = P^{(1)}(.|.) + Q^{(2)}(.|.). \tag{23}$$

It is possible to find the recurrence relations for the correlation functions of the N-step bilinear Markov chain. For this purpose, first of all we should calculate explicitly the average value of symbol $a_{r_1+\dots+r_{k-1}}$ in Eq. (5). Taking into account the equation $P(a = \alpha|\cdot) + P(a \neq \alpha|\cdot) = 1$, we can rewrite Eq. (5) for arbitrary $k \geq 2$ in the form

$$\begin{aligned} &\overline{[\delta(a_0,\beta_1) - p_{\beta_1}]\cdots[\delta(a_{r_1+\dots+r_{k-1}},\beta_k) - p_{\beta_k}]} \\ &\quad = \overline{[\delta(a_0,\beta_1) - p_{\beta_1}]\cdots[P(a_{r_1+\dots+r_{k-1}} = \beta_k|\cdot) - p_{\beta_k}]}, \end{aligned} \tag{24}$$

where the CPF $P(a_{r_1+\dots+r_{k-1}} = \alpha_k|\cdot)$ is given by Eq. (16). In that way, we can obtain the fundamental recurrence relation connecting the correlation functions of different orders k. Here we restrict ourselves by presenting these equations for the correlation functions $C_{\alpha\beta}(r)$ and $C_{\alpha\beta\gamma}(r_1,r_2)$ and the bilinear

CPF, Eq. (23),

$$C_{\alpha\beta}(r) = \sum_{r_1=1}^{N}\sum_{\gamma\in\mathcal{A}} C_{\alpha\gamma}(r-r_1)F_{\beta\gamma}(r_1) + \sum_{\gamma,\varepsilon\in\mathcal{A}}\ \sum_{1\leqslant r_1<r_2\leqslant N} C_{\alpha\varepsilon\gamma}(r-r_2,r_2-r_1)F_{\beta;\gamma\varepsilon}(r_1,r_2). \tag{25}$$

The equation for $C_{\alpha\beta\gamma}(r_1,r_2)$ reads

$$\begin{aligned}C_{\alpha\beta\gamma}(r_1,r_2) &= \sum_{\eta\in\mathcal{A}}\sum_{r_1'=1}^{N} C_{\alpha\beta\eta}(r_1,r_2-r_1')F_{\gamma\eta}(r_1')\\ &\quad+\sum_{\eta,\varepsilon\in\mathcal{A}}\ \sum_{1\leqslant r_1'<r_2'\leqslant N} F_{\gamma;\eta\varepsilon}(r_1',r_2')\\ &\quad\times[C_{\alpha\beta\varepsilon\eta}(r_1,r_2-r_2',r_2'-r_1')\\ &\quad-C_{\alpha\beta}(r_1)C_{\varepsilon\eta}(r_2'-r_1')].\end{aligned} \tag{26}$$

The other way to get Eqs. (25) and (26) is based on the minimization of the "distance," Eq. (12).

The system of equations (26) allows us to find the unknown memory functions $F_{\alpha\beta}(r)$ and $F_{\alpha\beta\gamma}(r_1,r_2)$ for consecutive construction of a representative random sequence with given correlation functions of second and third order. The memory functions should be expressed by means of the probability p_α and the correlation functions $C_{\alpha\beta}(r)$ and $C_{\alpha\beta\gamma}(r_1,r_2)$, which can be found numerically by means of an analysis of a given random chain.

Equations (25) and (26) enable us to understand that higher-order correlators $C^{(k)}$ ($k > 3$) and all correlation properties of higher order are not independent anymore. We cannot control them and reproduce correctly by means of the memory function $F_{\alpha\beta}(r)$ and $F_{\gamma;\eta\varepsilon}(r_1,r_2)$ because the latter is completely determined by the pair and third-order correlation functions.

We can make a similar conclusion about kth-order memory functions. The N-step Markov chain with a k-linear memory function allows us to reproduce correctly the chains up to the correlation function of $(k+1)$th order.

B. Approximate solution of equations

Equations (25) and (26) can be solved analytically only in some particular cases: for one- or two-step chains, for the Markov chain with a stepwise memory function, and so on. Here we give their approximate solution supposing that correlations in the sequence are not too strong (in amplitude, but not in length), and the alphabet $\mathcal{A}$ contains many letters. To formulate these conditions, we introduce the *normalized* symbolic correlation function defined by

$$K_{\alpha\beta}(r) = \frac{C_{\alpha\beta}(r)}{C_{\alpha\beta}(0)}, \quad C_{\alpha\beta}(0) = p_\alpha\delta(\alpha,\beta) - p_\alpha p_\beta. \tag{27}$$

If correlations in the random chain are not strong, it is plausible to suppose that all the components of the normalized correlation function with $r \neq 0$ are small with respect to $K_{\alpha\beta}(0) = 1$.

Neglecting the second term on the right-hand side of Eq. (25) [the correctness of this approximation is explained below, after Eq. (34)], we get Eq. (10). The solution of this equation can be written in the form

$$F_{\alpha\beta}(r) = \frac{1}{p_\beta}C_{\beta\alpha}(r) \tag{28}$$

if in definition (27) of $C_{\alpha\beta}(0)$ we can neglect the term $p_\alpha p_\beta$ with respect to p_α. This is possible if the dimension $|\mathcal{A}|$ of alphabet $\mathcal{A}$ satisfies the condition

$$|\mathcal{A}| = m \gg 1, \tag{29}$$

so that all probabilities p_α are small.

It is easy to see that after substituting Eq. (28) into (8), we can rewrite the additive conditional probability in the intuitively clear form

$$P^{(1)}\big(a_i=\alpha\big|a_{i-N}^{i-1}\big) = p_\alpha + \sum_{r=1}^{N}[P(a_i=\alpha|a_{i-r}) - p_\alpha], \tag{30}$$

which explains the probabilistic meaning of Eq. (28)—in this approximation, each symbol a_{i-r}, $1\leqslant r\leqslant N$, has its own, independent effect on the probability to generate a_i.

Our analysis of Eq. (26) shows that we can neglect the first term on the RHS of Eq. (26) with respect to the term $C_{\alpha\beta\gamma}(r_1,r_2)$ on the LHS because $F_{\gamma\eta}(r_1')$ contains only a nondiagonal small component of $C_{\eta\gamma}(r_1')$. For the same reason, the term $C_{\alpha\beta}(r_1)C_{\varepsilon\eta}(r_2'-r_1')$ is small with respect to $C_{\alpha\beta\varepsilon\eta}(r_1,r_2-r_2',r_2'-r_1')$. The last statement follows from estimation of the correlator $C_{\alpha\beta\varepsilon\eta}(r_1,r_2-r_2',r_2'-r_1')$. Its largest unique component satisfying the conditions $r_2'\geq r_1'+1, r_1'\geq 1$ is $C_{\alpha\beta\alpha\beta}(r_1,-r_1,r_1)$ at $r_1'=r_2, r_2'=r_1+r_2$. Thus, Eq. (26) reduces to

$$C_{\alpha\beta\gamma}(r_1,r_2) = F_{\gamma;\alpha\beta}(r_2,r_1+r_2)C_{\alpha\beta\alpha\beta}(r_1,-r_1,r_1). \tag{31}$$

Taking into account Eq. (29) and neglecting correlations while calculating $C_{\alpha\beta\alpha\beta}(r_1,-r_1,r_1)$, we get

$$F_{\gamma;\alpha\beta}(r_1,r_2) = \frac{1}{p_\alpha p_\beta}C_{\alpha\beta\gamma}(r_2-r_1,r_1). \tag{32}$$

Equation (13) for the conditional probability function of the symbolic high-order Markov chain with a bilinear memory function [in the first approximation with respect to the small parameters $|C_{\alpha\beta}(r)|\ll 1, r\neq 0$ and a multiletter alphabet, $p_\alpha \ll 1$] takes the form

$$\begin{aligned}P^{(2)}\big(a_i=\alpha\big|a_{i-N}^{i-1}\big) &\simeq p_\alpha\\ &+\sum_{\beta\in A}\sum_{r_1=1}^{N}\frac{1}{p_\beta}C_{\beta\alpha}(r_1)[\delta(a_{i-r_1},\beta)-p_\beta]\\ &+\sum_{\beta,\gamma\in\mathcal{A}}\ \sum_{1\leqslant r_1<r_2\leqslant N}\frac{1}{p_\beta p_\gamma}C_{\beta\gamma\alpha}(r_2-r_1,r_1)\\ &\times\{[\delta(a_{i-r_1},\beta)-p_\beta][\delta(a_{i-r_2},\gamma)-p_\gamma]\\ &-C_{\gamma\beta}(r_2-r_1)\}.\end{aligned} \tag{33}$$

C. k-linear form of the CPF

Equations (28) and (32) show that we can hope to obtain similar expressions for generalized memory functions of kth order expressed by means of a correlation function. The result of such calculations can be presented in the form

$$F_{\alpha;\alpha_1,\dots,\alpha_k}(r_1,\dots,r_k) = \frac{1}{p_{\alpha_1},\dots,p_{\alpha_k}} C_{\alpha_k,\dots,\alpha_1\alpha}(r_k - r_{k-1},\dots,r_2 - r_1,r_1). \quad (34)$$

To obtain this result, let us summarize the main steps of this procedure: (i) Calculate the correlation function of $(k+1)$th order, $C_{\alpha_1,\dots,\alpha_{k+1}}(r_1,\dots,r_k)$, $1 \leqslant k \leqslant N$. (ii) While calculating, use $P(.|.) = P^{(N)}(.|.) = \sum_{k=0}^{N} Q^{(k)}(.|.)$. (iii) As a result, the correlation function is presented as a sum of the memory functions of different order from 1 to N with coefficients $C_{\alpha'_1\alpha'_2\cdots}(r'_1,r'_k,\dots)$. (iv) In the main term of the sum containing $\sum_{r',\beta} F_{\alpha_{k+1};\beta_1,\dots,\beta_k}(r'_1,\dots,r'_k) C_{\alpha_1,\dots,\alpha_k\beta_k,\dots,\beta_1}(r_1,\dots,r_{k-1},r_k - r'_k,r'_{k-1} - r'_{k-2},r'_2 - r'_1)$, find the maximal term $C_{\alpha_1,\dots,\beta_1}(r_1,\dots,r'_2 - r'_1)$. It is maximal if in two increasing sequences $0,r_1,\dots,r_{k-1}$ and $r_k - r'_k,r'_{k-1} - r'_{k-2},r'_2 - r'_1$ (rewritten according to the collating procedure as $0,r_1,r_1+r_2,\dots,r_1+r_2+\dots+r_{k-1}$ and $r_1+\dots+r_k - r'_k,r_1+\dots+r_k-r'_{k-1},\dots,r_1+\dots+r_k-r'_1$), there is one-to-one correspondence among their terms: $r'_1 = r_k$, $r'_2 = r_k + r_{k-1}$, and $r'_k = r_k + \dots + r_1$. (v) Neglect correlations while obtaining $C_{\alpha_1,\dots,\beta_1}(r_1,\dots,r'_2 - r'_1) = \prod_{s=1}^{k} p_{\alpha_s}\delta(\alpha_s,\beta_s)$. (vi) All other terms containing $F_{\alpha_{s+1};\beta_1,\dots,\beta_s}(r'_1,\dots,r'_s)$ with $s \neq k$ are small with respect to that taken into account; they contain additional small factors $p^r, r \geq 1$.

Thus, the conditional probability function Eq. (13) for the symbolic high-order Markov chain in the first approximation with respect to the small parameters $|C_{\alpha\beta}(r)| \ll 1, r \neq 0$ and a multiletter alphabet, $p_\alpha \ll 1$, is expressed by means of "experimentally" measured quantities, i.e., the correlation functions. Taking into account property (21), $\sum_{\alpha_m} C_{\alpha_1,\dots,\alpha_k}(r_1,\dots,r_{k-1}) = 0$, it is convenient to present the final main result of the paper in the form of a series [Eq. (16)], where we should substitute Eq. (19) and replace the memory function by Eq. (34),

$$P(a_i = \alpha | a_{i-N}^{i-1}) = \sum_{k=0}^{N} \sum_{\beta_1,\dots,\beta_k \in \mathcal{A}} \sum_{1 \leqslant r_1 < \dots < r_k \leqslant N} \prod_{s=1}^{k} p_{\beta_s}^{-1} C_{\beta_k,\dots,\beta_1\alpha}(r_k - r_{k-1},\dots,r_2 - r_1,r_1) \times \left\{ \prod_{t=1}^{k} [\delta(a_{i-r_t},\beta_t) - p_{\beta_t}] - C_{\beta_k,\dots,\beta_1}(r_k - r_{k-1},\dots,r_k - r_1) \right\}. \quad (35)$$

Equation (35) provides a tool for constructing weak correlated sequences with given, prescribed correlation functions. Note that the i-independence of the function $P(a_i = \alpha|a_{i-N}^{i-1})$ provides homogeneity and stationarity of the sequence under consideration. According to the Markov theorem (see, e.g., Ref. [24]), the finiteness of N together with the strict inequalities

$$0 < P(a_{i+N} = \alpha | a_i^{i+N-1}) < 1,\; i \in \mathbb{N}_+ = \{0,1,2,\dots\} \quad (36)$$

provides ergodicity of the random sequence. Stationarity and ergodicity are sufficient conditions for the formulation of the asymptotic equipartition property (the Shannon-McMillan-Breiman theorem) [14] and Kac's lemma [25,26].

We see that if the correlations are weak, all terms of the CPF are independent of each other. If, e.g., we generate a sequence using the terms of zero order and the bilinear terms, we find that all correlators are equal to zero except third order correlators. In the general case, it is not correct. When, e.g., we generate a sequence with an additive memory function, there appear correlations of all orders, not only pair ones.

D. What to do if p_α are not small

If the real sequence under study does not satisfy the condition $p_\alpha \ll 1$, as, for example, in nucleotide DNA sequences, where all four probabilities p of the different nucleotide occurring are of the order $1/4$, we cannot apply Eq. (35) to obtain the CPF by means of correlation functions. To make this possible, we should decrease the probabilities p. For this purpose, we could use the idea proposed by Jiménez-Montaño, Ebeling, and others [27–29], who suggested coding schemes of the nonsequential recursive pair substitution. Each successive substitution is accompanied by a decrease in the probability $p_{\widetilde{\alpha}}$, where $\widetilde{\alpha}$ belongs to a new extended alphabet.

IV. ARTIFICIAL NEURAL NETWORK

In the previous section, we presented the analytical method of finding the memory functions $F_{\alpha;\beta_1,\dots,\beta_N}(r_1,\dots,r_N)$, and we expressed them in terms of correlation functions. In this section, we expose briefly the numerical method of network training, i.e., estimation of unknown parameters in a network. The result of this procedure should be the values of the matrix functions $F(.)$.

According to the definition given in Ref. [30], artificial neural networks (ANNs) are a family of connectionist models used to estimate or approximate functions [in our case, it is $P(a_i = \alpha|a_{i-N}^{i-1})$] that can depend on a large number of generally unknown parameters. Artificial neural networks are generally presented as systems of interconnected nodes or "neurons." The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning. These definitions correspond to our goal of numerically estimating the unknown function matrices $F_{\alpha;\beta_1,\dots,\beta_N}(r_1,\dots,r_N)$ depending on a great number of parameters, say $N \sim 10^5$ or more.

We mentioned that the equations for the memory functions $F_{\alpha;\beta_1,\dots,\beta_N}(r_1,\dots,r_N)$ can be obtained analytically by means of minimization of the distance (12) (known in the ANN theory as cost function or average system error) between desired and actual neuron output values, the elements of a real referent sequence, and the CPF. The same distance can be used for purposes of numerically finding an unknown quantity—generalized memory matrices $F(.)$—under the network training.

The considered problem with a (potentially) given random sequence falls within the paradigm of supervised learning, which can be thought of as learning with a "teacher." In supervised learning, each example is a pair consisting of an input vector object, a_{i-N}^{i}, and a desired output value, $a_i = \alpha$, or, more precisely, their conditional probabilities, $P(a_i = \alpha | a_{i-N}^{i-1})$. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class of memory functions.

A commonly used mean-square error tries to minimize the average squared error between the network's output and the target value over all the example pairs. When one tries to minimize this cost using gradient descent for the class of neural networks called multilayered perceptrons, one obtains the common and well-known backpropagation algorithm for training neural networks.

A number of supervised learning methods have been introduced in the past two decades. In the paper of Caruana and Niculescu-Mizil [31], the reader can find a large-scale empirical comparison between ten supervised learning methods: SVMs, neural nets, and so on.

V. NUMERICAL SIMULATIONS

In this section, to verify our analytical results, we give examples of numerical generation of random sequences with the state space of dimension $|\mathcal{A}| = 2$ [the symbols (numbers) of the sequence can only take on two values: 0 or 1]. Let us note that for a binary sequence, there is no distinction between symbolic and numeric approaches: all symbolic correlation functions can be expressed by means of numeric ones and vice versa. A more detailed explanation is outlined in the Appendix.

It is supposed that the modeled statistical properties of the random chain are determined by the probability of the symbols occurring, $p_\alpha = 1/2$, the additive part of the memory function:

$$F(r) = \begin{cases} 0.002r, & 1 \leqslant r \leqslant 5, \\ 0.02 - 0.002r, & 5 < r \leqslant 10, \\ 0, & 10 < r, \end{cases} \tag{37}$$

and the exponential (with respect to both arguments) bilinear part of the memory function:

$$F(r_1,r_2) = 0.5\exp(-0.5r_1)\exp(-0.5r_2), \tag{38}$$

with the truncated parameter $N = 20$ playing the role of the memory length.

These memory functions allow us to calculate the conditional probability function (13), consisting in this case of three parts: the zero-order function, $p_\alpha = 1/2$, the first-order additive function, $Q^{(1)}$, and the second-order bilinear function, $Q^{(2)}$, where the last two terms are given by expressions (18) and (22). Using p_α and Eqs. (37) and (38), it is possible to build up four different CPFs. Taking, e.g., the zero-order function only, we obtain an uncorrelated sequence.

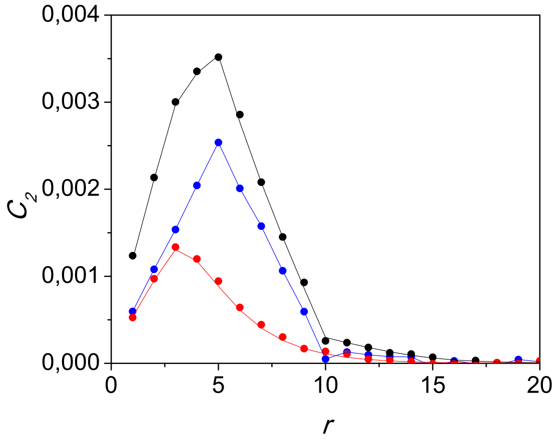

FIG. 1. The pair correlation functions for three random sequences constructed by means of the different conditional probability functions taken in the form of a combination of the linear and bilinear parts of memory functions (37) and (38). The lines are the correlation functions obtained by solving the system of Eqs. (25) and (26). The dots are direct calculations over generated sequences of binary symbols. The length of the sequences is 10^8.

We construct numerically three different random sequences of length 10^8. The first one is generated by the additive probability function $Q^{(1)}$ [Eq. (37)], the second is obtained with the bilinear part [Eq. (38)] (but without the additive part), and the third chain, the most general in our case, is obtained by the CPF containing both terms (37) and (38). All these CPFs contain, evidently, the zero-order term, p_α.

Since the sequences are prepared, we can calculate their correlation functions. In Fig. 1, the obtained correlators are presented by the dots. At the same time, using memory functions Eqs. (37) and (38) and solving iteratively the system of Eqs. (25) and (26) with respect to $C_{\alpha\beta}(r)$ and $C_{\alpha\beta\gamma}(r_1,r_2)$, we obtain the correlation functions presented in Fig. 1 [$C_{\alpha\beta}(r) = C_{11}(r) = C_2(r)$] by the curves.

The middle curve and dots correspond to the sequence generated with the additive memory function. The bottom curve and dots describe a sequence based on the bilinear part of the memory function. The upper curve and dots present C_2 of the sequence obtained with the additive and bilinear memory functions simultaneously. From Eqs. (25) and (26) and numerical simulations, it follows that the correlation functions are entangled and intricate. The additive part of the memory function is responsible for more than just the second-order correlation function, and the bilinear part of the memory function affects more than just the ternary correlation function—they are mixed up in the system (of equations): each memory function affects all correlators. The additive and bilinear parts of the memory function are mainly responsible for the second-order and third-order correlation functions, correspondingly, for a limiting case of small correlations predominantly. Note that the "small" values of the correlation functions C_2 of order of 10^{-3} are really not small. We chose the normalization coefficients of the memory functions (37) and

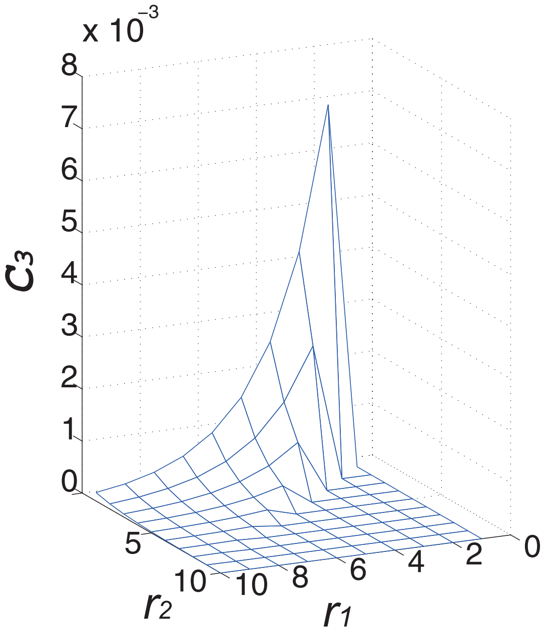

FIG. 2. The third-order correlation function $C(r_1,r_2)$ of the random binary sequence constructed by means of the bilinear exponential memory function (38).

(38) in such a way as to provide extremely strong persistent correlations with the CPF $P^{(2)}(a_i = \alpha|a_{i-N}^{i-1})$ [Eq. (23)].

In the case of a purely additive MF, the form of the correlator is close to the shape of the memory function (middle curve dots in Fig. 1), but it has a significant "tail" at a distance $10 \lesssim r \lesssim 15$. We have already seen this phenomenon of "lag" of a correlator with respect to a memory function in previous studies [18]. In the absence of the additive MF (bottom curve in Fig. 1), the pair additive correlator is not equal to zero, and it has a shape defined by the solution of Eqs. (25) and (26). In the case of simultaneous generation with additive and bilinear MFs, the pair correlator becomes significantly different from that in the previous cases.

The two-dimensional surface for the third-order correlator $C(r_1,r_2)$, obtained by a solution of Eqs. (25) and (26) with the bilinear part of the memory function [Eq. (38)] only, is shown in Fig. 2. A comparison of this function (line) and the numerically found third-order correlation function $C(r_1,r_2)$ (dots) of the binary sequence for the fixed coordinate $r_1 = 1$ is presented in Fig. 3. A good agreement between the analytical and numerical calculations of the correlation functions in all cases gives us a reason to believe that the system of equations and the generation procedure based on the bilinear form of the CPF are well grounded.

VI. CONCLUSION

We obtained the decomposition procedure for the CPF of symbolic random sequences [Eq. (16)], which were defined as high-order Markov chains. We represented the conditional probability function as the sum of multilinear memory function monomials of different orders [Eq. (19)]. This allowed us to build artificial sequences via a method of successive iterations, taking into account at each step increasingly high correlations among random elements. At weak correlations, the memory functions are expressed analytically in terms of high-order symbolic correlation functions [Eq. (35)]. Thus we have filled the theoretical gap between the methods of the additive Markov chain and the likelihood estimation.

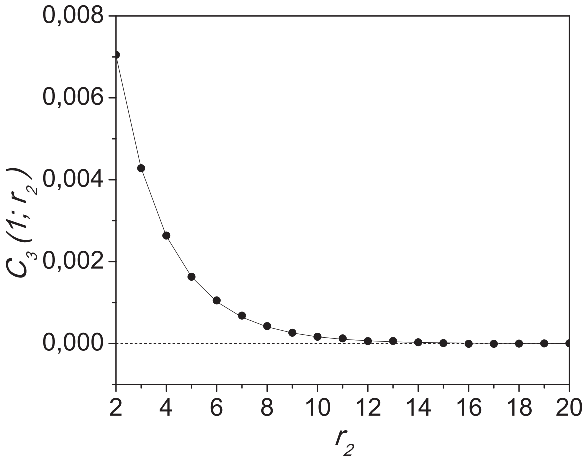

FIG. 3. A comparison of the third-order correlation functions $C(r_1,r_2)$ of the binary sequence [taking into account the additive (37) and bilinear (38) memory functions] prescribed by the memory function (line) and numerically found (dots) for the fixed coordinate $r_1 = 1$.

In this paper, we have considered the *microscopic* characteristics of random symbolic systems—the correlation functions. The conditional entropy is the most important *macroscopic* characteristic of the sequence. Our preliminary numerical simulations show that the analytical result obtained in this paper can be used successfully for numerical evaluation of the entropy of DNA nucleotide sequences and sequences obtained by dichotomization of a logistic map. We have seen that third-order correlations (in the framework of a bilinear Markov chain) can essentially lower the entropy calculated in the framework of the naive likelihood estimation and the additive Markov chain approach.

APPENDIX

Binary random sequences can always be considered to be numeric. Therefore, Eqs. (25) and (26) can be simplified considerably to a numeric form. Let us consider the symbolic correlator (5) taken at $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 1$. In this case, it is possible to set $\delta(a_r,\alpha) = a_r$, whereupon the symbolic correlation function takes on a form that is identical to the numeric one,

$$\begin{aligned} &C_{1,\dots,1}(r_1,r_2,\dots,r_{k-1}) \\ &= \overline{(a_0 - p_1)\cdots(a_{r_1+\dots+r_{k-1}} - p_1)} \\ &\equiv C(r_1,r_2,\dots,r_{k-1}). \end{aligned} \tag{A1}$$

To calculate a symbolic correlator with arbitrary indexes, it is sufficient to note that the term $\delta(a_r,\alpha) - p_\alpha$ changes its sign when one replaces 1 with 0 or 0 with 1, $\delta(a_r,\alpha) - p_\alpha = (1 - a_r) - p_0 = -(a_r - p_1)$. Each such replacement changes the sign of the correlator. Consequently, its value depends on the number q of symbols “0” among $\alpha_1,\dots,\alpha_k$,

$$C_{\alpha_1,\dots,\alpha_k}(r_1,r_2,\dots,r_{k-1}) = (-1)^q C(r_1,r_2,\dots,r_{k-1}). \tag{A2}$$

Using the same reasoning, we find a simplified expression for the numeric CPF,

$$P(a_i = 1|\cdots) = p_1 + \sum_{r=1}^{N} F(r)(a_{i-r} - p_1), \tag{A3}$$

where

$$F(r) = F_{11}(r) - F_{10}(r). \tag{A4}$$

Let us rewrite Eq. (25) for the element $C_{11}(r)$. We set $\alpha = \beta = 1$ and take a sum over $\gamma = \{0,1\}$. Then, on the LHS we obtain the numeric correlator $C(r)$; on the RHS, we express all symbolic correlators by means of numeric ones [Eq. (A2)]. After that, all symbolic memory functions turn out to be grouped in combinations, which result in numeric memory functions, e.g.,

$$\begin{aligned}&C_{10}(r-r_1)F_{1;0}(r_1) + C_{11}(r-r_1)F_{1;1}(r_1)\\ &\quad = C(r-r_1)[F_{1;1}(r_1) - F_{1;0}(r_1)] = C(r-r_1)F(r_1).\end{aligned} \tag{A5}$$

In the same way, we manipulate using the second term in Eq. (25) and introduce the second-order numeric memory function,

$$\begin{aligned}F(r_1,r_2) = {}& F_{100}(r_1,r_2) - F_{101}(r_1,r_2)\\ &- F_{110}(r_1,r_2) + F_{111}(r_1,r_2).\end{aligned} \tag{A6}$$

As a result, Eqs. (25) and (26) take on the following simplified forms, which are convenient for describing binary sequences:

$$\begin{aligned}C(r) = {}& \sum_{r_1=1}^{N} C(r-r_1)F(r_1)\\ &+ \sum_{r_1=1}^{N-1}\sum_{r_2=r_1+1}^{N} C(r-r_2,r_2-r_1)F(r_1,r_2),\end{aligned} \tag{A7}$$

$$\begin{aligned}C(r_1,r_2) = {}& \sum_{r_1'=1}^{N} C(r_1,r_2-r_1')F(r_1')\\ &+ \sum_{r_1'=1}^{N-1}\sum_{r_2'=r_1'+1}^{N} F(r_1',r_2')[C(r_1,r_2-r_2',r_2'-r_1')\\ &- C(r_1)C(r_2'-r_1')].\end{aligned} \tag{A8}$$

[1] D. Salomon, *A Concise Introduction to Data Compression* (Springer, Berlin, 2008).

[2] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval* (Cambridge University Press, Cambridge, 2008).

[3] A. D. Wissner-Gross and C. E. Freer, Phys. Rev. Lett. **110**, 168702 (2013).

[4] F. M. Izrailev, A. A. Krokhin, and N. M. Makarov, Phys. Rep. **512**, 125 (2012).

[5] P. Ehrenfest and T. Ehrenfest, *Encyklopädie der Mathematischen Wissenschaften* (Springer, Berlin, 1911), p. 742, Bd. II.

[6] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding* (Cambridge University Press, Cambridge, 1995).

[7] H. Kantz and T. Schreiber, *Nonlinear Time Series* (Cambridge University Press, Cambridge, 1997).

[8] R. Hosseinia, N. Leb, and J. Zideka, J. Stat. Theor. Practice **5**, 261 (2011).

[9] A. Raftery, J. R. Stat. Soc. B **47**, 528 (1985).

[10] M. Seifert, A. Gohr, M. Strickert, and I. Grosse, PLoS Computat. Biol. **8**, e1002286 (2012).

[11] P. C. Shields, *The Ergodic Theory of Discrete Sample Paths*, Graduate Studies in Mathematics Vol. 13 (American Mathematical Society, Providence, RI, 1996).

[12] O. V. Usatenko, S. S. Apostolov, Z. A. Mayzelis, and S. S. Melnik, *Random Finite-Valued Dynamical Systems: Additive Markov Chain Approach* (Cambridge Scientific, Cambridge, 2010).

[13] O. V. Usatenko and V. A. Yampol’skii, Phys. Rev. Lett. **90**, 110601 (2003).

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley, New York, 2006).

[15] https://math.dartmouth.edu/doyle/docs/random/random.pdf

[16] V. A. Uspensky and A. Shen, Math. Syst. Theor. **29**, 271 (1996).

[17] S. S. Melnyk, O. V. Usatenko, V. A. Yampol’skii, and V. A. Golick, Phys. Rev. E **72**, 026140 (2005).

[18] S. S. Melnyk, O. V. Usatenko, and V. A. Yampol’skii, Physica A **361**, 405 (2006).

[19] S. S. Melnik and O. V. Usatenko, Computat. Biol. Chem. **53**, 26 (2014).

[20] S. S. Melnik and O. V. Usatenko, Phys. Rev. E **93**, 062144 (2016).

[21] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, EURASIP J. Appl. Signal Process. **1**, 13 (2004).

[22] S. S. Apostolov, Z. A. Mayzelis, O. V. Usatenko, and V. A. Yampol’skii, Int. J. Mod. Phys. B **22**, 3841 (2008).

[23] J. Besag, J. R. Stat. Soc. B **36**, 192 (1974).

[24] A. N. Shiryaev, *Probability* (Springer, New York, 1996).

[25] M. Kac, Bull. Am. Math. Soc. **53**, 1002 (1947).

[26] F. Cecconi, M. Cencini, M. Falcioni, and A. Vulpiani, Am. J. Phys. **80**, 1001 (2012).

[27] W. Ebeling and M. A. Jiménez-Montaño, Math. Biosci. **52**, 53 (1980).

[28] M. A. Jiménez-Montaño, Bull. Math. Biol. **46**, 641 (1984).

[29] P. E. Rapp, I. D. Zimmermann, E. P. Vining, N. Cohen, A. M. Albano, and M. A. Jiménez-Montaño, Phys. Lett. A **192**, 27 (1994).

[30] S. Haykin, *Neural Networks: A Comprehensive Foundation* (Prentice-Hall, Englewood Cliffs, NJ, 1998).

[31] R. Caruana and A. Niculescu-Mizil, *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, edited by A. Nicholson and P. Smyth (AUAI Press, Corvallis, OR, 2013).