Information temperature and macroscopic descriptors of additive *N*th-order Markov chains

O. V. Usatenko,[1,2] G. M. Pritula,[1] and S. S. Melnyk[1]

[1]*O. Ya. Usikov Institute for Radiophysics and Electronics, Ukrainian Academy of Science, 12 Proskura Street, 61805 Kharkiv, Ukraine*

[2]*Center for Theoretical Physics, Polish Academy of Sciences, Aleja Lotników 32/46, 02-668 Warsaw, Poland*

(Received 30 March 2026; accepted 17 June 2026; published 10 July 2026)

Large-scale language models (LLMs) operate in extremely high-dimensional state spaces, where both token embeddings and their hidden representations create complex dependences that are not easily reduced to classical Markov structures. In this paper we explore a theoretically solvable approximation of LLM dynamics using *N*th-order additive Markov chains. Such models allow the conditional probability of the next token to be decomposed into a superposition of contributions from multiple historical depths, reducing the combinatorial explosion typically associated with high-order Markov processes. The main result of the work is the establishment of a correspondence between an additive multistep chain and a chain with a stepwise memory function. This correspondence allows the introduction of the concept of information temperature not only for stepwise but also for additive *N*th-order Markov chains.

DOI: 10.1103/tdml-7tk4

I. INTRODUCTION

Large-scale language models (LLMs) have become the dominant paradigm in contemporary natural language processing [1], providing state-of-the-art performance across a wide range of linguistic and reasoning tasks. Despite their empirical success, the internal statistical structure of LLMs remains only partially understood. In particular, it is still unclear which classes of stochastic processes best approximate the sequence distributions generated by LLMs and how these processes relate to classical probabilistic models of text. Developing a mathematically transparent framework that connects LLMs with established sequence models is therefore a pressing and foundational problem (see, in this connection, Refs. [2,3], where this point of view is most clearly expressed).

A natural starting point for such an analysis is provided by Markov chains. Classical *N*th-order Markov chains allow one to describe correlations extending over long histories but suffer from an exponential growth of parameters with increasing memory depth. This rapid growth represents a canonical manifestation of the curse of dimensionality in sequence modeling, rendering high-order Markov models impractical for analytical or statistical purposes. Additive *N*th-order Markov chains offer a principled reduction of this complexity by decomposing the conditional probability of the next symbol into a superposition of contributions from different delay positions. As a result, the number of parameters grows only linearly with the memory length, allowing long-range correlations to be represented in a compact and analytically tractable form.

The curse of dimensionality also plays a central role in modern machine learning systems, including LLMs, where extremely-high-dimensional embeddings and large parameter spaces coexist with tractable generation and sampling procedures. While LLMs are not Markovian models in a strict probabilistic sense, their output sequences exhibit strong correlations across multiple temporal scales. This motivates the study of simplified stochastic models in which the role of long memory and structural constraints can be analyzed explicitly, without attempting to reproduce the full architectural complexity of neural language models.

A central insight of statistical physics is that systems with many microscopic degrees of freedom can admit an effective description in terms of a small set of macroscopic parameters. For stochastic symbolic processes, such reduced descriptions may emerge from the correlation structure and entropy of the generated sequences, providing compact measures of their complexity. Large-scale language models naturally produce such stochastic sequences, and their output is typically controlled by a sampling temperature that modulates the entropy of the softmax distribution. While this parameter is widely used in practice, its status as an intrinsic characteristic of the underlying generative process rather than an external control remains conceptually unresolved.

In the present work we further develop the concept of information temperature, first introduced in [4] for binary Markov chains of *N*th-order with stepwise memory and shown in [5] to quantify the complexity of random symbolic sequences. We establish a correspondence between additive *N*th-order Markov chains and chains with a stepwise memory function, thereby extending the definition of information temperature to additive *N*th-order Markov models. In this framework, temperature emerges as a macroscopic parameter determined by the correlation structure of the process and characterizes the balance between entropy and informational constraints imposed by the prescribed correlation structure.

While our analysis is carried out for binary sequences, this restriction isolates the essential role of memory and additivity and enables exact analytical results. The resulting framework provides a controlled setting in which the emergence of macroscopic parameters in correlated symbolic systems can be studied. From this perspective, the temperature parameter used in LLM sampling may be viewed as an operational analog of information temperature, acting on an implicit, learned logit-based energy landscape rather than on explicitly defined stochastic correlations. This interpretation does not assert an equivalence between LLMs and additive Markov chains but suggests a principled connection at the level of macroscopic statistical descriptors.

The structure of the paper is as follows. Section II introduces the key concepts and definitions required for the subsequent derivation and discussion of the main results. Section III presents the central result of this work, namely, the establishment of a correspondence between an additive multistep Markov chain and a chain with a stepwise memory function. In Sec. IV we summarize the basic results concerning information temperature. Owing to the established equivalence between the two classes of chains, the concept of temperature is extended from stepwise to additive N-step Markov chains. Section V illustrates the obtained analytical results. Sections VI and VII conclude the paper with a summary of the results and a discussion of perspectives for future research, respectively.

II. BACKGROUND: BASIC DEFINITIONS AND CONCEPTS

In this section we recall the background information about two higher-order Markov chain models, i.e., the stepwise Markov chain and the additive Nth-order Markov chain, required for the subsequent analysis.

A. Nth-order Markov chains and their models

The notion of information temperature arises naturally in the context of a coarse-grained description of dynamical systems. In particular, a finite-alphabet symbolic system can be reduced to a binary representation, which allows for a simpler mathematical treatment. The resulting binary sequence is then modeled as an Nth-order Markov chain.

Consider an infinite random stationary ergodic sequence $\mathbb{S}$ of symbols a_i,

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

taken from the binary alphabet $a_i \in \mathcal{A}$:

$$\mathcal{A} = \{0, 1\}, \quad i \in \mathbb{Z} = \{\dots, -1, 0, 1, 2, \dots\}. \tag{2}$$

The sequence $\mathbb{S}$ is said to be an Nth-order Markov chain if the probability that symbol a_i takes a certain value $a \in \mathcal{A}$, given that all previous symbols are fixed, depends only on the N preceding symbols,

$$P(a_i = a| \dots, a_{i-2}, a_{i-1}) = P(a_i = a|a_{i-N}, \dots, a_{i-2}, a_{i-1}). \tag{3}$$

To ensure irreducibility and ergodicity, the conditional probability distribution function (CPDF) $P(\cdot|\cdots)$ of the chain has to satisfy the condition

$$0 < P\big(a_i = a|a_{i-N}^{i-1}\big) < 1, \tag{4}$$

where we use the concise notation $a_{i-N}^{i-1} = a_{i-N}, \dots, a_{i-1}$.

1. Stepwise Nth-order Markov chain

The statistical properties of a higher-order binary Markov chain can be described using a coarse-grained approximation: the Markov chain with a stepwise memory function, first introduced in [6]. Within this model, the conditional probability of observing the next symbol is independent of the ordering of the preceding N symbols and depends only on the number k of 1's among them,

$$\begin{aligned} P\big(a_i = a|a_{i-N}^{i-1}\big) &= P(a|k) \\ &= p(a) + \mu(2a-1)\left(\frac{2k}{N} - 1\right), \end{aligned} \tag{5}$$

where N is the order of the chain and μ is the correlation parameter $-\frac{1}{2} < \mu < \frac{1}{2}$.

The stepwise chain (5) with

$$p(a) = \tfrac{1}{2} + (2a-1)\nu \tag{6}$$

is usually referred to as a biased chain with bias parameter ν (see, e.g., [7]). In such chains the term ν describes the difference between the average frequencies of symbols 0 and 1.

Chains with stepwise memory belong to the class of so-called permutative Markov chains [8], in which the transition probability is independent of the ordering of symbols in the preceding Nth-order word. Remarkably, this simple model captures certain statistical features observed in literary texts and DNA sequences.

2. Additive Nth-order Markov chain

A particular form of the CPDF $P(a_i|a_{i-N}^{i-1})$, for a binary N-step Markov chain, is given by [9]

$$P\big(a_i = 1|a_{i-N}^{i-1}\big) = \overline{a} + \sum_{r=1}^{N} F(r)(a_{i-r} - \overline{a}). \tag{7}$$

Sequences satisfying this relation are referred to as additive Markov chains and $F(r)$ is called the memory function. It characterizes the strength of the influence of the previous symbol a_{i-r} $(1 \leqslant r \leqslant N)$ on the generated symbol a_i.

The relation between the correlation function

$$K(r - r') = \overline{a_{i-r}a_{i-r'}} - \overline{a}^2 \tag{8}$$

and the memory function $F(r)$,

$$K(r) = \sum_{r'=1}^{N} F(r')K(r - r'), \quad r \geqslant 1, \tag{9}$$

makes it possible to determine the memory function by solving this equation and consequently to construct a binary sequence with a prescribed correlation function. The overline in (8) denotes averaging; due to ergodicity [Eq. (4)], it is immaterial whether the average is taken over the ensemble or along the chain.

B. Chapman-Kolmogorov equation

For a binary stationary ergodic Markov chain, the probability $P(a_1a_2,\ldots,a_N)$ of occurrence of a given word $(a_1,a_2,\ldots,a_N)$ satisfies the compatibility condition given by the Chapman-Kolmogorov equation (see, for example, Ref. [10])

$$P(a_1,\ldots,a_N)=\sum_{a=0,1}P(aa_1,\ldots,a_{N-1})\times P(a_N|a,a_1,\ldots,a_{N-1}). \quad (10)$$

C. Equivalence between Markov and two-sided random chains

For a two-sided random chain equivalent to a binary Nth-order Markov chain, the conditional probability that the symbol a_i equals unity, under the condition that the rest of the symbols in the chain are fixed, can be written in the form [11]

$$P(a_i=1|A_i^-,A_i^+)=\frac{P(a_i=1,A_i^+|A_i^-)}{P(a_i=1,A_i^+|A_i^-)+P(a_i=0,A_i^+|A_i^-)}, \quad (11)$$

where $A_i^-=(a_{i-N},\ldots,a_{i-2},a_{i-1})$ and $A_i^+=(a_{i+1},a_{i+2},\ldots,a_{i+N})$ denote the previous and subsequent words of length N with respect to the symbol a_i. Here the two-sided conditional probability $P(a_i=1|A_i^-,A_i^+)$ is expressed in terms of the Markov-like probability functions $P(a_i,A_i^+|A_i^-)$.

D. Entropy rate

The Shannon entropy of a subsequence of symbols of length L (see, e.g., Ref. [12]) is

$$H_L=-\sum_{a_1,\ldots,a_L\in\mathcal{A}}P(a_1,\ldots,a_L)\ln P(a_1,\ldots,a_L). \quad (12)$$

Here $P(a_1,\ldots,a_L)$ is the probability of observing the length-L word $(a_1,\ldots,a_L)$ in the sequence. Equation (12) allows one to introduce the entropy rate (or conditional entropy), i.e., the entropy per symbol,

$$h_L=H_{L+1}-H_L. \quad (13)$$

The source entropy is the entropy per symbol in the asymptotic limit $\lim_{L\to\infty}h_L$. It measures the average information per symbol when all correlations are taken into account.

III. MACROSCOPIC PARAMETERS OF MARKOV CHAINS

In this section we establish the correspondence between stepwise and additive Markov chains.

Equivalence between additive and stepwise Markov chain models

To introduce the notion of the temperature for the random binary additive chain with the CPDF (7), first let us rewrite it in the more general form, representing the probability of generating the symbol a, after the word a_{i-N}^{i-1} as

$$P_{\rm ad}\left(a_i=a|a_{i-N}^{i-1}\right)=p_a+(2a-1)\sum_{r=1}^{N}F(r)(a_{i-r}-\overline{a}), \quad (14)$$

where $p_1=\overline{a}$ and $p_0=1-\overline{a}$.

We want to establish its correspondence with the biased stepwise chain characterized by the CPDF (5) with (6),

$$P_{\rm sw}(a_i=a|k)=p_{\nu,a}+(2a-1)\mu\left(\frac{2k}{N}-1\right), \quad (15)$$

$$p_{\nu,0}=\frac{1}{2}+\nu,\quad p_{\nu,1}=\frac{1}{2}-\nu,\quad k=\sum_{r=1}^{N}a_{i-r}, \quad (16)$$

for which the temperature can be introduced as discussed below Eq. (37). This correspondence can be formulated as a minimization of the distance $\mathcal{D}$,

$$\mathcal{D}=\sum_{a=\{0,1\}}\overline{\left[P_{\rm sw}(a_i=a|k)-P_{\rm ad}\left(a_i=a|a_{i-N}^{i-1}\right)\right]^2}, \quad (17)$$

between the conditional probabilities of two chains from which we should find the parameters ν and μ of a sequence with a stepwise memory function. The averaging in Eq. (17) is performed over the stationary distribution of the initially given additive chain determined by Eq. (14).

The parameters of the CPDF in (14), i.e., $p_0=1-p_1$, $p_1=\overline{a}$, and $F(r)$, are supposed to be known and the parameters μ and ν in (14) should be determined from the minimization equations. In the absence of correlation, such a connection is obvious:

$$\mu=F=0\Rightarrow p_{\nu,1}=\tfrac{1}{2}-\nu=p_1=\overline{a}. \quad (18)$$

If the memory function in Eq. (14) is constant, $F(r)=F_0=\text{const}$, then the additive Markov chain reduces to a Markov chain (15), with a stepwise memory function $F_0=2\mu/N$,

$$F(r)=F_0\Rightarrow\mu=NF_0/2. \quad (19)$$

The result of $\mathcal{D}$ calculations is

$$\mathcal{D}=2g_0^2+2\sum_{r,r'=1}^{N}g_rg_{r'}K(r-r'), \quad (20)$$

where

$$g_0=\left(\overline{a}-\tfrac{1}{2}\right)(1-2\mu)+\nu,\quad g_r=F_r-\frac{2\mu}{N}. \quad (21)$$

Using Eq. (20), we obtain from minimization equations the following results for the parameters μ and ν:

$$\frac{\partial\mathcal{D}}{\partial\mu}=0\Rightarrow\mu=\frac{1}{2}\frac{\langle K\star F\rangle}{\langle\langle K\rangle\rangle}, \quad (22)$$

$$\frac{\partial\mathcal{D}}{\partial\nu}=0\Rightarrow\nu=\tfrac{1}{2}(1-2\overline{a})(1-2\mu). \quad (23)$$

Here

$$\langle K\star F\rangle=\frac{1}{N}\sum_{r,r'=1}^{N}K(r-r')F(r'), \quad (24)$$

$$\langle\langle K\rangle\rangle=\frac{1}{N^2}\sum_{r,r'=1}^{N}K(r-r'). \quad (25)$$

Thus, Eqs. (22) and (23) define the parameters ν and μ of a Markov chain with a stepwise memory function. They are expressed in terms of the microscopic parameters of the additive Markov chain. The parameter ν has a nearly trivial meaning; it can be determined from a simple comparison of

the one-point distribution functions contained in the CPDFs (14) and (15). Comparing two results of the $\bar{a}$ calculation with Eq. (14),

$$\bar{a} = p_a = \overline{P_{\rm ad}(a_i = 1|a_{i-N}^{i-1})},$$

and with Eq. (15),

$$\bar{a} = \overline{P_{\rm sw}(a_i = 1|k)} = \tfrac{1}{2} - \nu + \mu(2\bar{a} - 1),$$

we find ν [Eq. (23)].

In a certain sense, the parameter μ is a more important parameter than ν, since it determines the correlation property of the two chains. This is what we will discuss in more detail below.

As expected, in the absence of memory $F(r) = 0$, we have, from Eq. (22), $\mu = 0$.

Let us note another useful form of the parameter μ presentation with using the relation (9),

$$\mu = \frac{1}{2}\frac{\langle K\rangle}{\langle\langle K\rangle\rangle}, \quad \langle K\rangle = \frac{1}{N}\sum_{r=1}^{N} K(r). \tag{26}$$

Despite the apparent similarity of the expressions for $\langle K\rangle$ and $\langle\langle K\rangle\rangle$ in the numerator and denominator of Eq. (26), the sets of correlation functions in them are different; in the numerator, the function represents a set of $K(r)$ with $r = 1, 2, \ldots, N$, and in the denominator, this is a set of $K(r)$ with $r = 0, 1, \ldots, N-1$, which provides the important inequality $|\mu| < \frac{1}{2}$ due to the property $|\langle\langle K\rangle\rangle| > |\langle K\rangle|$.

Another form of Eq. (22), expressed in terms of the variance of the random variable k,

$$D(N) = \overline{(k - \bar{k})^2} = \sum_{r,r'=1}^{N} K(r - r'), \tag{27}$$

is

$$\mu = \frac{N^2}{2}\frac{\langle K\rangle}{D(N)}. \tag{28}$$

In all the expressions presented above, $K(r)$ denotes the correlation function of the additive process that we approximate. However, if we assume that the additive and stepwise processes are similar in their correlation properties, then the correlation function of the step process can be used as $K(r)$. Then we can use asymptotic expressions for $K(r)$ and $D(N)$ calculated for the stepwise chain in Refs. [6,13],

$$D(N) = \begin{cases} \frac{N}{4(1-2\mu)}, & n \gg 1 \\ \frac{N^2}{4} - \frac{nN(N-1)}{2}, & n \ll 1. \end{cases} \tag{29}$$

Here the parameter n, defined by

$$n = \frac{N(1-2\mu)}{4\mu}, \tag{30}$$

determines and separates the strength of persistent ($\mu > 0$) correlation from weak ($n \gg 1$ and $\mu \to 0$) to strong ($n \ll 1$ and $\mu \to \frac{1}{2}$).

In the case of weak memory and correlation $|F(r)| \ll 1$ and $|\langle K\rangle| \ll 1$, we can use $D(N) \approx N/4$. Then

$$\mu \approx 2N\langle K\rangle, \quad |\mu| \ll 1. \tag{31}$$

In the opposite case of the maximal persistence of correlations, the correlator $K(r) \to \frac{1}{4}$, $\langle K\rangle \to \frac{1}{4}$, and the variance $D(N) \to N^2/4$. Then it follows from (28) that $\mu \to \frac{1}{2}$. Similarly, for the limiting antipersistent correlations, we have $\langle K\rangle \to -\frac{1}{4}$ and $D(N) \to N^2/4$; therefore, we get $\mu \to -\frac{1}{2}$.

IV. BASIC RESULTS ON INFORMATION TEMPERATURE

One of the methods for introducing information temperature, which we call the method of equivalent correspondences (ECs), is based on the fact that every binary Markov chain is equivalent to some binary two-sided chain [11] which in turn can be viewed as the Ising sequence for which the probability of a configuration is given by the Boltzmann distribution explicitly containing the temperature T.

The second method makes use of the traditional entropy-based thermodynamic definition of temperature with direct calculation of the block entropy and some fictive energy of the Markov chain. This method does not suppose a mapping of random sequence on the Boltzmann exponential distribution and describes a broader class of random sequences as compared to the previous EC method.

Both approaches yield identical results in the case of nearest-neighbor spin or symbol interactions; however, the method based on the correspondence between Markov and Ising chains becomes increasingly cumbersome for chain orders $N \geqslant 3$. While the first approach to introducing temperature appears more natural, the second may be viewed as heuristic or axiomatic. We present both methods to demonstrate the consistency of the results and to justify the use of auxiliary fictive energy in the second approach.

In both cases, information temperature is understood as an effective thermodynamic quantity that measures the degree of correlation or ordering in a sequence.

Keeping in mind the role of the temperature parameter in LLMs, it is natural to consider the possibility of introducing the information temperature of Markov chains within a maximum-entropy framework, in analogy with the softmax construction. This possibility is discussed in the Appendix.

A. Temperature from equivalence of the Ising and Markov chains

Order $N = 1$. Using the definition (5) for the CPDF of the ordinary one-step Markov chain, $N = 1$, considering the equivalence of Markov chain to the two-sided random sequence (11), and comparing them with the Boltzmann distribution for the corresponding Ising model, we get the information temperature as a function of the parameter μ,

$$\mu = \frac{1}{2}\tanh\left(\frac{1}{\tau}\right), \quad \frac{1}{\tau} = \frac{\varepsilon}{T} = \frac{1}{2}\ln\frac{1+2\mu}{1-2\mu}. \tag{32}$$

In the calculations we supposed that the binary variables $a_i = \{0, 1\}$ of the Markov chain are related to the Ising spin variables $s_i = \{-1, 1\} = \{\downarrow, \uparrow\}$ by the equality $s_i = 2a_i - 1$, and the energy of the spins interaction is of the form $\varepsilon_{\uparrow\uparrow} = \varepsilon_{\downarrow\downarrow} = -\varepsilon$ and $\varepsilon_{\uparrow\downarrow} = \varepsilon_{\downarrow\uparrow} = \varepsilon > 0$. The result (32) corresponds to our intuitive understanding of the notion of temperature. In fact, the definition (5) at $\mu \to 0$ describes disordered sequences and $\mu \to \frac{1}{2}$ describes strongly

correlated chains. So for $\tau \to \pm\infty$ we have $\varepsilon/T \simeq 2\mu \to 0$, and $\tau \to \pm 0$ when $\mu \to \pm\frac{1}{2}$. The negative value of τ describes an antiferromagnetic ordering of spins or symbols 0 and 1. We can say that τ is the temperature T measured in units of energy ε, $\tau = T/\varepsilon$.

B. Temperature from the entropy

The second method for introducing the information temperature is based on the thermodynamic definition of temperature in terms of entropy. In this approach the temperature is obtained from the dependence of the block entropy on the average energy of the symbolic sequence. The energy of the random chain is introduced within the pair interaction approximation.

Importantly, this entropy-based construction of information temperature is conceptually independent of the method of equivalent correspondences discussed above. Whereas the equivalent correspondence approach introduces temperature through a mapping of the symbolic sequence onto an effective Boltzmann distribution, the present method derives it directly from the statistical properties of the sequence via the relation between block entropy and the average interaction energy.

Within this analogy, a symbolic sequence is interpreted as a thermodynamic system. Each binary symbol $a_i = \{0, 1\}$ corresponds to a particle state; a finite block of symbols $\{a_i, \ldots, a_{i+L-1}\}$ may then be viewed as a microscopic state, interacting with the remaining part of the chain (a thermostat).

For an Nth-order Markov chain, the probability of the next symbol depends only on the N preceding symbols. Consequently, the statistical structure of the process is completely determined by the probabilities of blocks of length $N+1$. These $(N+1)$-length words can therefore be regarded as the microscopic states of the subsystem. In the stationary regime, the distribution of such blocks is invariant along the sequence, allowing the subsystem to be treated, within the thermodynamic analogy, as being in statistical equilibrium with its environment.

To introduce the temperature, we associate the auxiliary pair interaction energies ε_r with the symbols a_i and a_{i+r}. These energies do not affect the statistics of the sequence and serve only as a formal tool for defining thermodynamic quantities. The entropy of the subsystem is characterized by the Shannon block entropy (12), while the average energy is calculated from the assumed pair interactions. The temperature is then defined in the usual thermodynamic way as the derivative of entropy with respect to the average energy.

For simplicity, we restrict the analysis to the unbiased chain, Eq. (5) with $p(a) = \frac{1}{2}$. Details of the calculations can be found in Refs. [5,6]; here we present only the final results. The probabilities of the $N+1$ blocks are determined from the Chapman-Kolmogorov equation (10).

Order $N = 2$. Using Eqs. (10) and (5) for $N = 2$, we determine the probabilities of two-symbol words and the corresponding three-point probabilities of symbols 0 and 1. Introducing auxiliary interaction energies ε_1 and ε_2 for nearest- and next-to-nearest-neighbor symbol pairs, we calculate the average energy and the entropy per two bonds of the random chain. The temperature obtained from the thermodynamic relation between entropy and energy is

$$\frac{1}{\tau} = \frac{1}{4} \ln \frac{1+2\mu}{1-2\mu}. \tag{33}$$

Here

$$\frac{1}{\tau} = \frac{\langle\varepsilon\rangle}{T}, \quad \langle\varepsilon\rangle = \frac{2\,\varepsilon_1 + \varepsilon_2}{3}. \tag{34}$$

The quantity $\langle\varepsilon\rangle$ arises naturally in the calculation and can be interpreted as an effective average interaction energy. It also sets the natural unit for measuring the temperature. Choosing $\varepsilon_1 = \varepsilon_2 = 1$ yields the natural unit of information temperature, $\tau = T$. Remarkably, the final result (33) does not depend on the particular values of the auxiliary interaction energies.

Order $N = 3$. Applying the same entropy-based procedure to a random sequence with memory length $N = 3$ leads to a more cumbersome expression for the temperature. Its asymptotic behavior is

$$\lim_{\mu\to 0}\left(\frac{1}{\tau}\right) \simeq \frac{2\mu}{3}, \quad \lim_{\mu\to\pm 1/2}\left(\frac{1}{\tau}\right) \propto \ln\frac{1+2\mu}{1-2\mu}. \tag{35}$$

Order $N \gg 1$. For the stepwise chain in the high-temperature limit (large N or small μ), it was shown in Ref. [4] that the information temperature takes the simple form

$$\frac{1}{\tau} = \frac{2\mu}{N}. \tag{36}$$

C. Ansatz

Considering the obtained relations, one can propose the generalized expression

$$\frac{1}{\tau} = \frac{1}{2N} \ln \frac{1+2\mu}{1-2\mu}. \tag{37}$$

A comparison with Eqs. (32), (33), (35), and (36) shows that this expression reproduces Eq. (32) for $N = 1$ and Eq. (33) for $N = 2$. For $N = 3$, it agrees asymptotically with Eq. (35) in the limits $\mu \to 0$ and $\mu \to \pm\frac{1}{2}$, and it also reproduces the asymptotic behavior given by Eq. (36) as $\mu \to 0$ for large N.

These observations justify using Eq. (37) as a definition of the temperature of a random Nth-order Markov chain with the stepwise CPDF (5).

D. Temperature of the additive chain

Thus, the analysis carried out shows that the parameters μ and ν satisfy all the necessary properties and are uniquely determined by the parameters $p_{\nu,a}$ and $F(r)$ of the additive Nth-order Markov chain. Therefore, Eq. (37) can be considered as the information temperature of the additive Nth-order Markov chain.

The reduction of the Nth-order additive Markov chain to its stepwise representation can be viewed as an information-theoretic analog of statistical averaging in thermodynamics. In statistical physics, microscopic variables describing individual particle states are replaced by their ensemble averages, leading to macroscopic quantities such as temperature and pressure. Similarly, in the coarse-grained approximation of

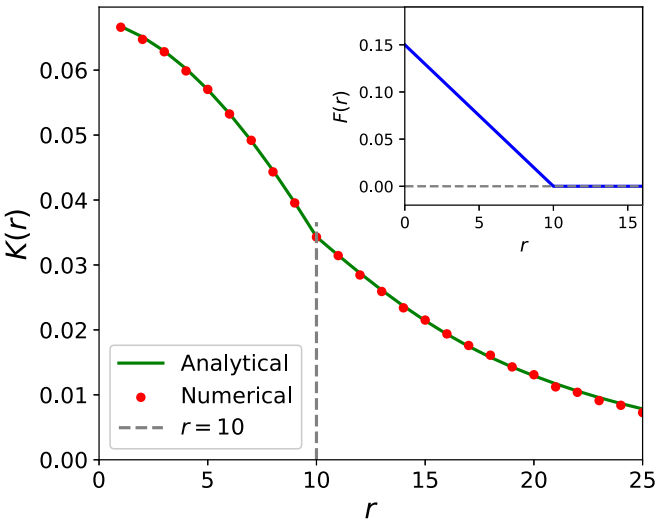

FIG. 1. Correlation function $K(r)$ of the additive Markov chain constructed using the memory function $F(r)$ [Eq. (38)] (shown in the inset) with memory length $r = N = 10$ and the parameters $\bar{a} = \frac{1}{2}$ and $F_0 = 0.15$. The solid line represents the numerical solution of Eq. (9). The dots represent the calculations by the definition (8) of generating a numerical sequence with the CPDF (7).

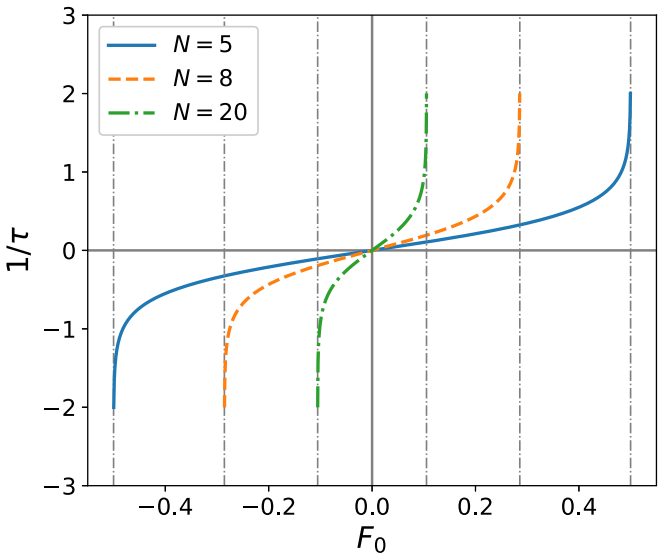

FIG. 2. Dependence of inverse temperature τ^{-1} defined by Eqs. (37) and (26) for the additive Markov chains with the CPDF (14) and memory function (38) for $N = 5, 8, 20$ (the corresponding lines are marked in the legend). The values of the parameter F_0 when the inverse temperature goes asymptotically to infinity are determined by the condition (4), i.e., $|F_0| \sum_{r=1}^{N}(1 - \frac{r}{N}) = 1$.

symbolic dynamics, the detailed dependence of the conditional probability on all previous symbols is replaced by an effective macroscopic parameter μ that represents the average correlation strength within the sequence. This "information averaging" provides the basis for introducing the concept of information temperature as a thermodynamiclike measure characterizing the global degree of order and randomness in the symbolic system.

V. NUMERICAL SIMULATIONS

This section illustrates the analytical results obtained. Using the construction of an additive Markov chain with a linearly decreasing memory function, we demonstrate how this chain is related to a stepwise Markov chain and present the temperature calculation results.

It is supposed that the modeled statistical properties of the random chain are determined by the probability of the symbol 0 or 1 occurring, $p_\alpha = \frac{1}{2}$, and the additive memory function

$$F(r) = \begin{cases} F_0\left(1 - \frac{r}{N}\right), & 1 \leqslant r < N \\ 0, & r \geqslant N. \end{cases} \tag{38}$$

The results of calculating the correlation function $K(r)$ by two different methods are shown by the solid curve and dots in Fig. 1.

Using the calculated values of the correlation function, we find the parameter μ and then the temperature of the additive Markov chain. The result of these calculations is shown in Fig. 2 for several values of the parameter F_0 of the additive Markov chain memory function.

The asymptotic increase of τ^{-1} toward infinity as F_0 approaches the ergodicity bound signifies a transition from a stochastic process to a deterministic one. Physically, this blowup represents the limit of maximum informational order, analogous to the absolute zero temperature in a physical system where fluctuations are completely suppressed by strong correlations.

It is clear that a Markov chain with a stepwise memory function is a coarse-grained description of an additive Markov chain. This coarsening is accompanied by a loss of information, which is reflected in an increase in the source entropy, as follows from Fig. 3.

However, we can achieve equality of the entropies of these two chains by fitting the parameters N and μ of the stepwise

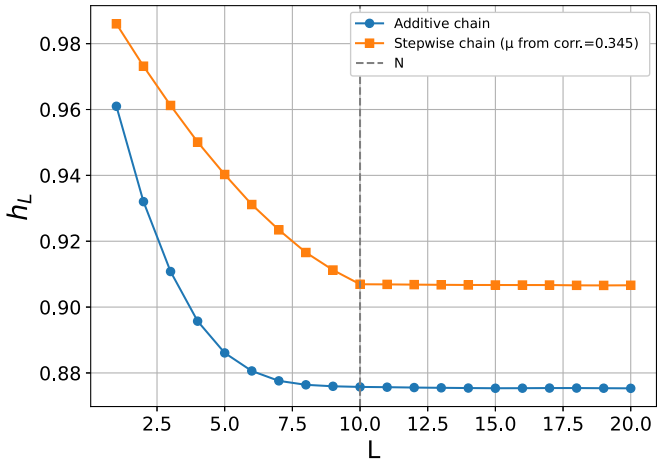

FIG. 3. The lower curve is the dependence of conditional entropies defined by Eqs. (12) and (13) for the additive Markov chains with the CPDF (14) and the memory function (38) with $N = 10$ and $F_0 = 0.15$. The calculated parameter $\mu = 0.345$, defined by Eq. (26), gives the entropy of the stepwise chain represented by the upper curve.

Markov chain for the fixed parameters of the additive Markov chain. This can be viewed as another principle of equivalence of random sequences, based on the equality of the source entropies.

VI. CONCLUSION

In this work we have analyzed additive multistep Markov chains as a class of stochastic processes capable of capturing long-range correlations while avoiding the exponential parameter growth characteristic of classical high-order Markov models. By establishing a correspondence between additive Nth-order binary Markov chains and chains with a stepwise memory function, we have extended the concept of information temperature to additive processes with arbitrary memory depth.

Within this framework, information temperature emerges as an intrinsic macroscopic parameter determined by the correlation structure of the symbolic sequence. It characterizes the balance between entropy and informational constraints in a manner closely analogous to temperature in statistical physics. The additive structure of the conditional probability allows this balance to be analyzed explicitly, providing a transparent example of how macroscopic descriptors can arise from microscopic stochastic dynamics in high-dimensional systems.

Although the present analysis does not model large-scale language models directly, it offers a conceptual bridge to their probabilistic behavior. In LLMs, temperature enters through the softmax function as a control parameter regulating the entropy of the output distribution under fixed logits. From a variational viewpoint, this corresponds to selecting a Gibbs distribution that minimizes a free-energy functional at a prescribed temperature. In contrast, the information temperature introduced here is not imposed externally but emerges from the internal correlation structure of the stochastic process.

Despite this difference in origin, both notions of temperature play the same mathematical role as Lagrange multipliers governing the tradeoff between entropy and energetic constraints. This correspondence suggests that sampling temperature in LLMs may be interpreted as an operational counterpart of information temperature, acting on an implicit energy landscape learned during training rather than on explicitly specified transition probabilities. Such an interpretation provides a theoretically grounded perspective on the widespread use of temperature in generative models, without requiring a direct identification between neural architectures and Markovian dynamics.

The results presented here contribute to a broader effort to understand high-dimensional symbolic systems through reduced, macroscopic descriptions. Future work may extend this approach to larger alphabets, nonadditive interactions, or data-driven estimation of information temperature in empirical sequences, including those generated by modern language models.

VII. PERSPECTIVES

Several directions for future research emerge naturally from our findings. Extending the temperature formalism beyond binary alphabets to multisymbol systems remains an important challenge, particularly for applications to natural language. Another promising direction is the quantitative comparison of empirical LLM-generated sequences with the predictions of additive Nth-order Markov models, which could clarify the degree to which LLM statistical behavior can be approximated by low-dimensional macroscopic parameters. Finally, the temperature-based characterization of sequence complexity may offer new diagnostic and interpretability tools for large-scale generative models.

The next step in studying and developing the concept of information temperature and other macroscopic parameters for description is its application to the analysis of symbolic random sequences with an arbitrary finite state space and various memory functions.

The concept of temperature cannot always be introduced into thermodynamics, and even if it is, a system is not always characterized by a single temperature. Similarly, for stationary random sequences, it is important to consider the conditions under which the concept of information temperature can be introduced.

Investigating the informational meaning of the introduced temperature is of interest. In particular, the question arises whether temperature can characterize the academic level of a text or serve as an indicator of the author's brain activity. The macroscopic interpretation of temperature in information systems could thus serve as a quantitative bridge between physical and artificial intelligence paradigms.

Future work should focus on exploring whether the information temperature can characterize not only structural complexity but also semantic richness in natural language data, or cognitive activity in generative processes. Such studies could deepen our understanding of how thermodynamic principles manifest in symbolic and cognitive systems. These directions may lead to a deeper mathematical theory connecting LLM architectures with interpretable stochastic processes and may help clarify the fundamental mechanisms driving high-dimensional sequence modeling.

ACKNOWLEDGMENTS

The authors thank V. E. Vekslerchik for discussions during this work. O.V.U. is grateful to the Center for Theoretical Physics, Polish Academy of Sciences, Warsaw for financial support and hospitality.

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

APPENDIX: MAXIMUM-ENTROPY FORMULATION

Consider the joint probability distribution $P(a_1, \ldots, a_{N+1})$ over $N+1$ blocks of a symbolic sequence. In addition to the normalization condition, we impose a constraint on a macroscopic quantity, namely, the average energy

$$\langle E \rangle = \sum_{a_1,\ldots,a_{N+1}} P(a_1, \ldots, a_{N+1}) E(a_1, \ldots, a_{N+1}). \quad \text{(A1)}$$

Maximizing the block entropy (12) under this constraint using Lagrange multipliers yields

$$P(a_1, \ldots, a_{N+1}) = \frac{1}{Z} \exp[-\beta E(a_1, \ldots, a_{N+1})], \quad \text{(A2)}$$

where Z is the partition function ensuring normalization.

In this formulation, the information temperature τ appears as a macroscopic parameter that quantifies the balance between entropy and imposed constraints. The specific form of the energy function $E(a_1, \ldots, a_{N+1})$ is not unique and may encode different structural characteristics of the sequence, such as correlation functions or memory effects.

For the stepwise Markov chain (5) of the second order $N = 2$, using the three-point joint probabilities [4]

$$P_{000} = P_{111} = \frac{1+2\mu}{8(1-\mu)}, \quad \text{(A3)}$$

$$P_{001} = \cdots = P_{110} = \frac{1-2\mu}{8(1-\mu)} \quad \text{(A4)}$$

and the auxiliary energies of configurations within the pair interaction approximation,

$$\varepsilon(000) = \varepsilon(111) = -2\varepsilon_1 - \varepsilon_2, \quad \text{(A5)}$$

$$\varepsilon(010) = \varepsilon(101) = 2\varepsilon_1 - \varepsilon_2, \quad \text{(A6)}$$

$$\varepsilon(001) = \cdots = \varepsilon(110) = \varepsilon_2, \quad \text{(A7)}$$

one can derive from Eq. (A2) the expression (33) for the information temperature τ, as well as the equality of fictive energies

$$\varepsilon_2 = \varepsilon_1, \quad \text{(A8)}$$

which is consistent with the assumption of uniform interaction in the stepwise model.

From a variational perspective, the temperature parameter in LLMs can also be interpreted as a Lagrange multiplier. The softmax distribution arises as the solution of a maximum-entropy problem under a constraint on the expected logits, which act as an effective energy (up to a sign). This parallels the definition of information temperature, where the multiplier is associated with constraints imposed by the structure of the symbolic sequence. Although the underlying quantities differ, the role of temperature as a parameter governing the balance between entropy and imposed constraints is the same.

It is important to note that the definition of information temperature is not unique and depends on how the effective energy is introduced. In the maximum-entropy framework, the temperature arises as a Lagrange multiplier associated with a prescribed constraint on the average energy, which requires that the probability distribution admits a Gibbs-type representation with a well-defined energy function. However, for general symbolic sequences, including additive Markov chains, such a representation may not exist or may be only approximate, leading to an overdetermined system for the temperature parameter.

For example, in the case of the stepwise Markov chain of order $N = 3$ with the constraint (A1), the corresponding system does not admit a consistent solution for τ within the admissible range of μ.

In contrast, our entropy-based definition of information temperature does not rely on the existence of an explicit Gibbs structure. Instead, it is defined through the derivative of entropy with respect to a suitably chosen effective energy function, $\tau^{-1} = dH/dE$, and is therefore directly determined by the statistical properties of the sequence. As a result, this definition remains applicable even when a global exponential representation of the distribution is not available.

[1] W. X. Zhao *et al.*, A survey of large language models, arXiv:2303.18223.

[2] M. Phuong and M. Hutter, Formal algorithms for transformers, arXiv:2207.09238.

[3] O. Zekri, A. Odonnat, A. Benechehab, L. Bleistein, N. Boullé, and I. Redko, Large language models as Markov chains, arXiv:2410.02724.

[4] O. V. Usatenko, S. S. Melnyk, G. M. Pritula, and V. A. Yampol'skii, Information entropy and temperature of the binary Markov chains, Phys. Rev. E **106**, 034127 (2022).

[5] O. V. Usatenko and G. M. Pritula, Information temperature as a measure of complexity of random symbolic sequences, Chaos Soliton. Fract. **202**, 117452 (2025).

[6] O. V. Usatenko and V. A. Yampolskii, Binary N-step Markov chains and long-range correlated systems, Phys. Rev. Lett. **90**, 110601 (2003).

[7] Z. A. Mayzelis, S. S. Apostolov, S. S. Melnyk, O. V. Usatenko, and V. A. Yampol'skii, Additive N-step Markov chains as prototype model of symbolic stochastic dynamical systems with long-range correlations, Chaos Soliton. Fract. **34**, 112 (2007).

[8] O. V. Usatenko, S. S. Apostolov, Z. A. Mayzelis, and S. S. Melnik, *Random Finite-Valued Dynamical Systems: Additive Markov Chain Approach*, Kharkov Series in Physics and Mathematics (Cambridge Scientific, Cambridge, 2010), Vol. 1.

[9] S. S. Melnyk, O. V. Usatenko, and V. A. Yampol'skii, Memory functions of the additive Markov chains: Applications to complex dynamic systems, Physica A **361**, 405 (2006).

[10] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences* (Springer, Berlin, 1985).

[11] S. S. Apostolov, Z. A. Mayzelis, O. V. Usatenko, and V. A. Yampol'skii, Equivalence of the Markov chains and two-sided symbolic sequences, Europhys. Lett. **76**, 1015 (2006).

[12] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).

[13] O. V. Usatenko, V. A. Yampol'skii, K. E. Kechedzhy, and S. S. Mel'nyk, Symbolic stochastic dynamical systems viewed as binary N-step Markov chains, Phys. Rev. E **68**, 061107 (2003).