Вісник ХНУ, № 1076, серія «Фізика», вип. 19, 2013. с. 28-34

PACS: 05.40.-a, 87.10.+e
УДК: 519.218

# Энтропия случайных бинарных последовательностей с дальними корреляциями

С.С. Мельник, Г.М. Притула, О.В. Усатенко

*Институт радиофизики и электроники имени А. Я. Усикова НАН Украины,*
*ул. Ак. Проскуры, 12, г. Харьков, 61085, Украина*
*pritula.galina@gmail.com*

На основе метода аддитивных цепей Маркова вычислена дифференциальная энтропия N-шаговой бинарной стационарной эргодической случайной последовательности. В предположении малости корреляций энтропия представляется в виде функционала парной корреляционной функции. Предполагается также, что случайная последовательность полностью задана своими парными корреляционными функциями, а не вероятностями реализации блоков – подпоследовательностей конечной длины, как это имеет место в стандартном описании случайных последовательностей. В рамках этих приближений оказывается возможным вычислить энтропию последовательности на гораздо больших расстояниях, чем это можно сделать с помощью блочной энтропии. Показано, что для некоторых классов последовательностей имеет место самоподобие энтропии по отношению к процедуре прореживания – случайному или регулярному удалению из последовательности некоторой доли символов.

**Ключевые слова**: энтропия, корреляции, случайность, самоподобие.

На основі методу адитивних ланцюгів Маркова обчислена диференціальна ентропія N- крокової бінарної стаціонарної ергодичної випадкової послідовності. У припущенні малості кореляцій ентропія є функціоналом парної кореляційної функції. Передбачається також, що випадкова послідовність повністю задана своїми парними кореляційними функціями, а не ймовірностями реалізації блоків – підпослідовностей кінцевої довжини, як це має місце у стандартному описі випадкових послідовностей. У рамках цих наближень виявляється можливим обчислити ентропію послідовності на багато більших відстанях, ніж це можна зробити за допомогою блокової ентропії. Показано, що для деяких класів послідовностей має місце самоподібність ентропії по відношенню до процедури проріджування – випадковому або регулярному видаленню з послідовності деякої частки символів.

**Ключові слова**: ентропія, кореляції, випадковість, самоподібність.

Differential entropy of the N-step stationary ergodic random binary sequence is calculated with the use of the additive Markov chain method. In the assumption of weak correlation, the entropy is represented as a functional of the pair correlation function. The random sequence is also supposed to be completely defined by its pair correlation function, rather than by the block probabilities, as is the standard description of random sequences. Within these approximations, the entropy of the sequence can be calculated at a much longer distance than with the use of the block entropy. The self-similarity of the entropy with respect to the decimation procedure – random or regular removal of some portion of the sequence characters – is revealed for certain classes of chains.

**Keywords**: entropy, correlations, randomness, self-similarity.

## Введение

В настоящее время существует общепринятая точка зрения, что наш мир является сложной коррелированной системой. Среди наиболее ярких примеров этой концепции можно привести последовательности нуклеотидов в молекуле ДНК и последовательности аминокислот в белках, записи мозговой и сердечной активности, человеческое общение и информационный обмен в животном мире. Письменные тексты, написанные на естественных языках, солнечная активность, погода (хаотический характер атмосферы), потоки данных в компьютерных сетях, фондовые индексы дополняют эти примеры. Не удивительно поэтому, что системы с дальним

взаимодействием и естественные последовательности с нетривиальным информационным содержанием в последние десятилетия находятся в центре внимания большого количества исследований в различных областях науки. Неослабевающий интерес к системам с коррелированными флуктуациями определяется также широкими возможностями практического применения специфических свойств, которые они демонстрируют.

Одним из базовых объектов в теории сложных коррелированных систем являются случайные последовательности. Они возникают, например, в результате огрубления отображения эволюции хаотической динамической системы на конечный набор символов [1, 2] и являются предметом изучения алгоритмической сложности (Колмогорова-Соломонова-Чаитина), искусственного интеллекта, теории информации и сжатия цифровых данных, а также служат инструментом при проектировании устройств и приборов со случайными компонентами в их структуре (различные фильтры, дифракционные решетки, искусственные материалы, антенны, конверторы, линии задержки и т.д. [3]). Такие устройства могут обладать необычными свойствами и аномальными динамическими, кинетическими или транспортными характеристиками, контролируемыми соответствующим выбором беспорядка.

Информационная энтропия является одной из наиболее удобных и общеупотребительных характеристик случайной динамической системы [4]. Будучи мерой неопределенности и информационной избыточности в случайной цепи символов, она служит мощным и важным инструментом в описании сложных явлений и используется для анализа широкого класса явлений, протекающих в сложных динамических системах. Знание энтропии последовательности помогает оценить количественно степень сложности системы и возможности сжатия информации, содержащейся в цепи символов – тексте, классификации систем по степени их сложности и многим другим статистическим параметрам системы. Существует несколько способов численной оценки энтропии, которые, как правило, требуют обработки огромных объёмов данных. Теоретические модели для аналитической оценки энтропии случайных последовательностей практически отсутствуют.

Стандартный способ описания статистических свойств конкретной случайной последовательности данных заключается в следующем. Сначала анализируется последовательность с целью вычисления корреляционных функций или вероятностей появления фрагментов последовательности (слов), длина $L$ которых больше длины корреляции $R_c$, но меньше длины $M$ последовательности,

$$R_c < L \ll M. \tag{1}$$

В то же время, количество $d^L$ различных слов длины $L$, записанных в алфавите содержащем $d$ букв должно быть существенно меньше, чем полное число $M$-$L$ слов длины $L$ в последовательности длины $M$:

$$d^L \ll M - L. \tag{2}$$

Следующим шагом в процедуре описания последовательности является выражение её корреляционных свойств в терминах функции условной вероятности цепи Маркова (см. формулу (5) ниже). Естественно, что порядок $N$ марковской цепи – глубина памяти – должен быть больше, чем корреляционная длина $R_c$,

$$R_c < N. \tag{3}$$

Представляющие интерес естественные последовательности (например, письменные тексты или последовательности нуклеотидов ДНК) обычно имеют длины корреляции порядка длины самой последовательности. Длины типичных слов, позволяющие правильно оценить вероятность их появления, составляют 4-5 символов для реального текста длины $10^6$, написанного на естественном языке (с использованием алфавита, содержащего 27-30 букв и символов), или порядка 20 символов для огрублённого текста, представленного посредством двоичной последовательности, то есть количество различных слов длины $L=20$ тоже порядка длины самого текста, $2^{20}\sim10^6$. Таким образом, в приведенных примерах ни одно из неравенств (1) - (3) не может быть выполнено.

Чтобы преодолеть эти трудности, мы предлагаем другой метод решения проблемы, позволяющий найти энтропию рассматриваемой последовательности. В настоящей работе мы жертвуем знанием точной статистики коротких слов и воспроизводим слабую дальнюю память, которая может быть выражена в терминах парной корреляционной функции символов и найдена посредством численного анализа последовательности на расстоянии, сравнимом с длиной последовательности. После этого мы разрабатываем метод построения функции условной вероятности, представленной с помощью парного коррелятора, что делает возможным аналитический расчет энтропии последовательности.

Изложение статьи построено следующим образом. Сначала мы кратко обсуждаем модель $N$-шаговой аддитивной цепи Маркова и, предполагая, что корреляции между символами в цепи не слишком сильны, выражаем функцию условной вероятности

цепи через парную корреляционную функцию. В следующем разделе мы выражаем дифференциальную энтропию в терминах функции условной вероятности цепи Маркова и представляем её в виде суммы квадратов парных корреляторов. Затем мы обсуждаем некоторые свойства полученных результатов, в частности, свойство самоподобия энтропии по отношению к процедуре прореживания для некоторых конкретных классов цепей Маркова. После обсуждения применимости нашей модели к анализу литературных текстов мы предлагаем возможные варианты расширения и углубления дальнейших исследований.

**Аддитивная цепь Маркова**

Рассмотрим последовательность $\mathbb{A} = \ldots, a_{-1}, a_0, a_1, a_2, \ldots$ вещественных случайных величин $a_i$, взятых из конечного алфавита $A=\{0,1,\ldots,d-1\}$. Последовательность $\mathbb{A}$ является $N$-шаговой марковской цепью, если она обладает следующим свойством: вероятность того, что символ $a_i$ имеет определенное значение $a$ при фиксированных значениях всех остальных символов, зависит только от значений $N$ предыдущих символов,

$$P(a_i = a \,|\, \ldots, a_{i-2}, a_{i-1}) = \\ = P(a_i = a \,|\, a_{i-N}, \ldots, a_{i-2}, a_{i-1}). \tag{4}$$

Иногда номер $N$ также называют порядком или глубиной памяти цепи Маркова. Условная вероятность $P(a_i=a|a_{i-N}, \ldots, a_{i-2}, a_{i-1})$ полностью определяет статистические свойства марковской цепи и метод её итерационного численного построения. Если последовательность, статистические свойства которой мы хотим проанализировать, известна, функция условной вероятности $N$-го порядка может быть найдена по стандартной методике,

$$P_N(a_{N+1} = a \,|\, a_1, \ldots, a_N) = \\ = \frac{P(a_1, \ldots, a_N, a)}{P(a_1, \ldots, a_N)}, \tag{5}$$

где $P(a_1, \ldots, a_N)$ – вероятность появления $N$-слова $a_1, \ldots, a_N$. Марковская цепь, определённая выражением (4), является однородной последовательностью, так как её условная вероятность не зависит явно от $i$, т. е. не зависит от позиции символов $a_{i-N}, \ldots, a_{i-1}$, $a_i$ в цепочке. Она зависит только от значений, принимаемых символами $a_{i-N}, \ldots, a_{i-1}$, $a_i$, и их взаимного расположения. Однородная последовательность является стационарной: среднее значение любой зависящей от $s$ аргументов функции $f(a_{r_1}, a_{r_1+r_2}, \ldots, a_{r_1+\ldots+r_s})$,

$$\overline{f}(a_{r_1}, \ldots, a_{r_1+\ldots+r_s}) = \lim_{M\to\infty} \frac{1}{2M+1} \times \\ \times \sum_{i=-M}^{M} f(a_{i+r_1}, \ldots, a_{i+r_1+\ldots+r_s}), \tag{6}$$

зависит только от $s-1$ разностей индексов. Иными словами, все статистически усредненные функции случайных величин однородной последовательности инвариантны относительно сдвига.

Предполагается, что цепь эргодическая. Согласно теореме Маркова (см., например, [5]), это свойство справедливо для однородных цепей Маркова, если для всех возможных значений аргументов функции (4) выполняются строгие неравенства

$$0 < P(a_i = a \,|\, a_{i-N}, \ldots, a_{i-1}) < 1, \\ i \in \mathbb{Z} = \ldots, -1, 0, 1, 2, \ldots \tag{7}$$

Из эргодичности следует, что корреляции между любыми блоками символов в цепи стремятся к нулю, когда расстояние между ними стремится к бесконечности, и в качестве репрезентативного представителя ансамбля цепей можно использовать одну случайную последовательность, заменив усреднение по ансамблю усреднением по цепочке (6).

Ниже мы будем рассматривать важный класс бинарных случайных последовательностей, в которых каждый символ $a_i$ может принимать только два значения, например, 0 и 1, $a_i \in \{0,1\}$. Условная вероятность обнаружения $i$-го элемента $a_i=1$ для двоичной $N$-шаговой марковской последовательности в зависимости от $N$ предшествующих элементов $a_{i-N}, \ldots, a_{i-1}$ представляет собой набор $2^N$ независимых чисел:

$$p_i(1|N) = p(a_i = 1 \,|\, a_{i-N}, \ldots, a_{i-1}), \\ p_i(0|N) = 1 - p(a_i = 1 \,|\, a_{i-N}, \ldots, a_{i-1}). \tag{8}$$

Условная вероятность (8) двоичной последовательности случайных величин $a_i \in \{0,1\}$ может быть точно представлена в виде полиномиального разложения:

$$p_i(1|N) = \bar{a} + \sum_{r_1=1}^{N} F_1(r_1)(a_{i-r_1} - \bar{a}) + \\ \sum_{r_1,r_2=1}^{N} F_2(r_1,r_2)(a_{i-r_1}a_{i-r_2} - \overline{a_{i-r_1}a_{i-r_2}}) + \ldots + \\ \sum_{r_1,\ldots,r_N=1}^{N} F_N(r_1,\ldots,r_N)(a_{i-r_1}\ldots a_{i-r_N} - \overline{a_{i-r_1}\ldots a_{i-r_N}}), \tag{9}$$

где под статистическим усреднением имеется в виду усреднение (6) по последовательности, $F_s$ – семейство функций памяти и $\overline{a}$ – относительное среднее число единиц в последовательности. Возможность представления условной вероятности (8) в виде (9) следует из простых тождеств $a^2=a$ и $f(a)=af(1)+(1-a)f(0)$ для произвольной функции $f(a)$, определённой на множестве $a\in\{0,1\}$. Если в уравнении (9) оставить только первое слагаемое, то условная вероятность $p_i(1|N)$ будет описывать генерацию некоррелированной последовательности. С учётом второго члена, пропорционального $F_1(r)$, корреляции в цепи могут быть воспроизведены с точностью до второго порядка, при этом все корреляторы высших порядков будут выражаться через парные корреляционные функции. В дальнейшем мы будем использовать только два первых члена, которые определяют так называемую аддитивную цепь Маркова [6, 7].

Частным случаем функции условной вероятности аддитивной цепи Маркова является цепь со ступенчатой функцией памяти,

$$p_i(1|k)=\overline{a}+\mu\left(\frac{2k}{N}-1\right). \qquad (10)$$

Вероятность $p_i(1|k)$ реализации символа $a_i=1$ после $N$-слова $a_{i-N+1},\ldots,a_{i-1}$, содержащего $k$ единиц, $k=\sum_{l=1}^{N}a_{i-l}$, является линейной функцией $k$ и не зависит от расположения символов в слове $a_{i-N},\ldots,a_{i-1}$, параметр μ характеризует силу корреляций в системе.

Существует довольно простое соотношение между функцией памяти $F(r)$ (здесь и ниже мы будем опускать индекс 1 у функции памяти) и парной корреляционной функцией бинарной аддитивной цепи Маркова. Известны два метода нахождения $F(r)$ для последовательности с известной парной корреляционной функцией. Первый [6] основан на минимизации «расстояния» между цепью Маркова, генерируемой при помощи искомой функции памяти, и исходной заданной последовательностью символов с известной корреляционной функцией. Такая минимизация приводит к уравнению связи между корреляционной функцией и функцией памяти,

$$K(r)=\sum_{r'=1}^{N}F(r')K(r-r'),\quad r\geq 1, \qquad (11)$$

где нормированная корреляционная функция $K(r)$ имеет вид

$$C(r)=\overline{(a_i-\overline{a})(a_{i+r}-\overline{a})},\quad K(r)=\frac{C(r)}{C(0)}. \qquad (12)$$

Второй способ вывода уравнения (10) – полностью вероятностное непосредственное вычисление [8].

Несмотря на простоту, уравнение (11) может быть решено аналитически только в некоторых частных случаях: для одно- или двух-шаговой цепи, цепи Маркова со ступенчатой функцией памяти и так далее. Во избежание трудностей, возникающих при решении уравнения (11), мы предполагаем, что корреляции в последовательности малы (по амплитуде, но не по длине). Это означает, что все компоненты нормированной корреляционной функции являются малыми, $|K(r)|<<1$, $|r|\neq 0$, за исключением $K(0)=1$. Таким образом, принимая во внимание, что в правой части уравнения (11) главным является член $K(0)=1$, можно получить приближенное решение для функции памяти в виде ряда

$$F(r)=K(r)-\sum_{r'\neq r}^{N}K(r-r')K(r')+\sum_{r'\neq r}^{N}\sum_{r''\neq r'}^{N}K(r-r')K(r'-r'')K(r'')+\ldots \qquad (13)$$

Уравнение для функции условной вероятности в первом приближении по малому параметру $|K(r)|<<1$, $|r|\neq 0$ имеет вид:

$$p_i(1|N)\simeq\overline{a}+\sum_{r=1}^{N}F(r)(a_{i-r}-\overline{a})\simeq\simeq\overline{a}+\sum_{r=1}^{N}K(r)(a_{i-r}-\overline{a}). \qquad (14)$$

Эта формула даёт инструмент для построения последовательности с заданной парной корреляционной функцией. Очевидно, что мы можем рассматривать только последовательности с корреляционной функцией, определяемые $p_i(1|L)$, которая удовлетворяет неравенствам (7).

Как правило, корреляционные функции используются в качестве входных характеристик для описания коррелированных случайных последовательностей. Тем не менее, парная корреляционная функция описывает не только прямое взаимодействие элементов $a_i$ и $a_{i+r}$, но и учитывает их косвенное взаимодействие через все промежуточные элементы последовательности. Наш подход работает с «первичными» характеристиками системы, в частности, с функцией памяти. Корреляционные функции и функции памяти являются взаимодополняющими характеристиками случайной последовательности в следующем смысле. Численный анализ данной случайной последовательности позволяет непосредственно определить корреляционную

функцию, а не функцию памяти. С другой стороны, построение случайной последовательности возможно при помощи функции памяти, а не корреляционной функции. Таким образом, функция памяти позволяет полнее раскрыть внутренние свойства коррелированных систем. Уравнение (14) показывает, что в пределе слабых корреляций обе функции играют одну и ту же роль.

Концепция аддитивной цепи Маркова широко использовалась ранее для изучения случайных последовательностей с дальними корреляциями. Примеры и ссылки можно найти в [7].

**Дифференциальная энтропия**

Для оценки энтропии бесконечной стационарной последовательности $\mathbb{A}$ символов $a_i$ можно использовать блочную энтропию

$$H_L = -\sum_{a_1,\ldots,a_L} P(a_1,\ldots,a_L)\log_2 P(a_1,\ldots,a_L). \quad (15)$$

Дифференциальная энтропия, или энтропия, приходящаяся на один символ, определяется выражением

$$h_L = H_{L+1} - H_L, \quad (16)$$

и характеризует степень неопределенности в появлении *(L+1)*-ого символа, если предыдущие $L$ символов известны. Энтропия источника (или энтропия Шеннона) − это дифференциальная энтропия в асимптотическом пределе, $h = \lim_{L\to\infty} h_L$. Эта величина измеряет среднюю информацию, приходящуюся на один символ, при учёте всех корреляций в системе.

Дифференциальная энтропия может быть представлена в терминах функции условной вероятности. Чтобы показать это, мы должны переписать уравнение (15) для блока длиной $L+1$ и выразить $P(a_1,\ldots,a_{L+1})$ через условную вероятность. В результате получим

$$\begin{aligned} h_L = -\sum_{a_1,\ldots,a_L} & P(a_1,\ldots,a_L)\times \\ \times\sum_{a_{L+1}} & P(a_{L+1}\,|\,a_1,\ldots,a_L)\times \\ \times\log_2 & P(a_{L+1}\,|\,a_1,\ldots,a_L). \end{aligned} \quad (17)$$

Таким образом, дифференциальная энтропия $h_L$ случайной последовательности оказывается представленной как обобщение стандартной условной энтропии $H = -\sum_A P(A)\sum_B P(B\,|\,A)\log_2 P(B\,|\,A)$, но для многосимвольного события $a_1,\ldots,a_L$,

$$h_L = \sum_{a_1,\ldots,a_L} P(a_1,\ldots,a_L)h(a_{L+1}\,|\,L) = \overline{h(a_{L+1}\,|\,L)}, \quad (18)$$

где $h(a_{L+1}|L)$ – это количество информации, содержащейся в $(L+1)$-ом символе последовательности, обусловленном предыдущими $L$ символами,

$$\begin{aligned} h(a_{L+1}\,|\,L) = & -p_i(1\,|\,L)\log_2 p_i(1\,|\,L) - \\ & -(1-p_i(1\,|\,L))\log_2(1-p_i(1\,|\,L)). \end{aligned} \quad (19)$$

Условная вероятность $p_i(1|L)$, $L<N$, появления единицы на $i$-ом месте после $L$-слова $a_1,\ldots,a_L$

$$p_i(1\,|\,L) = \bar{a} + \delta; \quad \delta = \sum_{r=1}^{L} F(r)(a_{i-r} - \bar{a}), \quad (20)$$

в первом приближении по параметру $\delta$ получается из уравнения (14) и условия совместности условной вероятности и вероятностей появления $L$ -слов для уравнения Колмогорова-Чепмэна (см., например, [9]).

С учётом малости корреляций $|\,\delta\,| \ll 1$ уравнение (19) можно разложить в ряд Тейлора до второго порядка по $\delta$,

$$h(a_{L+1}\,|\,L) = h_0 + \delta\left.\frac{dh}{dp_i}\right|_{p_i=\bar{a}} + \frac{\delta^2}{2}\left.\frac{d^2h}{dp_i^2}\right|_{p_i=\bar{a}}, \quad (21)$$

где производные берутся в точке «равновесия» $p_i(1\,|\,L) = \bar{a}$ и $h_0$ – энтропия некоррелированной последовательности:

$$h_0 = -\bar{a}\log_2(\bar{a}) - (1-\bar{a})\log_2(1-\bar{a}). \quad (22)$$

Используя уравнение (18) для усреднения $h(a_{L+1}|L)$

$$\begin{aligned} h_L = \overline{h(a_{L+1}\,|\,L)} = h_0 + \\ + \bar{\delta}\left.\frac{dh}{dp_i}\right|_{p_i=\bar{a}} + \frac{\overline{\delta^2}}{2}\left.\frac{d^2h}{dp_i^2}\right|_{p_i=\bar{a}} \end{aligned} \quad (23)$$

и учитывая, что $\bar{\delta} = 0$, для условной энтропии последовательности получаем

$$h_L = \begin{cases} h_{L\le N} = h_0 - \dfrac{1}{2\ln 2}\displaystyle\sum_{r=1}^{L} F^2(r), \\ h_{L>N} = h_{L=N}. \end{cases} \quad (24)$$

Если длина блока превышает длину памяти, $L>N$, условная вероятность $p_i(1|L)$ зависит только от $N$ предыдущих символов, см. уравнение (4). Тогда, как следует из (18), для $N$-шаговой цепи Маркова дифференциальная энтропия остается постоянной

при $L \geq N$ . Второе из уравнений (24) согласуется с первым, поскольку в первом приближении по $\delta$ корреляционная функция обращается в нуль при $L>N$ вместе с функцией памяти. Окончательное выражение – основной результат работы – для дифференциальной энтропии стационарных эргодических бинарных слабо коррелированных случайных последовательностей имеет вид

$$h_L = h_0 - \frac{1}{2\ln 2}\sum_{r=1}^{L} K^2(r). \qquad (25)$$

**Обсуждение**

Из (25) следует, что корреляционная поправка к энтропии некоррелированной последовательности является отрицательной. Это ожидаемый результат — корреляции уменьшают энтропию. Вывод не чувствителен к знаку корреляции: персистентные корреляции, K>0, описывающие «притяжение» одинаковых символов, и анти-персистентные, K<0, соответствующие «отталкиванию» между 0 и 1, дают вклад одного и того же отрицательного знака. Если корреляционная функция постоянна при $0<r\leq N$, энтропия является линейной убывающей функцией аргумента L вплоть до L=N. Этот результат совпадает с полученным ранее в работе [10] для цепи Маркова со ступенчатой функцией памяти (10).

В качестве иллюстрации результата (25) на рис.1 приведен график зависимости дифференциальной энтропии от длины слова. Как численные, так и аналитические результаты представлены для степенной функции корреляции $K(r)=0.01/r^{1.1}$. Хорошее совпадение кривых подтверждает адекватность использования приближения аддитивной марковской цепи. Резкое отклонение графика блочной энтропии от аналитической при $L\sim10$ является результатом нарушения неравенства (2).

Известно, что цепи Маркова со ступенчатой функцией памяти [11] и более широкий класс перестановочных цепей [12] инвариантны относительно процедуры прореживания. Перестановочными называются цепи, функция условной вероятности которых не зависит от порядка символов в $N$-слове, предшествующих генерируемому символу. Прореживание – это преобразование случайной последовательности путём регулярного или случайного удаления некоторой части символов из цепи. Как было показано в [12], после прореживания корреляционные функции указанных классов последовательностей инвариантны с точностью до новой уменьшенной длины памяти $N^*=\lambda N$ , где $\lambda<1$ – относительная неудалённая часть символов в цепочке. Это значит, что уравнение не изменяет своей формы, но вместо $N$ мы должны подставить новую длину памяти, т.е. энтропия

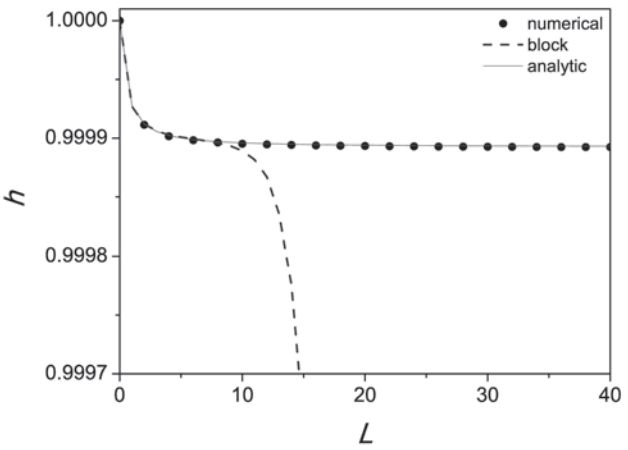


Рис. 1. Зависимость дифференциальной энтропии от длины слова. Сплошной линией представлен аналитический результат с корреляционной функцией $K(r)=0.01/r^{1.1}$ , точки соответствуют прямой численной оценке того же выражения для последовательности, численно построенной при помощи функции памяти (14) и численно найденной корреляционной функции построенной последовательности. Пунктирной линией представлен результат вычисления блочной энторпии (15) и (16) путём замены вероятности появления блоков относительной частотой их появления. Численные результаты получены для последовательности длины $M=10^8$.

таких последовательностей также обладает свойством самоподобия относительно процедуры прореживания.

**Применение к письменным текстам**

Теория аддитивных цепей Маркова с дальней памятью использовалась для описания свойств коррелированных литературных текстов [8]. Было показано, что огрубленные естественные письменные тексты являются сильно коррелированными последовательностями, которые обладают антиперсистентными свойствами на малых расстояниях (в области $L\leq300$ действия грамматических правил) и слабо персистентными последовательностями со степенным образом убывающей корреляционной функцией на больших расстояниях (в области $L\geq300$ действия семантических правил). Очевидно, что рассмотренная аддитивная марковская цепь может претендовать только на описание слабой степенной части энтропии, пропорциональной $L^{-\gamma}$.

Эбелинг и Николис [13] и Шурман и Грассбергер [14] предложили эмпирически найденную форму энтропии письменных текстов

$$h_L = h + c\frac{\log_2 L}{L^{\gamma}}, \quad \gamma > 0. \qquad (26)$$

Возникает естественный вопрос о природе этой

зависимости. Частичным ответом на этот вопрос может быть следующее объяснение. Энтропия марковской цепи со ступенчатой функцией памяти в пределе сильной корреляции, $N\ln N(1-2\mu) << 4\mu$, была получена в виде [10]

$$h_L = h + c\frac{\log_2 L}{L}. \qquad (27)$$

После сравнения результатов (24) и (27) с выражением (26) становится ясно, что член $\log_2 L$ описывает сильные ближние корреляции, а степенной член $L^{-\gamma}$ ответственен за слабые дальние корреляции. Таким образом, необходима комбинированная модель, которая может объединить два подхода для описания: аддитивной марковской цепи, представленный выше, и цепи Маркова со ступенчатой функцией памяти (10).

Вопрос о том, какая часть корреляционной функции, или функции памяти, отвечает за инвариантность относительно процедуры прореживания, остаётся открытым.

**Заключение и перспективы**

1. В настоящей статье мы рассмотрели бинарные случайные последовательности, однако практически все результаты могут быть обобщены на небинарные последовательности и применены к естественным письменным текстам (содержащим примерно 30 символов) и ДНК последовательностям, «написанным» на черырехсимвольном алфавите нуклеотидов.

2. Мы предполагали, что корреляции являются слабыми. Однако наше предварительное рассмотрение показывает, что при $L\to\infty$ сильные ближние корреляции изменяют в (25) множитель перед членом $\sum_{r=1}^{L} K^2(r)$.

3. Длина памяти может быть бесконечной. В этом случае мы должны наложить условие на скорость убывания функции условной вероятности.

**Благодарности**

Мы выражаем благодарность В.А. Ямпольскому и С.В. Денисову за полезное и плодотворное обсуждение.

1. P. Ehrenfest, T. Ehrenfest. Encyklopädie der Mathematischen Wissenschaften, Springer, Berlin (1911), 742p.
2. D. Lind and B. Marcus. An Introduction to Symbolic Dynamics and Coding, Cambridge University Press (1995), 499p.
3. F.M. Izrailev, A.A. Krokhin, N.M. Makarov. Physics Reports, 512, 3, 125 (2012).
4. C. E. Shannon and W. Weaver. The Mathematical Theory of Communication, University of Illinois Press, Urbana, IL (1949), 117p.
5. A. N. Shiryaev, Probability, Springer, New York (1996), 621p.
6. S. S. Melnyk, O. V. Usatenko, V. A. Yampol'skii. Physica A, 361, 405 (2006).
7. O. V. Usatenko, S. S. Apostolov, Z. A. Mayzelis, and S. S. Melnik. Random finite-valued dynamical systems: additive Markov chain approach, Cambridge, Cambridge Scientific Publishers (2009), 171p.
8. S. S. Melnyk, O. V. Usatenko, V. A. Yampol'skii, V. A. Golick. Phys. Rev. E., 72, 026140 (2005).
9. C. W. Gardiner. Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences, Springer Series in Synergetics, v. 13, Springer-Verlag, Berlin ; New York (1985), 442p.
10. S. V. Denisov, S. S. Melnik, A. A. Borisenko, O. V. Usatenko, and V. A. Yampolsky. Entropy of complex symbolic sequences: Exact results, to be published.
11. O. V. Usatenko, V. A. Yampol'skii. Phys. Rev. Lett., 90, 110601 (2003).
12. S. S. Apostolov, Z. A. Mayzelis, O. V. Usatenko, V. A. Yampol'skii. Int. J. Mod. Phys. B,. 22, 3841 (2008).
13. W. Ebeling and G. Nicolis. Europhys. Lett., 14, 191 (1991).
14. T. Schürmann and P. Grassberger. Chaos, 6, 414 (1996).