Fractal Binary Sequences: 
Tsallis Thermodynamics and Zipf Law.

S. Denisov  
Department of Mathematics, Institute for Electromagnetic Research,  
P. O. Box 10857, 310022 Kharkov, Ukraine.

Аннотация

It is pointed out that the generalized thermodynamics introduced by Tsallis [13]) provides a natural frame for extension of thermodynamical formalism of symbolic dynamics. The statistical structure of 1D fractal long-range correlation sequences is investigated using n-tuple Zipf analysis. The generalized form of Zipf-Mandelbrot law can be obtained using maximum entropy principle within the frame of Tsallis thermodynamics, if the frequency of a word is associated with probability and the rank with the energy.

In recent years much attention is paid to physical systems with the properties intermediate between deterministic predictability and complete chaos. Characteristic distinction of such systems is the temporal and spatial long-range power law correlations, e.g. the processes of self-organized criticality [1], anomalous diffusion [2], strange kinetics [3], fractals [4], DNA sequences [5] and texts in natural and artificial languages [6].

The efficient method of investigation of such systems is based on the technique of symbolic dynamics. By using decomposition of phase space (or space of states) into a finite number of partitions and labeling each partition with a definite symbol (a letter from same alphabet), we can pass from the examination of real trajectories in phase space to the corresponding symbolic sequences. Each of that sequences can be treated as certain "text" with certain statistical structure. The statistical peculiarities of such "text" are reflecting robust invariant properties of corresponding dynamics to include the character of correlations. Thus, the problem can be formulated as follows: to characterize and evaluate the symbolic sequence global properties such as correlations and statistical structure.

Retaining the generality the current study addresses the case of binary sequences, i.e. sequences with alphabet \{0,1\}. Recently the statistical structure of binary sequences with short-range (exponential law) markovian and long-range correlations has been investigated using an n-tuple Zipf analysis [7]. This method is based on calculations of the normalized frequency \( \omega(R) \) of "word \( n \)-tuple" as a function of its rank \( R \). In event of long-range correlation sequences in a wide range of \( R \) (except values close to borders \( R = 1 \) and \( R = 2^n \)) the normalized frequency \( \omega(R) \) has been found to decrease with rank \( R \) approximately as a power law:

\[
\omega \sim R^{-\xi}
\]

(1)

where \( \xi \) has been termed Zipf exponent. It can be found experimentally as the inclination coefficient of Zipf plot (log-log plot \( \omega \) vs \( R \)). Also the simple relation approximately linear, between the exponent \( \alpha \) [7], characterizing long-range correlations and Zipf exponent \( \xi \), has been found. Short-range correlation Markovian sequences did not show power-law Zipf properties and numerical results seem to suggest an exponential decrease of \( \omega(R) \) with respect to rank \( R \).

From thermodynamical standpoint a binary sequence can be treated as classical one-dimensional lattice gas, where "0" corresponds to vacancy and "1" corresponds to atom occupied position. In certain conditions the problem of statistical structure and correlations can be solved by using the concepts of "interaction" and "energy" within the framework of Gibbs-Boltzmann classical thermodynamics. This constituted the base of thermodynamical formalism of Sinai-Bowen-Ruelle (SBR-formalism) [8].

The energy of state \( x_0 \) of infinite binary sequence \( X = \ldots x_{-2} x_{-1} x_0 x_1 x_2 \ldots = \{x_i\}_{i=-\infty}^{\infty}, x_i \in \{0,1\} \), i.e. the contribution to the whole energy of "system"-sequence connected with appearance of state \( x_0 \) can be introduced by means of interaction potential:

\[
\varphi(x_0) = \varphi(X) = U(x_0, x_1, x_{-1}, x_2, x_{-2}, \ldots, x_n, x_{-n}, \ldots)
\]

(2)

The character of decrease of potential \( U \) is closely related to the nature of correlations. Thus the following series can be written \( \alpha = \{\alpha_n\}_{1}^{\infty}, \alpha_n \to 0, \)
Fractal Binary Sequences: Tsallis Thermodynamics and Zipf Law.

It means that the value of interaction can change no more than on $2C\alpha_n$ with arbitrary variations of variables $x_m, |m| \geq n + 1$. In event of short-range correlations $\alpha_n = \rho^k, 0 \leq \rho < 1$, while in event of long-range correlations $\alpha_n = n^{-s}, s > 0$. The existence of invariant statistical structure of a sequence means that condition (3) must be obeyed for any $x_i$ which is identical to the requirements of translation-invariant structure of equilibrium state.

There are two requirements to admissible sequences in the frame of SBR-formalism: (a) $\sum_{n=1}^{\infty} \alpha_n < \infty$ and (b) $\sum_{n=1}^{\infty} n\alpha_n < \infty$. The sequences satisfying both conditions are, by Ruelle, the systems "in which the interactions are such that the system display the thermodynamical behavior" [9]. Equilibrium states for this systems are constructively defined and their structure in thermodynamical limits ($n \to \infty$) is determined by Gibbs distributions:

$$p_j = \frac{e^{-\beta E_j}}{\sum_{j=1}^{\infty} e^{-\beta E_j}}$$

(4)

The existence and uniqueness of distribution are guaranteed by conditions (a) and (b) respectively. Such sequences can be viewed as "canonical ensembles" of "microsystems" $n$-tuples.

In case of unabised Markovian sequences the energy spectrum is completely defined by values of conditional probabilities $\varrho(0,0) = \varrho(1,1) = p$ and $\varrho(0,1) = \varrho(1,0) = q$ with $p + q = 1$. The frequency $\omega$ of a given $n$-tuple only depends on the number $k$ of consecutive digit pairs in that word with both digits different: $\omega = 5^k p^n q^{n-k}$. One can derive the statistics (4) by defining the energy spectrum as follows: $E_k = -\log q^k p^{n-k}$ which corresponds to the nearest-neighbor interaction with the pair potential:

$$U(x_i, x_{i+1}) = -\log(x_i, x_{i+1}).$$

(5)

In this case energy is an additive value and obviously, binary Markovian sequences are the best examples of "thermodynamical" sequences. Fig. 1 depicts typical Zipf plot of Markovian sequence. The steplike character of plot occurs because of strong degeneracy of energy spectrum.

There are three different situations for the case of long-correlation sequences in accordance to conditions (a) and (b) fulfillment. If $s > 2$ then corresponding sequence is "thermodynamical" and admissible in the framework of SBR-formalism. If $1 < s < 2$ then the condition (b) is not valid. This case has been studied by Dyson [10]. It has been shown, in particular, that in this case, in contrast to the "thermodynamical" sequence case, it is the phase transition of some kind connected with appearance of long order is possible.

In the most interesting third case $s \leq 1$ both conditions are not valid. Interactions have nonintegrable form $U(r) \sim r^{-s}, s \leq 1$ and corresponding sequences are not "thermodynamical systems" in classical sense. There are close relations between such sequences, processes of anomalous diffusion and fractal curves. This relations can be realized by using procedure of binary sequence mapping into a one-dimensional random walk (RW - algorithm) [11] and can be characterized by correlation exponent $\alpha$ [12]. In addition, the dimension of corresponding self-affine fractal curve is [4]:

$$D = 2 - \alpha.$$  

(7)

For $\alpha$ close to 0.5 there is no correlation or only short-range correlation in the sequences. If $\alpha$ deviates essentially from 0.5 then it indicates long-range correlations with $s \approx 2 - 2\alpha$. The existence of statistical convergence to non-Gibbs' distribution (1) on the one hand and "anomalous" character of long-correlation sequences on the other, allows to suggest that a natural frame for description of statistical structure is provided by generalized statistical thermodynamics (GST) proposed by Tsallis [13]. Recently this formalism has been used for the construction of thermodynamical theory of anomalous diffusion [14] and fractal random walk [15].
Рис. 2. Zipf plot of a fractal binary sequences with $n = 12$: $\alpha = 0.9(D = 1.1)$. The straight dashed line corresponds to the power-law function $\omega(R) \sim R^{-\xi}, \xi = 2\alpha - 1$.

GST consists in alternative defining the equilibrium entropy of a system whose $i$th microscopic state has probability $p_i$ as

$$S_q = k \left(1 - \frac{\sum p_i^q}{q - 1}\right),$$

(8)

where $k$ is a positive constant and the index $q$ defines a particular statistics. For $q \to 1$ and $k = k_b$ the usual Boltzmann definition is reobtained. Applying the maximum entropy principle to $S_q$ under conditions of probability normalization, i.e. $\sum p_i = 1$ and generalized form of energy average, $\sum E_i p_i^q = E_q$, we get the following form of distribution:

$$p_i = p(E_i) = \frac{1}{Z_q} \left[1 + \beta(q - 1)E_i\right]^{-\frac{1}{q-1}},$$

(10)

where $Z_q = \sum \left[1 + \beta(q - 1)E_i\right]^{-\frac{1}{q-1}}$ and $\beta$ designate positive constant and play the role of Lagrange multiplier. Note that $q$ must be greater or equal than unity. For $q = 1$ we get the Gibbs distribution (4) that corresponds to case of short-range correlations.

Assume that the energy spectrum is increasingly ordered. In this case the distributions (10) coincides with Mandelbrot’s generalization of Zipf law [4]

$$\omega(R) = \frac{A}{(D + R)^\xi},$$

(11)

if we associate $\omega$ with probability $p$ and put $E_i \propto R_i$, where $\xi = \frac{1}{q-1}$.

However in this case we encounter certain problems. Namely, how we can find the energy of $n$-tuple in the long-correlation sequence avoiding the divergence and to obtain precise formula for energy spectrum as in Markovian sequence case. However, this difficulties may be bypassed supposing that for sufficiently large $n$ (such that $N = 2^n \gg 1$) we can replace the discreet spectrum with continuous one.

We generate the long-range correlation fractal sequences by using the procedure of consequent random additions, proposed by Voss [16]. This procedure generates the parametrized fractal curves

П. 3. Zipf plot of a fractal binary sequences with $n = 12$: $\alpha = 0.7(D = 1.3)$. The straight dashed line corresponds to the power-law function $\omega(R) \sim R^{-\xi}, \xi = 2\alpha - 1$.

П. 4. Zipf plot of a fractal binary sequences with $n = 12$: $\alpha = 0.5(D = 1.5)$. The dashed line corresponds to Markovian approximation with $\varphi(1, 1) = 0.545$. Similar behaviour is obtained when $n = 8, 10$ and 14.
Fractal Binary Sequences: Tsallis Thermodynamics and Zipf Law.

which can be considered also as the trajectories of fractional Brownian motion (fBm). The first step is introduction of the real coordinates $y(t_i), t_i = 0, \frac{1}{2}, 1$, where $t_i$ - parameter (time in the case of fBm). The values of coordinates are determined as a random numbers chosen from the normal distribution with zero average and and dispersion $\sigma^2 = 1$. The intermediate values $t_i = \frac{1}{4}, \frac{3}{4}$ is considered as the additional points; the values of coordinates in them are calculated using the linear interpolation. At the second step $t_i = 0, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, 1$ and the random numbers with dispersion $\sigma^2 = (\frac{1}{2})^{2n}$ add to the corresponding coordinates. Iterating this procedure $n$ times we obtain $(1 + 2^n)$ values of $y(s_i)$. The dispersion of $n$-th generation terms is equal $\sigma^2 = (\frac{1}{2})^{2n}$. Fractal dimension of self-affine curves obtained is equal to $D = 2 - \alpha$ and dispersion of corresponding fBm is $\langle (y(t) - y(0))^2 \rangle > \propto t^{2\alpha}$. From the sequence of real numbers $y(t)$ sampled with interval $\Delta s$ we obtain corresponding binary sequence $x_i$ using inverse RW-algorithm:

$$x_i = \begin{cases} 1 & \text{if } y((i+1)\Delta t) \geq y(i\Delta t), \\ 0 & \text{if } y((i+1)\Delta t) < y(i\Delta t). \end{cases} \quad (12)$$

We have examined sequences with length $L = 10^7$ bits and number of steps of Voss procedure $M = 18$. It should be mentioned, however, that linear character of Zipf plot exhibits starting from the length of order $10^4, M = 12$ and uncertainty of $\xi$ is smaller than 3% for $L > 10^5$ and $M > 16$.

The Zipf plot of fractal sequences for various values of $\alpha$ is shown at Fig.2. For $\alpha \to 0.5$ we get the distributions corresponding to the absence of long-range correlations which are well described by Markovian approximations. Numerical studies show that in a wide range of values the relationship between $\alpha$ and $\xi$ is described with fairly high accuracy by

$$\xi = 2\alpha - 1 = 3 - 2D \quad (13)$$

Long-range correlation sequence (12) has a scale-invariant structure due to self-affinity of corresponding fractal curves. The statistical properties of such sequences are invariant with respect to coarse-graining procedure:

$$\{111,110,101,011\} \to 1, \{000,001,010,100\} \to 0,$$

as it corresponds to transition $\Delta t \to 3\Delta t$ for fractal curves. From thermodynamical standpoint it means the homogeneity of thermodynamical functions such as entropy and free energy. Hence the equilibrium distribution (10) is invariant with respect to renormalization procedure. Figure 3 shows Zipf plot of fractal sequence after few numerical coarse-graining procedure.

The formalism of generalized thermodynamics is closely related to the theory of multifractal measure $[17]$ and Tsallis entropy and can be associated with the fractal dimension spectrum $D_q$. From this point of view fractal binary sequence do "support" the measure, which defined on the set of all possible $n$-tuples. However it should be noted that the meaning of exponent $q$ is quite different from momentum order: it is not free but characteristic parameter.

All above is in agreement with Mandelbrot hypothesis that the fulfillment of Zipf law is related to the hierarchical structure of language [18]. Despite that the Markovian sequences can mimic Zipf law with small $\xi$ which changes with $R$ being quite slow [19], the fundamental difference of long-range correlation sequences from short-range correlation ones is their nonextensive character. Interactions in those sequences are such that they are structurally ordered units. This is closer to the properties of real texts, such as DNA molecules and texts in natural or artificial languages.

The construction of rigorous formalism can provide insight in a number of aspects of statistical properties and nature of systems with complex hierarchical structure.

The author wishes to thank Professor Oleg V. Usatenko (Kharkov State University) for valuable discussions.
Список литературы


