On a certain misunderstanding in the interpretation of the Canonical distribution of the symbol sequences

Ilya Shesterikov

December 9, 2025

Abstract

The canonical (Gibbs) distribution is widely used in statistical physics to describe the probabilities of microscopic states characterized by an energy value. In symbolic dynamics and the study of symbolic sequences generated by nonlinear dynamical systems, an analogous construction is frequently applied: the probability of observing a particular symbol sequence is assumed to depend exponentially on an associated "energy", often defined through a cylinder length or a Jacobian-based quantity. While this analogy is technically appealing and mathematically consistent, it has led to a persistent conceptual misunderstanding. The confusion arises when the discrete cylinder lengths (ℓ_i) are mistakenly interpreted as samples from a continuous distribution, leading to the use of probability density functions where **only discrete probabilities** are appropriate. In this paper, we analyze the origin of this misunderstanding, clarify the correct interpretation of the canonical distribution in symbolic dynamics, and provide practical guidance for avoiding associated pitfalls. We further illustrate the issue with examples, graphical explanations, and a discussion of implications for numerical studies of chaotic systems.

1 Introduction

Symbolic dynamics offers a compact and powerful framework for studying the structure of complex dynamical systems by encoding trajectories as sequences of discrete symbols. This work utilizes the fully chaotic **Logistic map** as a foundational model. For a given dynamical system equipped with a generating partition, each finite sequence of length (N) corresponds to a unique cylinder set in phase space. The size or "length" of this cylinder set, denoted (ℓ_i), reflects local properties of the dynamics such as stretching rates or Jacobian factors along the trajectory associated with that sequence. In systems with sufficient hyperbolicity, these lengths typically admit an exponential representation of the form

$$\ell_i \sim \exp(-\epsilon_i) \tag{1}$$

,

where the quantity (ϵ_i) plays a role analogous to an energy. This analogy becomes even more explicit when the natural invariant measure is written in the familiar Gibbs form

$$P_i = \frac{\ell_i}{\sum_j \ell_j} = \frac{\exp(-\epsilon_i)}{Z}, \tag{2}$$

where (Z) is the corresponding normalization factor. **Because the exponential dependence emerges so naturally, the theoretical framework used in symbolic dynamics is frequently phrased in the language of the canonical ensemble from statistical mechanics.**

Despite the mathematical consistency of this analogy, it has led to a conceptual misunderstanding that persists throughout the literature. The confusion begins with the observation that, for a sequence length (N), there are (2^N) possible sequences. When one computes and plots the corresponding cylinder lengths (ℓ_i), the resulting collection of values often appears to follow a continuous distribution: histograms of (ℓ_i) may resemble a Gaussian, log-normal, or heavy-tailed or some irregular distribution, depending on the system and the symbolic partition. This visual resemblance tempts researchers to treat the set of lengths as if it were produced by a continuous random variable, implicitly assuming the existence of a probability density $(\rho(\ell))$. Such an assumption leads naturally to expressions of the form

$$\int \rho(\ell) d\ell = 1, \tag{3}$$

and encourages interpretations grounded in continuous probability theory.

However, this line of reasoning is fundamentally incorrect. The cylinder lengths (ℓ_i) do not arise as random samples drawn from a continuous distribution. Each value (ℓ_i) is a discrete, deterministic quantity associated with a specific symbol sequence. The apparent smoothness of histograms does not justify the introduction of a probability density; it simply reflects the large number of discrete combinatorial objects when (N) is moderately large. The proper probabilistic structure of the system is fully discrete, and the correct normalization is

$$\sum_i P_i = 1, \tag{4}$$

not an integral over a density. Treating (ℓ_i) as if it had an underlying density obscures the fact that symbolic sequences form a finite set and that their measures are assigned individually through the dynamics of the system, not through sampling theory.

The misunderstanding is reinforced by the superficial similarity between the exponential representation ($\ell_i \sim \exp(-\epsilon_i)$) and the energy dependence in canonical statistical mechanics. In physics, energy values are typically continuous, and thus probability densities are not only natural but necessary. In symbolic dynamics, by contrast, the exponential dependence simply expresses the rapid decrease in cylinder size with increasing sequence length; it does not imply that (ϵ_i) varies continuously. Yet numerical simulations—through histograms and smoothed plots—often blur this distinction and unintentionally encourage a continuous interpretation.

Recognizing this distinction is essential. The canonical distribution in symbolic dynamics is a **discrete** probability distribution defined over a finite set of sequences; it is not a continuous distribution over the cylinder lengths themselves. Any analysis invoking probability densities, differential normalization conditions, or distributional "shapes" of (ℓ_i) risks introducing conceptual errors. By properly acknowledging the discrete structure of the problem, one preserves both the mathematical integrity of the thermodynamic formalism and the conceptual clarity needed to interpret numerical data correctly.

2 Background: Logistic Map, Symbol Sequences, Cylinder Sets, and Measures

The Logistic map provides an ideal and straightforward paradigm for developing foundational insights into the study of chaotic phenomena. This map is defined by the following recurrence relation:

$$x_{n+1} = r x_n (1 - x_n) \tag{5}$$

The map defines a process of iteration, where the output of the function at step n becomes the input for step $n + 1$. The sequence of values generated, $\{x_0, x_1, x_2, \dots\}$, is called an ***orbit***. State Variable (x_n) - This is the variable whose value evolves over time. Mathematically, it is a real number restricted to the unit interval $I = [0, 1]$. This restriction is necessary to ensure that the next value, x_{n+1}, also remains within $[0, 1]$ when the parameter r is in its relevant range $[0, 4]$. The function $f_r(x)$ is a quadratic polynomial, making the system nonlinear. This nonlinearity is the source of its complex dynamics. Control parameter (r) is a positive real number, typically restricted to $r \in [0, 4]$. It acts as the "tuning knob" that determines the dynamical properties of the entire system.

To apply symbolic dynamics, the continuous state space (the interval $[0, 1]$) is divided into a finite number of regions, and each region is assigned a **unique symbol**. An orbit, which is a sequence of real numbers $x_0, x_1, x_2, \dots$, is then converted into a **symbolic sequence** by recording the symbol of the region the trajectory visits at each time step.

A partition is classified as a **generating partition** if the correspondence between the infinite symbolic sequence and the original orbit is **one-to-one**.

This means that studying the symbolic sequence alone can reveal the essential topological and measure-theoretic properties of the chaotic dynamical system. For one-dimensional maps like the Logistic Map, the boundary points of the partition are determined by the critical points and their preimages under the map.

For the fully chaotic Logistic Map with $r = 4$ (i.e., $x_{n+1} = 4x_n(1 - x_n)$), the state space $I = [0, 1]$ has a single critical point (the maximum of the parabola) at $x_c = 0.5$. The simplest and most fundamental **generating partition** is the **binary partition** defined by this critical point:

- **Region 0 (Symbol L or 0)**: $I_0 = [0, 0.5)$
- **Region 1 (Symbol R or 1)**: $I_1 = [0.5, 1]$

The critical point $x_c = 0.5$ is the *partition boundary.*

The symbolic sequence $S = s_0 s_1 s_2 \ldots$ corresponding to an orbit $x_0, x_1, x_2, \ldots$ is constructed using the following rule:

$$s_n = \begin{cases} 0 \text{ (or L)} & \text{if } x_n \in [0, 0.5) \\ 1 \text{ (or R)} & \text{if } x_n \in [0.5, 1] \end{cases}$$

For example, if a trajectory is $x_0 = 0.2, x_1 = 0.64, x_2 = 0.92, x_3 = 0.29, \ldots$, its symbolic sequence would be $0, 1, 1, 0, \ldots$.

Consider a dynamical system with a symbolic encoding defined by a generating partition. For a sequence of length (N), there are (2^N) possible binary words:

$$s_i = (s_{i1}, s_{i2}, \ldots, s_{iN}), \quad s_{ij} \in 0, 1. \tag{6}$$

Each sequence corresponds to a cylinder set in phase space, whose measure or length (ℓ_i) depends on the dynamics. For uniformly hyperbolic systems, this length typically relates to Lyapunov exponents or derivatives of the underlying map, such as:

$$\ell_i \approx \prod_{k=1}^{N} \frac{1}{|f'(x_{ik})|}. \tag{7}$$

Thus, cylinder lengths encapsulate the contraction or expansion along trajectories. Let the length be expressed as:

$$\ell_i = \exp(-\epsilon_i), \tag{8}$$

where (ϵ_i) plays the role of an energy. The probability of observing sequence (i) is determined by its expansion rate, or energy (ϵ_i):

$$P_i = \frac{\exp(-\epsilon_i)}{Z}, \quad Z = \sum_i \exp(-\epsilon_i). \tag{9}$$

This is fully analogous to the canonical ensemble of statistical mechanics. However, unlike physical energies, which vary continuously, the set (ϵ_i) is **finite and discrete**, corresponding to the **finite number of possible symbol sequences**.

3 Origin of the Misunderstanding

Because the number of sequences grows exponentially with (N), the set of values (ℓ_i) becomes large. When plotted as a histogram, the values may visually appear to follow some continuous distribution (Gaussian-like, log-normal, or heavy-tailed).

Such a histogram might for instance resemble the Figure 1. This figure illustrates how the set of cylinder lengths (ℓ_i) often appears when plotted as a histogram for a moderately large sequence length (N). The data shown in this figure are synthetic and serve only to demonstrate the typical visual appearance of cylinder-length histograms. Although each (ℓ_i) is a discrete quantity associated uniquely with a particular symbol sequence, the large number of possible sequences—(2^N) in the case of binary dynamics—produces a visual distribution that resembles a smooth continuous curve. In this example, the histogram of (ℓ_i) values takes on a shape similar to a Gaussian profile, which may create the misleading impression that the system generates (ℓ_i) as samples from an underlying continuous probability distribution. The figure therefore highlights the key source of the common misunderstanding: while the graphical representation looks statistically smooth, the underlying objects are entirely discrete, and no probability density $(\rho(\ell))$ exists in a rigorous sense. The smooth envelope simply reflects the combinatorial abundance of cylinder sets of differing sizes and should not be interpreted as evidence for a true continuous distribution.

This visual similarity to a standard continuous distribution invites the false conclusion that there exists a **probability density** $(\rho(\ell))$. But such a density would require (ℓ) to be drawn from a continuous random variable, which is not the case. The misunderstanding arises because in thermodynamics, energies are usually continuous, but in symbolic dynamics the analogous quantity is tied to a finite combinatorial set. This confusion leads to inconsistencies in interpretation and normalization. The correct normalization for discrete probabilities is:

$$\sum_i P_i = 1. \tag{10}$$

But if one incorrectly assumes a density, one may write:

$$\int \rho(\ell) d\ell = 1, \tag{11}$$

which is mathematically inconsistent with the discrete nature of the symbol sequences.

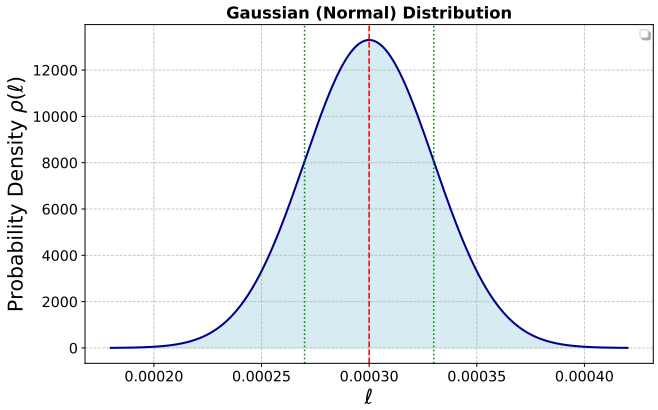

Figure 1: Histogram of cylinder lengths (ℓ_i) for symbol sequences of length (N). This figure is an artificial example included solely for illustrative purposes. Although the individual values of (ℓ_i) are discrete and deterministically assigned to specific symbol sequences, the large number of possible sequences produces a visually smooth distribution that might, for instance, resembles a Gaussian curve. This apparent smoothness should not be interpreted as a continuous probability density but arises solely from the combinatorial structure of the symbolic dynamics.

4 Correct Interpretation of the Canonical Distribution and Its Consequences

The correct interpretation of the canonical distribution in symbolic dynamics begins with recognizing that cylinder lengths are fundamentally **discrete** quantities. Each (ℓ_i) is tied to a specific symbol sequence and should therefore be seen as an indexed value, not as a sample drawn from an underlying continuous distribution. This distinction is central: a continuous probability density function would require (ℓ) to be a variable that assumes values over an interval, but in symbolic dynamics the number of possible cylinder lengths is finite and determined entirely by the finite set of symbol sequences of length (N). Thus, the notion of a density $(\rho(\ell))$ does not apply in this setting. What we have instead is a discrete family of probability values (P_i), each of which corresponds exactly to one sequence. The corresponding probability values is written in the familiar Gibbs form:

$$P_i = \frac{\ell_i}{\sum_j \ell_j} = \frac{\exp(-\epsilon_i)}{Z}, \tag{12}$$

This observation leads directly to the correct form of normalization. While a continuous distribution must satisfy an integral condition over an interval, the canonical distribution of symbol sequences normalizes via a simple sum over all sequences, $(\sum_i P_i = 1)$. Any attempt to replace this with an integral over (ℓ)

implicitly assumes a continuous structure that is absent. Despite this, confusion often arises due to the way cylinder lengths appear in numerical experiments: when many values of (ℓ_i) are plotted in a histogram, the resulting picture may resemble a continuous distribution such as a Gaussian or log-normal curve. This visual impression can be misleading. A histogram does not reveal a true underlying probability density; it merely shows how the discrete values are spread across bins. The apparent smoothness comes from the fact that for large (N), the number of distinct values becomes very large, **which is a combinatorial property of the symbolic space rather than evidence of an underlying continuous variable**. Since each cylinder length may be written as $(\ell_i = \exp(-\epsilon_i))$, the probability of each sequence assumes the Gibbs-like form $(P_i \propto \exp(-\epsilon_i))$. This dependence is exact and does not rely on any assumptions about continuity. Graphical representations of this relationship—such as plots of (P_i) against (ℓ_i) or of (P_i) against $(-\epsilon_i)$—simply illustrate how probabilities scale with the corresponding "energies", and these plots should always be interpreted as representations of discrete objects.

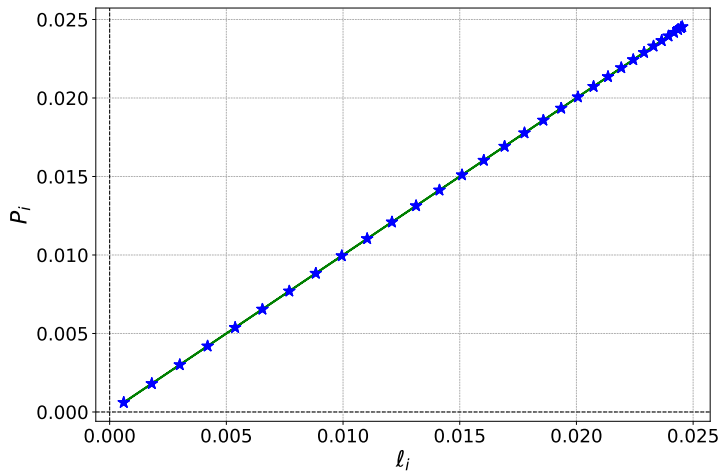

Figure 2: Discrete probabilities (P_i) plotted as a function of their corresponding cylinder lengths (ℓ_i). The linear trend reflects the relation $(P_i \propto \ell_i)$, which follows directly from the definition of the canonical measure. Each point corresponds to a distinct symbol sequence, highlighting that the probability distribution is discrete and not a continuous function of (ℓ).

Figure 2 depicts the discrete probability values (P_i) plotted against their corresponding cylinder lengths (ℓ_i). Unlike Figure 1, which presents a histogram of the distribution of lengths, this figure shows the direct linear relationship $(P_i \propto \ell_i)$ that arises from the normalization definition of the invariant measure. Each point represents one symbol sequence, and the vertical axis displays the probability assigned to that specific sequence. The plot emphasizes that (P_i)

is fundamentally a discrete probability associated with a single sequence rather than with an interval of (ℓ). The scatter of points does not form a continuous curve but instead reflects the discrete nature of the symbolic space. This visualization reinforces the central argument of the paper: although cylinder lengths may appear to form a smooth distribution in aggregate, the measure constructed from them is purely discrete and assigns probability mass only to individual sequences.

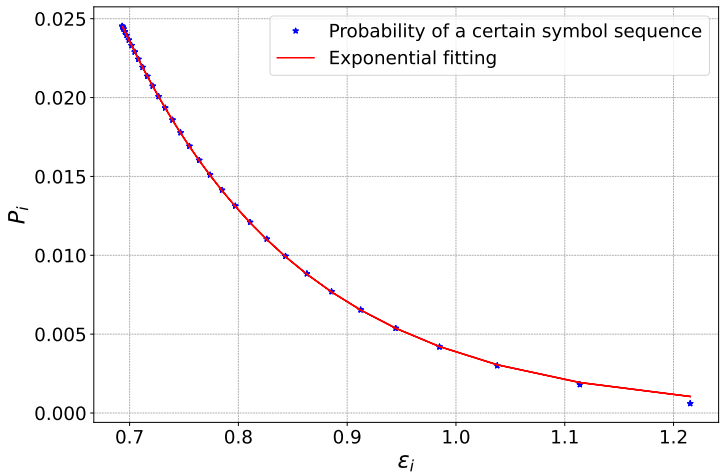

Figure 3: Canonical distribution of probabilities (P_i) plotted against the corresponding energy values ($\epsilon_i = -\ln \ell_i$). The exponential decay visible in the plot reflects the Gibbs-type relation ($P_i \propto \exp(-\epsilon_i)$). Each point represents a single symbol sequence, underscoring the discrete nature of the probability assignment even though the overall trend resembles the continuous Boltzmann distribution.

Figure 3 presents the canonical distribution in its exponential form by plotting the discrete probabilities (P_i) against the associated "energies" ($\epsilon_i = -\ln \ell_i$). When expressed in this representation, the expected exponential decay becomes evident: the probability of a symbol sequence decreases exponentially with increasing energy, mimicking the structure of the Boltzmann–Gibbs distribution familiar from statistical mechanics. The data points align closely with an exponential fit, demonstrating the consistency of the thermodynamic formalism when applied to symbolic dynamics. Importantly, the figure again illustrates that each plotted value corresponds to a single symbolic sequence. The exponential envelope that emerges is not a continuous probability density but a discrete mapping of sequences to probability weights derived from their cylinder lengths.

Understanding the discrete nature of the canonical distribution also clarifies several issues encountered in numerical and analytical studies. In computational work, it is common to estimate the distribution of cylinder lengths using his-

tograms. While histograms are useful visualization tools, they should never be interpreted as revealing a probability density function for (ℓ). If one mistakenly assumes such a density exists, the subsequent analysis may involve integrals instead of sums, leading to inconsistencies in normalization or even incorrect conclusions about the presence of tails or specific functional forms in the distribution. Such artifacts often arise solely from the choice of bin size or from the smoothing inherent in numerical plotting, not from any actual structural property of the dynamical system.

The misinterpretation also has implications for entropy calculations and for the thermodynamic formalism. Entropy in symbolic dynamics is a combinatorial concept: it reflects the exponential growth rate of the number of admissible sequences and the distribution of their corresponding probabilities. Treating (ℓ_i) as a continuous variable can distort estimates of quantities such as Rényi entropies or the topological entropy, especially when numerical methods are used. Theoretical constructs from thermodynamics remain valid, but only when applied within the proper discrete framework. For example, although $(-\ln(\ell_i))$ may appear approximately Gaussian for large (N)—a consequence of central-limit-like behavior in chaotic systems—this does not imply that (ℓ_i) itself should be treated as a random variable with a continuous density.

Finally, distinguishing correctly between discrete and continuous interpretations improves the reliability of algorithms used to approximate invariant measures or equilibrium states in symbolic and chaotic systems. Numerical methods often rely on discretization or coarse-graining, and the temptation to interpret histograms as densities can obscure the fundamentally combinatorial nature of the underlying objects. By maintaining a clear conceptual separation, one avoids misinterpretations and ensures that numerical results remain consistent with the mathematical structure of symbolic dynamics and with the proper use of the canonical distribution.

5 Conclusion

The canonical distribution has long served as a conceptual bridge between statistical physics and symbolic dynamics, providing a formally elegant way to assign probabilities to symbol sequences based on an energy-like quantity. Yet this analogy works correctly only when one preserves the fundamental distinction between discrete and continuous structures. The frequent misinterpretation of cylinder lengths as samples drawn from a continuous probability distribution illustrates how easily this distinction becomes blurred, especially in numerical studies. Large collections of discrete values often resemble smooth, continuous histograms, and their visual similarity to classical probability densities can tempt researchers into applying inappropriate continuous formalism.

Recognizing that cylinder lengths (ℓ_i) are discrete objects tied to specific symbolic sequences eliminates this ambiguity. Each value has meaning only within the combinatorial framework of symbolic dynamics, and the associated probabilities must therefore be treated as a discrete set (P_i), normalized by a

finite sum rather than an integral. The fact that these discrete probabilities may mimic the exponential form familiar from the Gibbs distribution does not imply the existence of a genuine probability density over (ℓ); rather, it reflects the underlying dynamical structure and the exponential sensitivity characteristic of chaotic systems.

Clarifying this conceptual point has practical consequences. It prevents erroneous normalization procedures, avoids misleading interpretations of histogram shapes, and ensures that entropy estimates and other quantities derived from thermodynamic formalism remain mathematically meaningful. Importantly, this clarification does not diminish the value of the thermodynamic analogy. Instead, it strengthens its foundation by placing it on correct mathematical footing.