Statistical Physics of Coding for the Integers

Neri Merhav

The Andrew & Erna Viterbi Faculty of Electrical Engineering, Technion, Haifa 32000, Israel.
E–mail: merhav@ee.technion.ac.il

Abstract. We study a paradigm of coding for compression of the natural numbers via the zeta distribution and develop a statistical-mechanical interpretation, both in terms of Hagedorn systems and a Bose gas with energy levels given by logarithms of prime numbers. We also propose a simple coding scheme for the zeta distribution that nearly achieves the ideal code length. For block coding of vectors of natural numbers, we derive the micro-canonical entropy function and demonstrate its asymptotic linearity implying that its behavior is analogous to that of a Hagedorn system. We also derive the large deviations rate function, and provide a formula for the best coding parameter in the large deviations sense. We show that due the Hagedorn-type phase transition there is only partial equivalence of ensembles, due to the degeneration of the domain of the partition function.

Keywords: zeta distribution, Zipf law, Hagedorn system, phase transition, ensemble equivalence.

1. Introduction

One of the basic problems in information theory is the assignment of code lengths to elements of a countable set for the purpose of data compression, in particular, compact representation for all the positive integers by the shortest possible binary strings. Coding schemes for the integers play a central role in a variety of settings where the objects to be described are drawn from a countably infinite or a priori unknown domain. This situation arises ubiquitously in data compression, where algorithms such as the Lempel-Ziv algorithm in its various versions (LZ77 [1], LZ78 [2] and others) must encode dictionary indices, match lengths, and offsets that grow with the data, making fixed-length representations infeasible. More broadly, integer coding underlies universal modeling and inference: in the minimum description length (MDL) principle (see, e.g., [3] and references therein), model classes are penalized via the code-lengths of integer-valued parameters (Markov order, number of states, etc.), while in Kolmogorov complexity, self-delimiting descriptions require efficient encodings of lengths and indices. Practical systems likewise rely on compact representations of counts, gaps, and run-lengths in compressed indexes and streaming protocols. From a probabilistic perspective, integer codes also provide near-optimal representations for heavy-tailed distributions (e.g., Zipf-like laws to be discussed below), thereby bridging combinatorial structure and statistical modeling. Collectively, these applications highlight that efficient, prefix-free coding of the integers is not merely a technical convenience but a fundamental building block in information theory, compression, and learning.

Consider any uniquely decodable code that assigns code lengths $\ell(x)$ to the positive integers, $x = 1, 2, \ldots$, arranged in non-decreasing order, $\ell(1) \leq \ell(2) \leq \ell(3) \leq \cdots$. As will be shown below, in order that a the code for the natural numbers would be uniquely decodable, the code-length for the integer x is lower bounded by

$$\ell(x) \geq \log x. \tag{1}$$

Thus, even in the absence of any probabilistic assumptions, there is an intrinsic logarithmic growth of code length with the index, which cannot be avoided. This reflects a basic combinatorial constraint: describing larger and larger objects necessarily requires increasing resources.

A natural question which may arise at this point is whether this lower bound can be approached in a universal manner, without assuming any specific probability distribution over the integers. A classical answer was provided by Elias [4] who proposed several simple and efficient coding schemes which assign to each positive integer a binary representation whose length behaves as

$$\ell(x) = \log x + O(\log \log x). \tag{2}$$

These constructions achieve, up to lower-order corrections, the minimal growth compatible with the lower bound (11), at least asymptotically for large x. In this sense, they realize a nearly optimal way of encoding the integers in a distribution-free setting.

Once a probabilistic structure is introduced, the problem acquires an additional layer. If the integers are generated according to a given probability distribution, $\{P(x),\ x = 1, 2, \ldots\}$, then neglecting integer length constraints, the optimal code length is given by $\ell(x) = -\log P(x)$ [5]. With this fact in mind, it is readily observed that Elias coding for the integers corresponds to a probability assignment of the form

$$P(x) = \frac{1}{2^{\log x + O(\log\log x)}} = \frac{1}{x \cdot 2^{O(\log\log x)}}, \tag{3}$$

namely, a distribution that decays at a rate that is slightly faster than the rate of $\frac{1}{x}$ (see also [3]). But this extra speed of convergence beyond $\frac{1}{x}$ is inevitable since the sum of the infinite series $\sum_{x=1}^{\infty} \frac{1}{x}$ diverges, and so, a distribution proportional to $\frac{1}{x}$ cannot exist, unless truncated to a finite range, $x \in \{1, 2, \ldots, J\}$. A simple natural remedy, in the case of infinite support, is to introduce a power parameter $\beta > 1$, namely, to let $P(x)$ be proportional to $\frac{1}{x^\beta}$, whose sum is convergent, as will be discussed shortly.

Interestingly, distributions of this type are not merely of theoretical interest. They are closely related to empirical regularities such as Zipf's law (see, e.g., [6], [7] and references therein), according to which the frequency of an event is inversely proportional to its rank, and to its refinement, the Zipf–Mandelbrot law (see, e.g., [8]), which incorporates a shift parameter and possibly a power parameter, that is,

$$P(x) \propto \frac{1}{(x+\alpha)^\beta}, \qquad \alpha \geq 0,\ \beta > 1. \tag{4}$$

Such probability laws have been empirically observed across a wide range of systems, including natural language, biological data, and complex networks. From the coding perspective, they correspond to situations in which large integers occur with non-negligible probability, making logarithmic-length codes not only necessary but essentially optimal.

This class of heavy tailed probability distributions are referred to as *power law distributions*. A particularly important subclass of (4) is given by the special case of $\alpha = 0$, namely,

$$P_\beta(x) = \frac{x^{-\beta}}{\zeta(\beta)}, \qquad \beta > 1, \quad x = 1, 2, \ldots \tag{5}$$

where $\zeta(\beta)$ is the normalization constant,

$$\zeta(\beta) = \sum_{x=1}^{\infty} \frac{1}{x^\beta}, \tag{6}$$

which is the well known *Riemann zeta function*. This leads to code lengths of the form $\ell(x) \approx \beta \log x$ (for large x), which respect the fundamental logarithmic scaling dictated by counting, while introducing a parameter that controls the relative weight assigned to large integers. Baer [9], [10] suggested structured, efficient algorithms of coding for the zeta distribution. In Appendix A, we propose an alternative algorithm, which we believe is simpler and more explicit. In this context, it should be noted that Elias coding can be viewed as corresponding to a choice of β that depends on x, in particular, $\beta_x = 1 + O(\log(\log x)/\log x)$.

The normalization of the power-law distribution, which is the zeta function, $\zeta(\beta)$, can be naturally displayed as a partition function,

$$\zeta(\beta) = \sum_{x=1}^{\infty} e^{-\beta \ln x}. \tag{7}$$

Interpreting $\ln x$ as an Hamiltonian, $\mathcal{H}(x)$ associated with state x, this defines a statistical-mechanical model with an unbounded state space. According to this model, the energy, $\mathcal{H}(x) = \ln x$, associated with state x is proportional to the cost of its code length, and coding with respect to the distribution P_β corresponds to the canonical ensemble associated with $\mathcal{H}(x)$. The divergence of $\zeta(\beta)$ at $\beta = \beta_{\rm c} = 1$ signals a critical point at which the normalization breaks down, reflecting the overwhelming contribution of high-energy states.

In this paper, we first demonstrate that the above described statistical-mechanical model is actually analogous to a Hagedorn system [11], [12], [13], which is a system whose density of states grows exponentially with the energy, up to a possible sub-exponential multiplicative factor (see Appendix B for more detailed background). The above mentioned critical point $\beta_{\rm c} = 1$ below which $\zeta(\beta)$ diverges is exactly the Hagedorn inverse temperature associated with this model.

It is important to emphasize that this phenomenon is not a consequence of any specific interaction or "potential" in the usual sense, but rather of the relation between the indexing of states and their assigned energy. In our setting, the energy variable is naturally identified with the code length, which, in accordance with the fundamental counting constraint, grows logarithmically with the integer label, $\mathcal{H}(x) \sim \log x$. As a result, the number of states with energy about E scales as

$$|\{x : \mathcal{H}(x) \approx E\}| \sim \exp\{\beta_{\rm c} E\}, \tag{8}$$

i.e., the density of states grows exponentially with the energy. This exponential proliferation of states is precisely the mechanism underlying Hagedorn behavior, leading to a finite radius of convergence of the partition function and a critical point at which normalization breaks down. In other words, in contrast to more traditional settings, where such growth arises from intricate microscopic structure, here it is a direct consequence of the logarithmic scaling imposed by coding considerations. This provides a particularly transparent realization of a Hagedorn transition, rooted in the combinatorial structure of the integers.

Moreover, owing to Euler's product form of the zeta function, there is also another analogy, pertaining to the grand-canonical ensemble of the Bose gas, whose energy levels are $E_p = \ln p$, where p runs over all primes, i.e., $p = 2, 3, 5, \ldots$. As β descends towards $\beta_{\rm c} = 1$, the total number of bosons in this model goes to infinity.

All the above holds true even for a single 'particle' at state x. But to complete the picture, we also consider a system with N independent such particles, $(x_1, \ldots, x_N)$, and focus on its density of states for large N, showing that for large per-particle energy ϵ, it behaves roughly like $\exp\{N\beta_{\rm c}\epsilon\}$ (where again, $\beta_{\rm c} = 1$), and so, the Hagedorn system characteristics are of course manifested here too. It will be interesting to note that

here, there is partial equivalence between the canonical and the micro-canonical model, which holds only above the Hagedorn inverse temperature ($\beta > \beta_{\rm c}$). By contrast, if the support of P_β is truncated into a finite range, $x \in \{1, 2, \ldots, J\}$, then there is no longer a critical point and there is full equivalence between the ensembles.

Finally, from the data compression perspective, we also study the behavior of code lengths for sequences of independent samples. Our focus is on the large deviations properties of the total code length and on the interplay between the coding parameter and the statistics of rare events. We show that the optimal parameters governing these deviations are driven toward the critical point $\beta_{\rm c} = 1$, leading to asymptotic behavior analogous to that of systems with Hagedorn-type transitions.

These results provide a concrete setting in which fundamental ideas from information theory —- universal coding for natural numbers, heavy-tailed distributions, and large deviations—interact with concepts from statistical mechanics such as partition functions, criticality, and ensemble equivalence.

2. Background on Uniquely Decodable Data Compression Codes

A code for lossless data compression is a mapping between a finite or countable alphabet of symbols of the data to be compressed (e.g., latin letters, digits, or any other characters) into a set of binary strings, which may be of different lengths. For the data to be perfectly recoverable from its compressed representation, this mapping must be one-to-one, not only in the level of single symbols and their codewords, but also when the code is used repeatedly and the decoder receives the concatenation of the compressed binary strings corresponding to the successive data symbols, because the parsing of the codewords is not provided to the decoder.

A simple method to ensure this unique decodability property is to design the code by maintaining the prefix-free condition, namely, to keep the rule that no codeword would be a prefix of any other codeword. For example, if the alphabet is $\mathcal{X} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ and the corresponding codewords are '0', '10', and '11', the prefix condition holds, and there is only one way (and hence - the right way) to parse any given compressed bit-stream and thereby to decode it. But it is not necessary to maintain the prefix condition for the code to be uniquely decodable (UD), as there are UD codes which are not prefix-free. Whether the prefix condition is met or not, it is clear that a necessary condition for a code to be UD is that the codewords of the code are sufficiently long in some collective sense. To give this (admittedly vague) statement a clear significance, let us first define the *length function*, $\ell(x)$ ($x \in \mathcal{X}$) of the given code to designate the length (in bits) of the codeword for x. In the above example, $\ell(\mathrm{A}) = 1$, and $\ell(\mathrm{B}) = \ell(\mathrm{C}) = 2$. One of the fundamental principles in information theory is a necessary and sufficient condition for the existence of a UD code with a given length function, $\ell(x)$. This necessary and sufficient condition is the well-known Kraft-McMillan (KM) inequality (see, e.g., [5]):

$$\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1. \tag{9}$$

In the above example,

$$2^{-\ell(\mathrm{A})} + 2^{-\ell(\mathrm{B})} + 2^{-\ell(\mathrm{C})} = 2^{-1} + 2^{-2} + 2^{-2} = 1,$$

and so, the KM inequality is saturated in this case. Note that whenever the KM inequality is met with equality, the terms of the KM sum, $\{2^{-\ell(x)},\ x \in \mathcal{X}\}$ can be thought of as probabilities, $\{P(x),\ x \in \mathcal{X}\}$, as they are non-negative reals that sum to unity. Conversely, every probability distribution with dyadic probabilities (integer powers of $\frac{1}{2}$) correspond to a length function, $\ell(x) = -\log P(x)$ of some UD code. Thus, lossless compression and probability assignment are two sides of the same coin.

Consider now the case where the alphabet is $\mathcal{X} = \mathcal{N} \equiv \{1, 2, 3, \ldots\}$, namely, the set of all natural numbers, and we wish to design a data compression code for $\mathcal{N}$. It is conceivable that large natural numbers would need to be mapped to longer codewords than those of small natural numbers, and therefore, $\ell(1) \leq \ell(2) \leq \ell(3) \leq \cdots$.‡ As observed in [3], the KM inequality and the monotonicity of $\ell(x)$ imply the following:

$$1 \geq \sum_{x'=1}^{\infty} 2^{-\ell(x')} \geq \sum_{x'=1}^{x} 2^{-\ell(x')} \geq x \cdot 2^{-\ell(x)}, \tag{10}$$

and so,

$$\ell(x) \geq \log x. \tag{11}$$

In other words, no UD code can assign to any $x \in \mathcal{X}$ a codeword shorter than the base 2 logarithm of the rank of x in the order of non-decreasing $\ell(x)$. This is a fundamental limitation of data compression, that stems from purely combinatorial considerations, regardless of any probability distribution that may govern the data, and even in the absence of any probabilistic model. Therefore, any concrete coding scheme whose length function comes close to $\log x$ is optimal or nearly optimal.

As mentioned in the Introduction, the simple codes proposed by Elias [4] come close as they all have length functions of the form $\ell(x) = \log x + O(\log(\log x))$, and so, they are nearly optimal at least for large x. Obviously, there is no UD code that meets the lower bound (11) for all x since $\sum_{x=1}^{\infty} 2^{-\log x} = \sum_{x=1}^{\infty} \frac{1}{x} = \infty$. In other words, the corresponding KM sum is not only larger than unity, but it diverges altogether. Accordingly, considering the above-mentioned correspondence between compression and probability assignment, there is no probability distribution, $P(x)$, that is proportional to $2^{-\log x} = \frac{1}{x}$ for all $x \in \mathcal{N}$, since this series diverges and therefore it is not normalizable. A natural compromise that comes to mind is then to resort to a probability distribution that is quite similar, namely, one where

$$P(x) \propto \frac{1}{x^{\beta}}, \qquad \beta > 1 \tag{12}$$

which is the so-called zeta distribution:

$$P(x) = \frac{1}{\zeta(\beta) x^{\beta}}, \tag{13}$$

‡ Even if this is not the case, one can always rearrange the elements of $\mathcal{X}$ according to non-decreasing values of $\ell(x)$ and relabel the members of this alphabet.

where the normalization constant is recognized as the Riemann zeta function (6). We henceforth denote the zeta distribution by $P_\beta(x)$ instead of the generic notation $P(x)$, that is,

$$P_\beta(x) = \frac{x^{-\beta}}{\zeta(\beta)}, \tag{14}$$

in order to maintain the dependence on β in the notation. Here the parameter β can be used as a regularization parameter that controls the degree of proximity to $\frac{1}{x}$ and the closely related Zipf distribution, the Zipf-Mandelbrot distribution, etc., as discussed in the Introduction. Ignoring integer length constraints, the length function pertaining to the zeta distribution is given by

$$\ell_\beta(x) = -\log P_\beta(x) = \beta \log x + \log \zeta(\beta). \tag{15}$$

When β comes close to 1 (from above), $\ell_\beta(x)$ comes close to the lower bound, $\log x$, at least for large x, with $\log x \gg \log \zeta(\beta)$. As mentioned earlier, in Appendix A, we propose a simple coding scheme whose length function is essentially the same as $\ell_\beta(x)$. The optimal choice of β is associated with a compromise since the first term of $\ell_\beta(x)$ is increasing with β, whereas the second term is decreasing. If the expectation μ of $\log x$ is known, or can be estimated empirically from past data, then the optimal choice of β is readily seen to be the one for which the expectation of $\log x$ under P_β is equal to μ.

3. The Zeta Partition Function and its Hagedorn Behavior

3.1. Single Particle

We now focus on the normalization constant of $P_\beta(\cdot)$, which is $\zeta(\beta)$, and view it as a partition function in the form of eq. (7). Let $\Delta > 0$ be arbitrarily small, and for every positive integer i, consider the range $\mathcal{N}_i = \{x \in \mathcal{N} :\ (i-1)\Delta \le \ln x < i\Delta\}$. Note that for large i, and ignoring integer-value constraints, the cardinality of $\mathcal{N}_i$ is given by

$$|\mathcal{N}_i| = |\{x \in \mathcal{N} :\ e^{i-1)\Delta}| \le x < e^{i\Delta}| = e^{i\Delta} - e^{(i-1)\Delta} = e^{i\Delta}(1 - e^{-\Delta}). \tag{16}$$

Let us examine tail of the zeta function, $\zeta_n(\beta) \equiv \sum_{x=n}^{\infty} e^{-\beta \ln x}$ for a given $n \gg 1$, which constitutes the contribution of high energies to the partition function, $\zeta(\beta)$. Then, denoting $\lambda_n = \ln(n)/\Delta$

$$\begin{aligned}
\zeta_n(\beta) &= \sum_{x=n}^{\infty} e^{-\beta \ln x} \\
&= \sum_{i \ge \lambda_n} \sum_{x \in \mathcal{N}_i} e^{-\beta \ln x} \\
&\le \sum_{i \ge \lambda_n} \sum_{x \in \mathcal{N}_i} e^{-\beta (i-1)\Delta} \\
&= \sum_{i \ge \lambda_n} |\mathcal{N}_i| e^{-\beta (i-1)\Delta} \\
&= (1 - e^{-\Delta}) \cdot \sum_{i \ge \lambda_n} e^{i\Delta} \cdot e^{-\beta (i-1)\Delta}
\end{aligned}$$

$$= (1 - e^{-\Delta})e^{\beta\Delta} \cdot \sum_{i \geq \lambda_n} e^{i\Delta(1-\beta)}. \tag{17}$$

Likewise,

$$\begin{aligned} \zeta_n(\beta) &= \sum_{i \geq \lambda_n} \sum_{x \in \mathcal{N}_i} e^{-\beta\Delta} \\ &\geq (1 - e^{-\Delta}) \cdot \sum_{i \geq \lambda_n} e^{i\Delta} \cdot e^{-i\beta\Delta} \\ &= (1 - e^{-\Delta}) \cdot \sum_{i \geq \lambda_n} e^{i\Delta(1-\beta)}. \end{aligned} \tag{18}$$

Both the upper bound and the lower bound on $\zeta_n(\beta)$ converge to the same limit as $\Delta \to 0$. For small Δ, $1 - e^{-\Delta} \approx \Delta$ and $e^{\beta\Delta} \approx 1$, and so, the common limit becomes

$$\int_{\ln n}^{\infty} e^{E(1-\beta)} \mathrm{d}E, \tag{19}$$

which is equal to $\frac{n^{1-\beta}}{\beta-1}$ for $\beta > 1$ and it diverges for $\beta \leq 1$. This is exactly the same behavior as that of the partition function of a Hagedorn system (see Appendix B for details), where the density of states, e^E (namely, $e^{\beta_{\mathrm{c}}E}$ with $\beta_{\mathrm{c}} = 1$) 'competes' with the Boltzmann factor $e^{-\beta E}$ in the integration over large E. The critical inverse temperature (a.k.a. the Hagedorn inverse temperature), $\beta_{\mathrm{c}} = 1$, is of course the boundary point between convergence and divergence of the zeta function in the first place. It is well known that when β is just slightly above 1,

$$\zeta(\beta) \approx \frac{1}{\beta - 1} \tag{20}$$

in the sense that $\lim_{\beta \downarrow 1}(\beta - 1)\zeta(\beta) = 1$.

It should be noted that essentially the same effect occurs in physical models where the Hamiltonian consists of a potential function that is logarithmic in the position of the particle, similarly as in [14], [15], [16], [17]. However, here the logarithmic Hamiltonian was obtained from a fundamental information-theoretic consideration, as explained above.

3.1.1. The Bose Gas with Log-Prime Energy Levels Another interesting statistical-mechanical analogy becomes apparent once the zeta function is displayed in Euler's product form,

$$\zeta(\beta) = \prod_{p \in \mathcal{P}} (1 - p^{-\beta})^{-1} = \prod_{p \in \mathcal{P}} (1 - e^{-\beta \ln p})^{-1}, \tag{21}$$

where $\mathcal{P} = \{2, 3, 5, \ldots\}$ is the set of all primes. The second expression is readily recognized as the grand-canonical partition function of bosons with energy levels $E_p = \ln p$, $p \in \mathcal{P}$, namely, again logarithmic energy levels, except that here, not the logarithms of all positive integers are involved, but only those of the primes. The mean occupation of state p is therefore,

$$\bar{N}_p = \frac{1}{e^{\beta \ln p} - 1}, \tag{22}$$

and so, the expected total number of particles is

$$\bar{N} = \sum_{p\in\mathcal{P}} \frac{1}{e^{\beta \ln p} - 1}, \tag{23}$$

which tends to infinity as $\beta \downarrow 1$, since $\sum_{p\in\mathcal{P}} \frac{1}{p-1} \geq \sum_{p\in\mathcal{P}} \frac{1}{p} = \infty$, as was proved by Euler in 1737. It therefore appears that from the viewpoint of the Bose gas model, the critical behavior is manifested as a phenomenon of an unbounded increase in the number of particles.

To understand the relation between the two statistical-mechanical models, consider the factorization of the positive integer x of the first model into its prime factors,

$$x = \prod_{p\in\mathcal{P}} p^{N_p}, \tag{24}$$

where $\{N_p\}_{p\in\mathcal{P}}$ are non-negative integers, which designate the multiplicities of the various prime factors. Now,

$$\begin{aligned}
\zeta(\beta) &= \sum_{x=1}^{\infty} x^{-\beta} \\
&= \sum_{x=1}^{\infty} \prod_{p\in\mathcal{P}} p^{-\beta N_p} \\
&= \sum_{N_2=0}^{\infty} \sum_{N_3=0}^{\infty} \sum_{N_5=0}^{\infty} \cdots \prod_{p\in\mathcal{P}} p^{-\beta N_p} \\
&= \prod_{p\in\mathcal{P}} \left(\sum_{k=0}^{\infty} p^{-\beta k} \right) \\
&= \prod_{p\in\mathcal{P}} \left(1 - p^{-\beta}\right)^{-1},
\end{aligned} \tag{25}$$

where the third equality is due to the fact every positive integer x has a unique factorization into prime factors. Consider an infinite series of independent geometric random variables, $N_2, N_3, N_5, \ldots$, whose marginal distributions are

$$\Pr\{N_p = k\} = (1 - p^{-\beta})p^{-\beta k}, \quad k = 0, 1, 2, \ldots. \tag{26}$$

Then the random variable x (under P_β) can be represented as $\prod_{p\in\mathcal{P}} p^{N_p}$, where each N_p, $p \in \mathcal{P}\}$, is an independent geometric random variable with parameter $p^{-\beta}$. Accordingly, the occupation numbers, $\{N_p\}$, of this boson-gas model are simply the multiplicities of the various primes in the factorization of x. Likewise,

$$\log x = \sum_{p\in\mathcal{P}} N_p \log p = \sum_{p\in\mathcal{P}} N_p E_p, \tag{27}$$

and

$$\ell(x) = \beta \log x + \log \zeta(\beta) = \beta \sum_{p\in\mathcal{P}} N_p \log p - \sum_{p\in\mathcal{P}} \log(1 - p^{-\beta}). \tag{28}$$

3.2. Multiple Particles

We now return to the zeta function in its original form. For a single particle, we were able to observe an exponential density of states only at very high energy levels. To clarify the picture for all energy levels, we now examine the density of states, or more precisely, the micro-canonical entropy, in the case of N independent particles in the thermodynamic limit, $N \to \infty$. Clearly,

$$[\zeta(\beta)]^N = \left[\sum_{x=1}^{\infty} e^{-\beta \ln x}\right]^N = \sum_{x_1=1}^{\infty}\sum_{x_2=1}^{\infty}\cdots\sum_{x_N=1}^{\infty} \exp\left\{-\beta\sum_{i=1}^{N}\ln x_i\right\}. \tag{29}$$

Denoting the micro-state $(x_1, \ldots, x_N)$ by $\boldsymbol{x}$, the Hamiltonian is now given by

$$\mathcal{H}(\boldsymbol{x}) = \sum_{i=1}^{N} \ln x_i. \tag{30}$$

The specific entropy, $s(\epsilon)$, for a given energy per particle, ϵ, is given by the Fenchel-Legendre transform of $\ln\zeta(\beta)$:

$$s(\epsilon) = \inf_{\beta\geq 0}\{\beta\epsilon + \ln\zeta(\beta)\} = \inf_{\beta>1}\{\beta\epsilon + \ln\zeta(\beta)\}, \tag{31}$$

where the second equality is due to the fact that $\zeta(\beta)$ diverges for $\beta \in [0,1]$, and so, the infimum must be achieved (or at least approached) in the range $\beta > 1$. Fig. 1 depicts the function $s(\epsilon)$ across a certain range of ϵ, including small ϵ. As can be seen, for large ϵ, the curve becomes nearly linear. Indeed, for large ϵ, the minimizing β approaches 1 (from above), and then $\zeta(\beta) \approx \frac{1}{\beta-1}$. Therefore, for large ϵ,

$$s(\epsilon) \approx \inf_{\beta>1}\{\beta\epsilon - \ln(\beta-1)\}. \tag{32}$$

Taking the derivative of the right-hand side with respect to β and equating to zero, one readily finds that the minimizing β is about $\beta^* = 1 + \frac{1}{\epsilon}$, which upon substituting back into the expression $\beta^*\epsilon - \ln(\beta^* - 1)$ yields

$$s(\epsilon) \approx \epsilon + \ln(e\cdot\epsilon), \tag{33}$$

and so, the leading term is indeed linear in ϵ with coefficient $\beta_{\rm c} = 1$, as expected. The corresponding density of states for large ϵ, is therefore roughly proportional to $\epsilon^N e^{N\epsilon}$, and the Hagedorn-type behavior is observed once again.

Note that in this statistical-mechanical model there is only partial equivalence between the micro-canonical ensemble and the canonical one. The equivalence holds merely in the range $\beta > 1$. For the range $\beta \leq 1$, there is no matching energy levels. In the micro-canonical ensemble, the temperature is stuck at $T = T_{\rm c} = \frac{1}{\beta_{\rm c}} = 1$, whereas the canonical ensemble allows all $T < 1$.

By contrast, if the particle states are limited to a finite range, $\{1, 2, \ldots, J\}$, the curve of $s(\epsilon)$ would tend to a plateau at the level $\ln J$ and would no longer grow linearly. In this case, all temperatures are achievable and there are no critical phenomena.

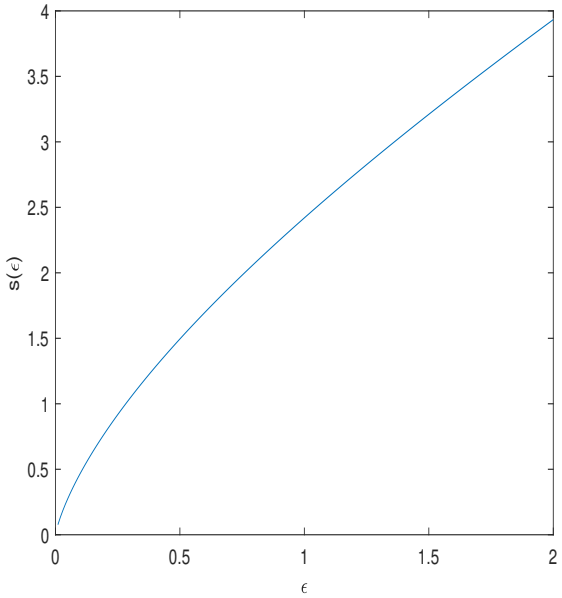

Figure 1. The entropy function $s(\epsilon) = \inf_{\beta>1}\{\beta\epsilon + \ln\zeta(\beta)\}$. Observe that for large ϵ, the function becomes nearly linear in ϵ.

4. Large Deviations Analysis

It is instructive to examine also the large deviations behavior associated with the zeta distribution: What is the exponential rate of the probability that $\sum_{i=1}^{N}\ln x_i$ would exceed $N\epsilon$, where $\epsilon > 0$ is a given constant, independent of N, and larger than the expectation of $\ln x_1$. This is readily accomplished by applying the Chernoff bound:

$$\begin{aligned}
&\Pr\left\{\sum_{i=1}^{N}\ln x_i \geq N\epsilon\right\}\\
&\leq \inf_{\lambda\geq 0}\boldsymbol{E}\left\{\exp\left[\lambda\left(\sum_{i=1}^{N}\ln x_i - N\epsilon\right)\right]\right\}\\
&= \inf_{\lambda\geq 0} e^{-\lambda\epsilon N}\boldsymbol{E}\left\{\exp\left[\lambda\sum_{i=1}^{N}\ln x_i\right]\right\}\\
&= \inf_{\lambda\geq 0} e^{-\lambda\epsilon N}\left[\boldsymbol{E}\left\{\exp\left(\lambda\ln x\right)\right\}\right]^N\\
&= \inf_{\lambda\geq 0} e^{-\lambda\epsilon N}\left[\boldsymbol{E}\{x^\lambda\}\right]^N\\
&= \inf_{0\leq\lambda<\beta-1} e^{-\lambda\epsilon N}\left[\frac{\zeta(\beta-\lambda)}{\zeta(\beta)}\right]^N\\
&= \exp\left\{-N\cdot\sup_{0\leq\lambda<\beta-1}\left[\lambda\epsilon - \ln\zeta(\beta-\lambda) + \ln\zeta(\beta)\right]\right\}. \qquad (34)
\end{aligned}$$

We note that as $\beta\downarrow 1$, the range of optimization of the Chernoff parameter λ shrinks to zero and the large deviations rate function vanishes. When ϵ is large, the optimal λ is

about $\lambda^* \approx \beta - 1 - \frac{1}{\epsilon}$, and the resulting large deviations rate function becomes nearly $\epsilon(\beta-1) - \ln(e\cdot\epsilon) + \ln\zeta(\beta)$, namely, essentially linear in ϵ for large ϵ.

The large deviations behavior is relevant also for the coding problem discussed earlier. Suppose that we encode a block $\boldsymbol{x} = (x_1, \ldots, x_N)$ of independent integers, all governed by P_β. What is the probability that the total code-length $\ell(\boldsymbol{x}) = -\log P_\beta(\boldsymbol{x}) = -\sum_{i=1}^N \log P_\beta(x_i)$ exceeds a certain threshold, nR. The motivation is clear: suppose we wish to store the compressed representation of $\boldsymbol{x}$ in a buffer of size nR bits and we are concerned by the unfortunate event of buffer overflow, which causes loss of information. Based on the structure of P_β, this is the event $\beta\sum_{i=1}\log x_i + N\log\zeta(\beta) \geq nR$, or equivalently, $\beta\sum_{i=1}^N \ln x_i + N\ln\zeta(\beta) \geq R\ln 2$. But when it comes to large deviations performance, the length function $\ell(\boldsymbol{x}) = -\log P_\beta(\boldsymbol{x})$ may not be optimal in the sense of maximizing the large-deviations rate function. It is instructive to examine whether there is another value of the parameter of the zeta distribution, call it θ, whose corresponding length function, $\ell_\theta(\boldsymbol{x}) = -\log P_\theta(\boldsymbol{x})$ would be better in that sense. For θ, the buffer overflow event becomes $\theta\cdot\sum_{i=1}^N \ln x_i + N\ln\zeta(\theta) \geq R\ln 2$, or equivalently,

$$\sum_{i=1}^N \ln x_i \geq \frac{R\ln 2 - \ln\zeta(\hat{\theta})}{\theta}, \tag{35}$$

and then we are back to the above derivation with the assignment

$$\epsilon = \frac{R\ln 2 - \ln\zeta(\theta)}{\theta}. \tag{36}$$

The optimal value of θ is the one that maximizes the right-hand side of the last equation. Interestingly, this value of θ depends merely on R, and not on the parameter β of the underlying zeta distribution. This means that when designing a code with good large deviations behavior, one needs to know merely the buffer size, but not the parameter β of the underlying source. Fig 2 depicts the curve of ϵ as a function of θ for $R\ln 2 = 3$. The maximum achieved is $\epsilon_{\max} = 1.3617$ at $\theta = 1.54$.

5. Summary and Outlook

We have investigated the problem of coding sequences of integers from the perspective of statistical mechanics, focusing on the interplay between heavy-tailed distributions, large deviations of code length, and the structure of the associated partition function. The starting point of our analysis is the elementary but fundamental observation that, for any prefix-free code on a countable alphabet, code lengths must grow at least logarithmically with the index. This intrinsic constraint naturally leads to coding schemes and probabilistic models in which $\log x$ plays the role of an energy variable.

Within this framework, power-law distributions emerge as canonical objects, both from a theoretical and an empirical standpoint. They are consistent with the minimal logarithmic scaling imposed by counting arguments and, at the same time, capture the heavy-tailed behavior observed in a wide variety of systems. The normalization of these distributions introduces a partition function with a finite radius of convergence, thereby

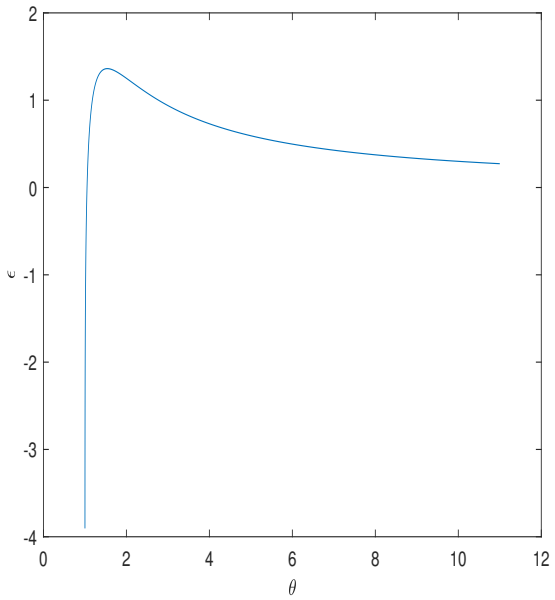

Figure 2. A graph of ϵ vs. θ for $R\log(e) = 3$.

placing the problem in close analogy with statistical-mechanical systems exhibiting critical phenomena.

Our main focus has been on the coding of i.i.d. sequences drawn from the zeta distribution and on the large deviations behavior of the total code length under mismatched coding. We have shown that the probability of atypically large code lengths admits a precise exponential characterization, governed by a rate function with a clear variational structure. This allows one to formulate and solve the problem of optimal mismatch in a sharp way: the coding parameter that controls rare events is determined by a nonlinear relation linking it to the typical energy of the source.

A central outcome of this analysis is that, in the regime of large deviations, the optimal coding parameter is driven toward the critical point at which the partition function diverges. We have quantified this behavior and shown that the approach to criticality is exponentially fast in the deviation level. In this regime, the normalization term dominates the code length, and the system exhibits features analogous to those of Hagedorn-type models, in which the exponential growth of the density of states leads to a limiting temperature.

This critical behavior has important structural consequences. In particular, it leads to a breakdown of full equivalence between the canonical and micro-canonical descriptions. While the two ensembles remain equivalent at the level of typical fluctuations, their correspondence becomes singular in the high-energy regime. The mapping between the inverse temperature and the mean energy ceases to be regular, and the canonical parameter is no longer able to parameterize large deviations in a smooth manner. This provides a concrete and analytically tractable example of partial ensemble equivalence arising from the divergence of the partition function.

Beyond the specific results obtained here, the present work highlights a broader perspective. Coding over countable alphabets, heavy-tailed probability distributions, and statistical-mechanical models with unbounded state spaces share a common structural core. Concepts such as energy, entropy, partition functions, and phase transitions arise naturally in the analysis of code lengths, even in purely information-theoretic settings. Conversely, ideas from information theory, such as optimal coding and large deviations, provide useful tools for understanding the behavior of systems near criticality.

Several directions for future work suggest themselves. One natural extension is to consider truncated or regularized versions of power-law distributions, for which the partition function remains finite, and to study how full equivalence of ensembles is recovered in this setting. Another direction is to investigate more general classes of heavy-tailed distributions and to determine the extent to which the Hagedorn-type behavior identified here is universal. It would also be of interest to explore connections with universal coding schemes and Bayesian mixture models, where coding for the integers arises naturally in the description of model complexity.

In summary, the problem of coding the integers provides a simple yet rich setting in which fundamental ideas from information theory and statistical mechanics meet. The emergence of criticality, the role of heavy tails, and the partial breakdown of ensemble equivalence all arise in a transparent and analytically accessible form, suggesting that this framework may serve as a useful laboratory for further explorations at the interface of these fields.

Appendix A

Practical Structured Coding for the Zeta Distribution

Consider the zeta (power-law) distribution on the positive integers,

$$P_\beta(x) = \frac{x^{-\beta}}{\zeta(\beta)}, \qquad \beta > 1. \tag{37}$$

The Shannon optimal code lengths for this distribution are

$$\ell^*(x) = -\log_2 P_\beta(x) = \beta \log_2 x + \log_2 \zeta(\beta), \tag{38}$$

indicating that, up to an additive constant, the optimal code-length grows as $\beta \log n$. However, a direct implementation of the Shannon code is impractical due to the lack of a simple constructive structure. In this appendix, we derive a structured prefix code that closely matches these optimal lengths while remaining simple to implement.

The key idea is to exploit the natural dyadic partition of the integers. For each $x \geq 1$, define

$$k = \lfloor \log_2 x \rfloor, \qquad r = x - 2^k, \tag{39}$$

so that x lies in the interval $\{2^k, 2^k + 1, \ldots, 2^{k+1}\}$ and $r \in \{0, 1, \ldots, 2^k - 1\}$. This representation separates the integer into a *scale* parameter k and an *offset* r within the

dyadic block. Under the zeta distribution, the induced distribution of k satisfies

$$P(k) = \sum_{n=2^k}^{2^{k+1}-1} P_\beta(n) \approx C \cdot 2^{-k(\beta-1)} \tag{40}$$

for large k, where C is a normalization constant that depends only on β (which can easily be found upon approximating the sum defining $P(k)$ by the integral of $u^{-\beta}/\zeta(\beta)$ across the interval $u \in [2^k, 2^{k+1}]$). Thus, the scale k is approximately geometrically distributed with parameter $2^{-(\beta-1)}$. Conditioned on k, the offset r is approximately uniform over its range.

Motivated by this structure, we define a two-part prefix code:

(i) **Encoding the scale k:** Since k is approximately geometric, we encode it using a Golomb code [18], [19] (or any near-optimal code for geometric distributions) with parameter $q = 2^{-(\beta-1)}$.

(ii) **Encoding the offset r:** Given k, the offset r is encoded using a fixed-length binary representation of k bits.

The resulting codeword for x is the concatenation of the codeword for k and the k-bit binary representation of r. Let $\ell(x)$ denote the length of the resulting code. The Golomb code for k achieves an average length $\ell_G(k) \approx (\beta-1)k + O(1)$, and therefore the total length satisfies $\ell(x) = \ell_G(k) + k \approx \beta k + O(1)$. Since $k = \lfloor \log x \rfloor$, we obtain $\ell(x) = \beta \log x + O(1)$, which matches the Shannon optimal length $\ell^*(x)$ up to an additive constant independent of x.

Appendix B

Background on Hagedorn Statistical Mechanics

In most familiar physical systems, the number of accessible microscopic configurations grows relatively slowly as energy increases. Typically, the density of states grows as a power law in energy, and thermodynamic quantities behave smoothly as the system is heated. However, certain systems display a fundamentally different statistical structure: the number of accessible states increases exponentially with energy. When this occurs, conventional thermodynamics predicts the appearance of a limiting temperature, known as the Hagedorn temperature.

The idea of a limiting temperature was first proposed in the 1960s in the context of high-energy particle physics. It was observed that the spectrum of hadronic particles—particles composed of quarks bound by the strong interaction—appears to grow exponentially with energy. If the density of states has the approximate form

$$\rho(E) \sim E^{-a} e^{\beta_{\mathrm{H}} E}, \tag{41}$$

then the thermodynamic properties of the system change qualitatively. In particular, the canonical partition function,

$$Z(\beta) = \int_0^\infty \rho(E) e^{-\beta E} \mathrm{d}E \tag{42}$$

converges only when the inverse temperature β exceeds a critical value β_{H}. In other words, thermal equilibrium in the canonical ensemble exists only for temperatures below a certain threshold T_{H}. When the temperature approaches this value from below, the partition function diverges, indicating that the usual canonical description of the system breaks down.

This unusual behavior reflects a simple statistical mechanism. Because the number of available states grows exponentially with energy, adding energy to the system primarily increases the number of accessible configurations rather than increasing the average kinetic energy of existing degrees of freedom. Consequently, the entropy grows approximately linearly with energy. Since temperature is defined thermodynamically through

$$\frac{1}{T} = \frac{\partial S}{\partial E}, \tag{43}$$

a linear dependence $S(E) \propto E$ implies that the derivative becomes constant. The temperature therefore approaches a fixed value rather than continuing to rise as energy increases.

Physically, this means that supplying additional energy does not significantly increase the temperature. Instead, the energy is absorbed through the creation of new states. In the context of hadronic matter, this was originally interpreted as the ultimate temperature of strongly interacting particles: heating the system further would simply produce more and more hadrons. Modern understanding, informed by quantum chromodynamics, interprets the Hagedorn temperature differently. It marks the transition from ordinary hadronic matter to a new phase in which quarks and gluons are no longer confined inside hadrons, forming a quark–gluon plasma.

The concept later appeared in other areas of theoretical physics as well. In string theory, for example, the number of vibrational modes of strings grows exponentially with energy, leading to a similar divergence in the partition function at a characteristic temperature. Near this temperature, long and highly excited strings dominate the statistical behavior of the system. The Hagedorn temperature thus emerges as a universal feature of string thermodynamics.

Connections have also been drawn between Hagedorn behavior and the thermodynamics of black holes. Black hole entropy is proportional to the area of the event horizon and corresponds to an enormous number of microscopic configurations. Although the physical origins differ, the underlying statistical pattern—rapid growth of the number of states with energy—resembles the behavior observed in Hagedorn systems. These parallels suggest that exponential state growth may be a common statistical feature of extreme physical regimes.

Importantly, the appearance of a Hagedorn temperature is not restricted to high-energy or relativistic systems. The essential requirement is simply the exponential growth of the accessible phase-space volume with energy. Whenever the number of micro-states increases exponentially, the same statistical mechanism can produce a limiting temperature. This opens the possibility of observing Hagedorn-like behavior in

much simpler physical systems, including classical models.

One can therefore ask whether analogous phenomena arise in non-relativistic settings where theoretical analysis and numerical experiments are easier to perform. Classical systems can indeed display similar statistical structures when their geometry or potential energy landscape causes the available phase space to expand exponentially with energy. In such cases, even simple particles can exhibit thermodynamic behavior reminiscent of high-energy particle systems.

The broader scientific interest in Hagedorn physics lies in understanding the interplay between energy, entropy, and state counting. When the growth of accessible configurations becomes sufficiently rapid, conventional thermodynamic intuition breaks down, and new phenomena emerge—such as limiting temperatures, divergences in specific heat, and unconventional dynamical behavior.

Thus, Hagedorn statistical mechanics provides a striking example of how the structure of the microscopic state space can fundamentally reshape macroscopic thermodynamics. What began as an observation in particle physics has evolved into a general statistical principle that applies across many areas of modern theoretical physics.

References

[1] Ziv J. and Lempel A. "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. IT–23, no. 3, pp. 337–343, May 1977.

[2] Ziv J. and Lempel A. "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT–24, no. 5, pp. 530–536, September 1978.

[3] Rissanen J. "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, vol. 11, no. 2, pp. 416–431, 1983.

[4] Elias P. "Universal codeword sets and representations of the integers," *IEEE Trans. Inform. Theory*, vol. IT–21, no. 2, pp. 194–203, March 1975.

[5] Cover T. M. and Thomas J. A, *Elements of Information Theory*, Second Edition, Wiley–InterScience, John Wiley & Sons, 2006.

[6] Powers D. M. W. "Applications and explanations of Zipf's law," *Joint conference on New Methods in Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, pp. 151-–160, 1998.

[7] Piantadosi S. "Zipf's word frequency law in natural language: A critical review and future directions," *Psychon Bull Rev.* vol. 21, no. 5, pp. 1112-–1130, 2014.

[8] Mandelbrot B. "Information Theory and Psycholinguistics". In R. C. Oldfield and J. C. Marchall (ed.). Language. Penguin Books, 1968.

[9] Baer M. B. "Prefix codes for power laws with countable support," `https://arxiv.org/pdf/cs/0611073`

[10] Baer M. B. "Prefix codes for power laws," *Proc. of the 2008 International Symposium on Information Theory* (ISIT 2008), pp. 2464–2468, Toronto, Canada, July 6–11, 2008.

[11] Hagedorn R. "Statistical thermodynamics of strong interactions at high energies," *Nuovo Cim. Suppl.*, vol. 3, pp. 147–186, 1965.

[12] Atick J. J. and Witten E. "The Hagedorn transition and the number of degrees of freedom of string theory," *Nuclear Physics B*, vol. 310, no. 2, 291, 1988.

[13] Cabibbo, N. and Parisi G. "Exponential hadronic spectrum and quark liberation" *Physics Letters B*, vol. 59, no. 1, pp. 67–69, 1975.

[14] Kessler D. A. and Barkai E. "Infinite covariant density for diffusion in logarithmic potentials

and optical lattices," *Phys. Rev. Lett.*, vol. 105, p. 120602, Sep 2010. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.105.120602

[15] Dechant A., Lutz E., Barkai E., and Kessler D. A. "Solution of the Fokker-Planck equation with a logarithmic potential," *Journal of Statistical Physics*, vol. 145, no. 6, pp. 1524-–1545, 2011. [Online]. Available: https://doi.org/10.1007/s10955-011-0363-z

[16] Hirschberg O., Mukamel D., and Schütz G. M. ,"Approach to equilibrium of diffusion in a logarithmic potential," *Phys. Rev. E*, vol. 84, p. 041111, Oct 2011. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.84.041111

[17] Hirschberg O., Mukamel D., and Schütz G. M. , "Diffusion in a logarithmic potential: scaling and selection in the approach to equilibrium," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, no. 2, p. P02001, Feb. 2012. [Online]. Available: https://doi.org/10.1088/1742-5468/2012/02/P02001

[18] Golomb S. W. "Run-length encodings," *IEEE Trans. Inform. Theory*, vol. IT–12, no. 3, pp. 399–401, May 1966.

[19] Gallager R. G. and van Voorhis D. C. "Optimal source codes for geometrically distributed integer alphabets," *IEEE Trans. Inform. Theory*, vol. 21, no. 2, pp. 228-–230, March 1975.