entropy

MDPI

Article

Maximum Configuration Principle for Driven Systems with Arbitrary Driving

Rudolf Hanel [1,2] and **Stefan Thurner [1,2,3,4],***

[1]   Section for the Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria; rudolf.hanel@meduniwien.ac.at
[2]   Complexity Science Hub Vienna, Josefstädterstrasse 39, 1080 Vienna, Austria
[3]   Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
[4]   International Institute for Applied Systems Analysis, Schlossplatz 1, 2361 Laxenburg, Austria
[*]   Correspondence: stefan.thurner@meduniwien.ac.at

check for updates

Abstract: Depending on context, the term entropy is used for a thermodynamic quantity, a measure of available choice, a quantity to measure information, or, in the context of statistical inference, a maximum configuration predictor. For systems in equilibrium or processes without memory, the mathematical expression for these different concepts of entropy appears to be the so-called Boltzmann–Gibbs–Shannon entropy, H. For processes with memory, such as driven- or self-reinforcing-processes, this is no longer true: the different entropy concepts lead to distinct functionals that generally differ from H. Here we focus on the maximum configuration entropy (that predicts empirical distribution functions) in the context of driven dissipative systems. We develop the corresponding framework and derive the entropy functional that describes the distribution of observable states as a function of the details of the driving process. We do this for sample space reducing (SSR) processes, which provide an analytically tractable model for driven dissipative systems with controllable driving. The fact that a consistent framework for a maximum configuration entropy exists for arbitrarily driven non-equilibrium systems opens the possibility of deriving a full statistical theory of driven dissipative systems of this kind. This provides us with the technical means needed to derive a thermodynamic theory of driven processes based on a statistical theory. We discuss the Legendre structure for driven systems.

Keywords: non-equilibrium; maximum configuration; maximum entropy principle; driven systems; statistical mechanics

1. Introduction

Entropy is a single word that refers to a number of very distinct concepts, a fact that has caused much confusion. These concepts cover areas as diverse as statistical mechanics, minimal code lengths, or the multiplicity of available choices in statistical inference. The distinction between the concepts underlying entropy is frequently not made properly. As long as we deal with equilibrium processes or processes with no memory (that can be modeled as i.i.d. sampling processes), formally it does not matter which entropy concept we use: they all lead to the same functional

$$H(p) = -k \sum_{i \in \Omega} p_i \log p_i \tag{1}$$

where k is a context- and unit-dependent positive constant, and $\Omega = \{1, 2, \cdots, W\}$ is the *sample space* containing W distinct states $i = 1, \cdots, W$ that occur (or are sampled) with relative frequencies p_i. In statistical mechanics, Ω is the set of all accessible microstates; in information theory, it is an alphabet

of W letters that can be coded in a code-alphabet using b code-letters (typically 0 and 1 for $b = 2$). Setting $k = 1/\log(b)$ often indicates that H is used as a measure of *information production*. If all we know about a class of messages is the frequency p_i of letters i, then H is a sharp lower bound on the average number of code-letters per letter, needed for coding messages in the original alphabet. In physics, k is the Boltzmann constant, measuring entropy in units of energy per degree Kelvin. We refer to H as the Boltzmann–Gibbs–Shannon entropy.

However, this degeneracy of entropy concepts with respect to H vanishes for processes with memory. For systems or processes away from equilibrium, such as for processes with memory, the different entropy (production) concepts lead to distinct, non-degenerate entropy functionals. "Entropy", in a way, becomes context-sensitive: the functional form of entropy measures differ from one class of processes to another. The different concepts link structural knowledge about a given system with the functional form of its corresponding entropy measure. We demonstrated this fact for self-reinforcing processes, Pólya urn processes in particular [1–3], for driven dissipative systems that can be modeled with slowly driven *sample space reducing* (SSR) processes [4–7] and for multinomial mixture processes [1]. Consequently, for non-equilibrium processes, we need to properly distinguish entropies depending on the context they are used for. In Reference [1], we distinguish three (there is no reason that there could not be more than three) entropy concepts:

- The concept of *maximum configuration* (MC) predictions defines entropy, S_{MC}, as the logarithm of the multiplicity of possible equivalent choices and plays a central role in predicting empirical distribution functions of a system. It plays a role in statistical mechanics [8,9] but can be extended naturally beyond physics in the context of statistical inference [10].
- The concept of entropy in *information theory* [11–13] can be identified as the *information production rate* (IP) of information sources. The corresponding entropy, S_{IP}, is an asymptotically sharp lower bound on the loss-free compression that can be achieved in the limit of infinitely long messages.
- The necessity of extensive variables (e.g., in thermodynamics) leads us to define an *extensive* entropy, S_{TD}, through an asymptotic *scaling expansion* that relates entropy with the phase-space volume growth as the size of a system changes [14,15]. (Consistent scaling of extensive variables is a prerequisite for thermodynamics to make sense. We therefore call the corresponding entropy measure "thermodynamic," S_{TD}, despite the fact that all of the three mentioned concepts may be required to characterize the full thermodynamics of a non-equilibrium system. However, to avoid confusion, we will use the term *thermodynamic scaling* entropy whenever it is necessary to distinguish the notion from Clausius' thermodynamic entropy, S_{Clausius}, for which, for reversible processes, $dS_{\mathrm{Clausius}} = dQ/T$, holds.)

Here we are interested in the description of driven statistical systems. The interest is to know the most likely probability distribution for finding the system in its possible states. Observables can then be computed. The maximum configuration entropy is a way to compute this probability distribution. In the following, we aim to understand entropies of driven dissipative systems that are relatively simple in the sense that they exhibit *stationary distributions*. In particular, we are interested in the functional form of their maximum configuration entropy, S_{MC}.

Driven systems are composed from antagonistic processes: a *driving process* that "loads" or charges the system (e.g., from low energy states to higher states) and a *relaxing* or discharging process that brings the system towards a "ground state" (equilibrium or sink state). (Both processes by themselves may be non-ergodic. By coupling both processes, the system may—especially when it assumes steady states—regain ergodicity.) The driving process induces a probabilistic "current" that *breaks detailed balance*. This creates a "preferred direction" in the dynamics of the system.

In previous work [4,5], we discussed driven processes in the limit of extremely slow driving rates. We find simple functional expressions for maximum configuration entropies that depend on the probabilities p_i only. These denote the probability for the system to occupy state i. The situation for general driving process—which constitutes the essence of this paper—is more involved. To characterize an arbitrary (stochastic) driving process, it is necessary to specify the probability with which the system

that is in a given state i at a given time t is driven to a "higher" state j at time $t + 1$. If we think of states being associated with energy levels, that would mean $\epsilon_j > \epsilon_i$. We use p_i^* to denote the *probability for the system to be "driven"* when it is in state i. It is clear that entropies for processes that are driven in a specific way will depend on the driving rate that is given by p_i^*. As a consequence, the entropy for driven dissipative systems is very different from those that were discussed in the literature of generalized entropies. We will see that the MC entropy of driven dissipative systems is neither of trace form, nor does it depend on one single distribution function only. The latter property means that they violate the first Shannon–Khinchin axiom, see, e.g., [14]. The adequate entropies of driven systems become measures that jointly describe the relaxation and the driving of the system; they depend on p_i and the (state-dependent) p_i^*.

Despite these complications, we will see that the *maximum configuration entropy* of driven systems remains a meaningful and useful concept for predicting distribution functions of driven systems. We further show that the maximum entropy principle for driven systems still follows the mantra: *maximize an entropy functional + constraints*. This functional corresponds (up to a sign) to the information divergence or relative entropy of the driven process. By identifying the constraint terms with cross-entropy, the known formula holds also within the driven context: entropy = cross entropy − divergence.

2. The Sample Space of Driven Dissipative Systems

Imagine a driven system. It is composed of a driving process and a relaxation process. The latter determines the dynamics of the system in the case of no driving—the system relaxes towards an attractor-, sink-, or equilibrium-state. The driving process can be thought of as a potential that keeps the system away from equilibrium. The system dynamics is then described by the details of the driving and relaxation processes. The collection of states i that a system can take is called the discrete *sample space*, Ω. For simplicity, we also assume discrete time steps. (All of the following can be formulated in continuous variables. A continuous sample space we call *phase space*.) The dynamics of a system is characterized by the temporal succession of states: initial state $\rightarrow i \rightarrow j \rightarrow k \cdots \rightarrow$ final state. If the system is stochastic, the transition probabilities, $p(j|i)$, represent the probability that the system will be found in state j in the next step, given that it is in state i now. States are characterized by certain properties that allows them to be sorted. In physical systems, such properties could be energy levels, ε_i, that are associated to state i. A relaxing process is a sequence of states that leads from high to lower energy levels. In other words, $p(j|i) \geq 0$ if $\varepsilon_i > \varepsilon_j$, and $p(j|i) = 0$ if $\varepsilon_i < \varepsilon_j$. This is what we call a *sample space reducing process*, a process that reduces its sample space over time [4,5]. The current sample space (and its volume) depends on the current state i the system is in. It is the set of states that can be reached from its current state i, $\Omega(i) = \{j | \varepsilon_j < \varepsilon_i\}$. If a system relaxes in a sequence $i \rightarrow j \rightarrow k \cdots$, then the sequence of sample spaces $\Omega(i) \supset \Omega(j) \supset \Omega(k) \cdots$ is nested. Generally, relaxation processes are sample space reducing processes.

A driving process typically brings a system from "low states" to "higher states"; or it brings low states to any other state. The sample space of the driving process can be the highest state, any state that is higher than the present state, or any state, i.e., the entire sample space, $\Omega_{\text{drive}} = \Omega$. In other words, it can (but does not have to) depend on the current state. A driving event is any event that allows the system to change from a low state to a higher state. A schematic view of a driven process is shown in Figure 1. This particular example consists of a relaxing (SSR) process, ϕ, and a driving process, ϕ^*, that places the system always in the highest possible state.
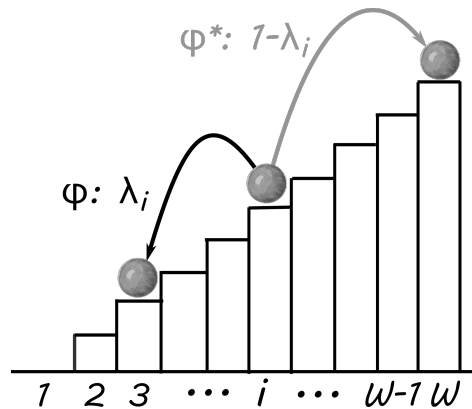
Figure 1. Illustration of a driven process as a combination of a relaxing process, ϕ, and a driving process, ϕ^*. Without driving, the ball jumps randomly to any lower stair—it never jumps upward. It follows a sample space reducing process until it hits the ground state $i = 1$, upon which it is restarted. Restarting, or driving, means that it becomes lifted to any higher state. The system is driven after it is fully relaxed to the ground state. This scenario we call slow driving. In a more general situation, at any state i there is a probability $1 - \lambda_i$ that the process experiences a driving event in the next update. Intuitively, with state-dependent driving, with probability λ_i, the process follows ϕ (black), and relaxes (can only sample states $j < i$, the ball moves downward), or the process faces a driving event with probability $1 - \lambda_i$, and follows the driving process, ϕ^* (grey).

2.1. Slowly Driven Sample Space Reducing Processes

Let us briefly recapitulate the properties of sample space reducing processes with a simple example. Given a sample space with W states, $\Omega = 1, 2, \cdots, W$, imagine that the process samples state $x_1 \in \Omega$. Then, in the next time step, the process can only sample from $\Omega(x_1) = 1, 2, \cdots, x_1 - 1 \subset \Omega$. First assume (without loss of generality [5]) that all of the states in $\Omega(x_1)$ are sampled with the same probability. With every sample x_n, the process is left with a smaller sample space, similar to a ball bouncing down a staircase, which can only sample stairs below its current position (see Figure 1). Once the process reaches the lowest possible state, $x_m = 1$ at some point in the process (think of time), m, then the sample space of the process becomes empty and the process stops at the sink state. The process can only continue if the sample space reducing process is restarted with a driving event that will allow it to sample from larger sample spaces again.

The simplest way to specify the driving process in this example is to allow the process to sample from the entire sample space Ω, once it has arrived at the sink state, $i = 1$. This is done by setting $\Omega(1) = \Omega$. In this case, the driving process becomes active after the system is completely relaxed. We call that the *slow driving* limit. Restarting the relaxing sample space reducing process produces an ongoing Markov process, and a stationary visiting distribution function of the states i, p_i. However, detailed balance is explicitly violated in driven processes.

For the slowly driven process, the transition probabilities are given by

$$p(i|j) = \Theta(j - i) \frac{q_i}{Q_{j-1}} + \delta_{j1} q_i \qquad (2)$$

where $\Theta(x)$ is the Heaviside function, δ_{ij} is the Kronecker delta, and $Q_j = \sum_{i=1}^{j} q_i$ is the cumulative distribution of weights (or prior probabilities, which determine how likely the process starts in state i) q. Let $p = (p_1, \cdots, p_W)$ be the visiting distribution of states, i.e., p_i is the relative frequency of state i being visited by the driven process, then the equation $p_j = \sum_{i=1}^{W} p(j|i) p_i$ holds and

$$p_i = p_1 \frac{q_i}{Q_i}. \qquad (3)$$

For the choice $q_i = 1/W$ (equidistributed prior probabilities), the transition probabilities read $p(i|j) = 1/(j-1)$, and we immediately obtain

$$p_i \sim \frac{1}{i}, \tag{4}$$

i.e., Zipf's law. Remarkably, also for almost any other choice of q_i, the same result is obtained, meaning that Zipf's law in the probability of state visits is an attractor of slowly driven systems, regardless of the details of the relaxing process. For details, see [4–6]. A physical example is given in [7], where the kinetic energy of a projectile passing through a box filled with balls—with which it collides elastically—reduces as an sample space reducing process. Zipf's law is found in the occurrence frequencies of projectiles with specific kinetic energies.

2.2. Driven Sample Space Reducing Processes with Arbitrary Driving

Driven sample space reducing processes on the states $i = 1, \cdots, W$ can be characterized as a mixture of the relaxing process ϕ, and a (usually stochastic) driving process, ϕ^*. As before, let us assume (without loss of generality) that the driving process ϕ^* samples among all possible states $i \in \Omega$ with probability q_i. In contrast to the slow driving case, a driving event can now occur at any time, not only after the system reaches its ground state, $i = 1$. Symbolically, the combination of driving and relaxing processes reads $\phi_\lambda = \lambda\phi + (1-\lambda)\phi^*$, where λ is the probability that the process follows ϕ (relaxes), and $1 - \lambda$ is the probability for a driving event. In the ground state, $i = 1$, ϕ, and ϕ^* behave identically. In the general case, λ may depend on the state i, i.e., $\lambda = (\lambda_1, \cdots, \lambda_W)$, and $1 - \lambda_i$ becomes a state-dependent driving rate. The transition probability from state j to state i is

$$p_\lambda(i|j) = q_i \left(\frac{\Theta(j-i)\lambda_j}{Q_{j-1}} + 1 - \Theta(j-1)\lambda_j \right). \tag{5}$$

Note that the value of λ_1 does not effect the transition probabilities of the process. Technically, we set $\lambda_1 = 0$. Using $p_i = \sum_{j=1}^{W} p(i|j)p_j$, one finds the exact asymptotic steady state solution:

$$\frac{p_i}{p_1} = \frac{q_i}{q_1} \prod_{j=2}^{i} \left(1 + \lambda_j \frac{q_j}{Q_{j-1}} \right)^{-1}. \tag{6}$$

For the case where all $\lambda_j = 0$, we get $p_i = q_i$, as expected for processes that are dominated by the driving process. For all $\lambda_j = 1$ (except for λ_1), we recover the slow driving limit of Equation (3). For the equidistributed prior probabilities, $q_i = 1/W$, the following relation between the state-dependent driving rate, $1 - \lambda(x)$, and the (stationary) distribution function exists

$$\lambda(x) = -x\frac{d}{dx} \log p(x). \tag{7}$$

Here, x labels the states in a continuous way. For details, see [7].

3. The Maximum Configuration Entropy of Driven Processes

For stationary systems, the probabilities for state visits can be derived with the use of the maximum entropy principle. We discuss the implications of driving a process for the underlying MC entropy functional. Let k_i count how often a particular state $i = 1, 2, \cdots, W$ has occurred in N sampling steps (think of a sequence of N samples). $k = (k_1, \cdots, k_W)$ is the histogram of the process, and $p_i = k_i/N$ are the relative frequencies.

To find the maximally likely distribution (maximum configuration) of a process, one considers the probability to observe the specific histogram, k, after N samples (steps), $P(k|\theta) = M(k)G(k|\theta)$. Here $M(k)$ is the multiplicity of the process, i.e. the number of sequences x that have the histogram k. $G(k|\theta)$

is the probability of sampling one particular sequence with histogram k, and *theta* parameterizes the process. Taking logarithms on both sides of $P(k|\theta) = M(k)G(k|\theta)$, we identify

$$\underbrace{\frac{1}{R} \log P(k|q)}_{-\text{relative entropy}} = \underbrace{\frac{1}{R} \log M(k)}_{\text{MC entropy } S_{MC}} + \underbrace{\frac{1}{R} \log G(k|q)}_{-\text{cross}-\text{entropy}}. \tag{8}$$

Here we scaled both sides by a factor, $R(N)$, that represents the number of degrees of freedom in the system. To arrive at relative frequencies, we use $p = k/N$. Given that N is sufficiently large, p approaches the most likely distribution function.

3.1. Slowly Driven Processes

Driving a process slowly means to restart it with a rate that is lower (or of the same order) than the typical relaxation time. In the above context of the sample space reducing process, it is restarted once the process has reached its ground state at $i = 1$. The computation of the MC entropy, together with its associated distribution functions, for slowly driven systems has been done in [1,2]. We include it here for completeness. For simplicity, we present the following derivation in the framework of sequences. They hold more generally for other processes and systems.

Let $x = (x_1, x_2, \cdots, x_N)$ be a sequence that records the visits to the states i. $\lambda = 1$ means that the process is slowly driven and restarts only in state $i = 1$. To compute the multiplicity, M, note that we can decompose any sampled SSR sequence $x = (x_1, x_2, \cdots, x_N)$ into a concatenation of shorter sequences x^r such that $x = x^1 x^2 \cdots x^R$. Each x^r starts directly after a driving event restarted the process, and ends when the process halts at $i = 1$, where it awaits the next driving event. We call x^r one "run" of the SSR process. Since every run ends in the ground state $i = 1$, we can be sure that, for the slowly driven process, we find $R = k_1$: number of runs is equal to the number of times we observe state 1. To determine M and the probability G, we represent the sequence x in a table, where we arrange the runs x^r in W columns and k_1 rows, such that the entry in the rth row and ith column contains the symbol "$*$" if run r visited state i, and the symbol "$-$" otherwise.

$r \times i$	W	$W-1$	$W-2$	$\cdots$	2	1	
1	$-$	$*$	$-$	$\cdots$	$*$	$*$	
2	$*$	$-$	$*$	$\cdots$	$-$	$*$	
3	$*$	$*$	$-$	$\cdots$	$-$	$*$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	(9)
$R-2$	$-$	$*$	$*$	$\cdots$	$-$	$*$	
$R-1$	$-$	$*$	$-$	$\cdots$	$*$	$*$	
R	$-$	$*$	$*$	$\cdots$	$-$	$*$	
	k_W	k_{W-1}	k_{W-2}	$\cdots$	k_2	k_1	

Recognizing that each column i is $k_1 = R$ entries long and contains $k_i \leq k_1$ items $*$, it follows that without changing a histogram k we can rearrange the items $*$ in column i in $\binom{k_1}{k_i} = k_1!/k_i!/(k_1 - k_i)!$ ways. M therefore is given by the product of binomial factors

$$M(k) = \prod_{i=2}^{W} \binom{k_1}{k_i}. \tag{10}$$

In other words, $M(k)$ counts all possible sequences x with histogram k. Using Stirling's approximation, the MC entropy, $S_{\text{MC}}(p) = \frac{1}{N} \log M(Np)$, is

$$S_{\text{MC}}(p) = -\sum_{i=2}^{W} \left[p_i \log \left(\frac{p_i}{p_1} \right) + (p_1 - p_i) \log \left(1 - \frac{p_i}{p_1} \right) \right]. \tag{11}$$

Note that for computing the MC entropy of the slowly driven process one only needs to know that the process gets restarted at $i = 1$. At no point did we need to know the transition probabilities from Equation (2).

This has an important implication: the entropy of the slowly driven process does not depend on how a process relaxes exactly, it only matters that the process *does* relax somehow (replacing the transition probabilities Equation (2) with other transition probabilities that respect the sample space reducing structure of the processes and restarting at $i = 1$ only changes the expression for the sequence probabilities, G) and is restarted in state 1.

Using Equation (2), the probability, G, for sampling a particular x can be computed in a similar way. Each visit to any state above the ground state, $i > 1$, in the sequence x contributes to the probability of the next visit to a state $j < i$ with a factor Q_{i-1}, independently from the state j that will be sampled next (only if $i = 1$ do we not obtain such a renormalizing factor, since the process restarts with all possible states $i \in \Omega$ as valid successor states of state 1, with prior probability q_i.). As a consequence, one obtains

$$G(k|q, N) = \prod_{i=1}^{W} q_i^{k_i} \prod_{j=2}^{W} Q_{i-1}^{-k_i}. \tag{12}$$

The cross-entropy $S_{\text{cross}}(p|q) = -\frac{1}{N} \log G(k|q, N)$ is given by

$$S_{\text{cross}}(p|q) = -\sum_{i=1}^{W} p_i \log q_i + \sum_{i=2}^{W} p_i \log Q_{i-1}, \tag{13}$$

and the maximum entropy functional of the slowly driven staircase process is given by $\psi = S_{\text{MC}} - S_{\text{cross}}$. Note that $-\psi(p|\theta)$ plays the role of a divergence, generalizing the Kullback–Leibler divergence.

To obtain the maximum configuration $\hat{k}$ or $\hat{p} = \hat{k}/N$, we can solve the maximum entropy principle by maximizing ψ with respect to p under the constraint $\sum_{i=1}^{W} p_i = 1$. As a result, one obtains, $p_i = p_1 \frac{q_i}{Q_i}$, which is exactly the solution that we obtain in Equation (3).

Note that MC entropy is not of trace form and the driving process "entangles" the ground state $i = 1$ with every other state. This entropy violates all but one Shannon–Khinchin axiom. For the uniform distribution $p_i = 1/W$, one obtains $S_{\text{MC}}(1/W, \cdots, 1/W) = 0$.

3.2. Arbitrarily Driven Processes

For exploring relaxing systems with arbitrary drivers, again, we have to compute the probability of finding the histogram, P, by following the generic construction rule for the MC divergence, entropy, and cross entropy (see [16]). We will see that, for arbitrarily driven processes, P factorizes again into a well defined multiplicity, M, and a sequence probability, G, just as for slowly driven processes.

To find P, we analyze the properties of a sampled sequence x. First note that, by mixing ϕ and ϕ^*, we now have to determine at each step, whether the process follows ϕ (relaxes) or ϕ^* (is reloaded). To model this stochastically, assume binary experiments, using, e.g., a coin ω_n for every sample x_n. If the coin samples $\omega_n = 0$ with probability λ_{x_n}, the process relaxes and follows ϕ. If the coin samples $\omega_n = 1$ with probability $1 - \lambda_{x_n}$, the process follows ϕ^*, and a driving event occurs. Let's call any instance, x_n, a *starting point of a run*, if (i) $n = 1$, (ii) $\omega_n = 0$, or (iii) $x_{n-1} = 1$. We call an instance, x_n, an *end point of the run*, if (i) $n = N$, (ii) $\omega_n = 1$, or (iii) $x_n = 1$. Except for $n = 1$ and $n = N$, all starting points directly follow an endpoint, and we can decompose the sequence x into R runs, x^r, as before, such that $x = x^1 x^2 \cdots x^R$. Each run, $x^r = (x_{\underline{r}}, \cdots, x_{\bar{r}})$, begins at a starting point $x_{\underline{r}}$, and ends in an endpoint, $x_{\bar{r}}$. Note that $\underline{1} = 1$, $\bar{r} + 1 = \underline{r+1}$, and $\bar{R} = N$. The average length of the sequences x^r is given by $\bar{\ell} = N/R$.

3.2.1. MC Entropy of Arbitrarily Driven Processes

Let $x^* = (x_{\bar{1}}, x_{\bar{2}}, \cdots, x_{\bar{R}})$ be the subsequence of x containing all endpoints of runs x^r. $k^* = (k_1^*, \cdots, k_W^*)$ is the histogram of x^*, with $\sum_i k_i^* = R$. We can now determine the number $M(k, k^*)$ of all sequences of runs $x = (x^1 x^2 \cdots x^R)$ with histograms k and k^*. Similar to the case of slow driving, this can be done by arranging runs in a table of R rows and W columns

$$
\begin{array}{c|c|c|c|c|c|c|c}
r \times i & W & W-1 & W-2 & \cdots & 2 & 1 & \\
\hline
1 & * & - & - & \cdots & \bullet & - & l(x^1) \\
2 & - & * & * & \cdots & \bullet & - & l(x^2) \\
3 & * & - & \bullet & \cdots & - & - & l(x^3) \\
\vdots & & \vdots & & & & \vdots & \vdots \\
R-2 & - & * & * & \cdots & - & \bullet & l(x^{R-2}) \\
R-1 & - & - & - & \cdots & \bullet & - & l(x^{R-1}) \\
R & - & - & * & \cdots & * & \bullet & l(x^R) \\
\hline
& k_W & k_{W-1} & k_{W-2} & \cdots & k_2 & k_1 &
\end{array}
\tag{14}
$$

We mark endpoints with the symbol "$\bullet$" (one per row), other sequence elements with "$*$", and the rest with "$-$". Having fixed k^*, we can distribute the elements of x^* in $\binom{R}{k^*} = R! / \prod_i k_i^*!$ different ways over R rows. Given this fact, we can reorder the rows of the table in increasing order of their endpoints $x_{\bar{r}}$ (compare with Figure 2). Obviously $k_1 = k_1^*$, since all visits to the ground state are endpoints. We note further that for a state $i > 1$, we can redistribute $k_j - k_j^*$ items $*$ over $\sum_{s=1}^{j-1} k_s^*$ available positions. We can now simply read off the multiplicity from the ordered table

$$
M(k^*, k) = \binom{R}{k^*} \prod_{j=2}^{W} \binom{\sum_{s=1}^{j-1} k_s^*}{k_j - k_j^*}
\tag{15}
$$

where the product over the states j are contributions of binomial factors counting the number of ways $k_i - k_i^*$ states can be distributed over $\sum_{s=1}^{j-1} k_s^*$ possible positions.
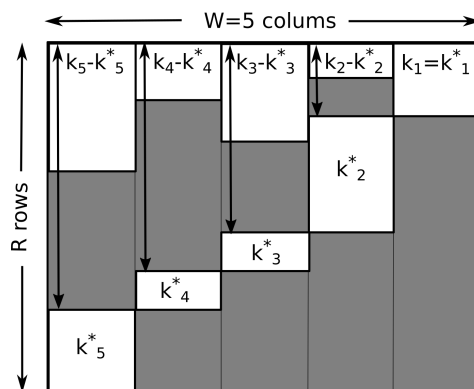

Figure 2. Demonstration of how the ordering of subsequences of a sequence x (in increasing order of their endpoints) leads to a diagram that allows us to identify the combinatorial factors that determine the multiplicity, M, of sequences with particular histograms k and k^*.

The multiplicity, M, does not explicitly depend on how exactly the system is driven. M contains no traces of the driving rates $1 - \lambda_i$, nor any dependencies on transition probabilities, meaning no dependencies on prior weights, q_i. This independence, however, is paid for with the dependence of M on a second histogram, k^*. M depends only on the *combinatorial relations* of the relaxing (sample space reducing) process with the driving process. M describes the multiplicity, i.e., the *a priori* chances of jointly observing k and k^* together. Note that k_i^* is the empirically observed number of driving

events observed in state i, k_i is the empirical number of visits of the process to state i, while $1 - \lambda_i$ is the probability of a driving event to occur when the system is in state i.

The "joint" MC entropy of the arbitrarily driven process, ϕ_λ, is the logarithm of M times a scaling factor. If we want an entropy per driving event, the factor is $1/R = \bar{\ell}/N$, and we obtain

$$S_{\mathrm{MC}}(p^*, p, \bar{\ell}) = \tfrac{1}{R} \log M(k^*, k) = H(p^*) + \sum_{j=2}^{W} \left(\sum_{s=1}^{j-1} p_s^* \right) H_2 \left(\frac{\bar{\ell} p_j - p_j^*}{\sum_{z=1}^{j-1} p_z^*} \right), \tag{16}$$

which depends on the average length of the runs, $\bar{\ell} = N/R$. The parameter $\bar{\ell}$ corresponds to the *average free relaxation time* of the process. Note that we can define the *average driving rate* by $1 - \bar{\lambda} = 1/\bar{\ell}$, which is nothing but the ratio R/N, i.e., the average number of driving events per process step. The MC entropy also depends on $p^* = k^*/R$ and $p = k/N$, the distributions of driving events and state visits, respectively. $H_2(x) = -x \log(x) - (1-x) \log(1-x)$ is the Shannon entropy of a binary process. From the fact that, by definition, $k_1 = k_1^*$, we obtain the constraint

$$p_1^* = \bar{\ell} \, p_1. \tag{17}$$

3.2.2. Cross-Entropy and Divergence of Arbitrarily Driven Processes

Similarly, we can collect the contributions to the sequence probability, G. Every time the sequence x follows the relaxing (or sample space reducing) process, we obtain a factor $\lambda_i q_i / Q_{i-1}$, for some $i = x_n$, and a factor $(1 - \lambda_i) q_i$ if the sequence reaches an endpoint, $i = x_n$, of a run x^r. As a consequence, the probability of observing a sequence x with histograms k and k^* is given by

$$G(k^*, k | q, \lambda) = q_1^{k_1^*} \prod_{i=2}^{W} \left(\frac{\lambda_i q_i}{Q_{i-1}} \right)^{k_i - k_i^*} \prod_{j=2}^{W} \left((1 - \lambda_j) \, q_j \right)^{k_j^*} \left(\frac{\lambda_j}{Q_{j-1}} \right)^{k_j - k_j^*}. \tag{18}$$

Note that $N \geq R \geq \max(k_1 \cdots, k_W)$. The joint probability of observing k^* and k together becomes

$$P_\lambda(k^*, k | q, \lambda) = M(k^*, k) G(k^*, k | q, \lambda). \tag{19}$$

This shows that the mixing of relaxing and driving processes, ϕ and ϕ^*, leads to a joint probability that depends on k and k^*. Noting that we can define $\lambda_1 = 0$, the cross entropy construction yields

$$\begin{aligned} S_{\mathrm{cross}}(p^*, p, \bar{\ell} | q, \lambda) &= -\tfrac{1}{R} \log G(k^*, k | q, \lambda) \\ &= -\sum_{j=1}^{W} p_j^* \left(\log(1 - \lambda_j) + \log q_j \right) - \sum_{j=2}^{W} \left(\bar{\ell} p_j - p_j^* \right) \log \left(\frac{\lambda_j q_j}{Q_{j-1}} \right). \end{aligned} \tag{20}$$

The MC functional, $\psi = \tfrac{1}{R} \log P$, is $\psi(p^*, p, \bar{\ell} | q, \lambda) = S_{\mathrm{MC}}(p^*, p, \bar{\ell}) - S_{\mathrm{cross}}(p^*, p, \bar{\ell} | q, \lambda)$. $-\psi$ can be interpreted as the *information divergence* or *relative entropy* of the process. Maximizing ψ with respect to p, p^*, and $\bar{\ell}$, under the constraints $\sum_i p_i^* = 1$, $\sum_i p_i = 1$, and $p_1^* = \bar{\ell} p_1$ (compare with Equation (17)), yields the maximum configuration $(\hat{p}, \hat{p}^*, \hat{\ell})$. The maximization of the MC functional,

$$\psi - \alpha \left(\sum_i p_i - 1 \right) - \alpha^* \left(\sum_i p_i^* - 1 \right) - \gamma \left(\bar{\ell} p_1 - p_1^* \right), \tag{21}$$

with respect to p, p^*, and $\bar{\ell}$, is the *maximum entropy principle* of a driven processes. α, α^*, and γ are Lagrangian multipliers.

By noting that derivatives of the MC functional Equation (21) behave differently with respect to p_1 and p_1^*, when compared with how they behave with respect to p_i and p_i^* ($i > 1$), one finds that $\alpha = \gamma = 0$, and, using the short hand for $\mu_i = \lambda_i q_i / Q_{i-1}$, one obtains the formulas

$$
\begin{aligned}
p_i &= \tfrac{1}{\bar{\ell}} \left[p_i^* + \left(\textstyle\sum_{s=1}^{i-1} p_s^* \right) \tfrac{\mu_i}{1+\mu_i} \right] \\
p_1^* &= \tfrac{1}{Z^*} \prod_{s=2}^{W} \tfrac{\mu_s}{1+\mu_s} \\
p_i^* &= \tfrac{(1-\lambda_i)q_i}{1+\mu_i} \tfrac{p_1^*}{q_1} \prod_{s=2}^{i-1} \tfrac{1}{1+\mu_s} \quad i > 1
\end{aligned}
\tag{22}
$$

These allow us to compute the solution for p_i recursively. The result is exactly the prediction from the eigenvector equation in Equation (6). This fact shows that, indeed, the maximum entropy principle yields the expected distributions. However, we learn much more here, since by construction, $\exp(-\psi(p, p^*, \bar{\ell}))$, is the probability measure of the process. This allows us to study fluctuations of p and p^* within a common framework. This framework opens the possibility of extending fluctuation theorems—which are of interest in statistical physics (see, e.g., [17])—to systems explicitly driven away from equilibrium. We can immediately predict parameters of driven processes, such as the "average free relaxation time" of the process, $\bar{\ell} = p_1^* / p_1$.

4. Legendre Structure and Thermodynamic Interpretation

In the standard equilibrium MC principle, one maximizes the functional $H(p) - \alpha' \sum_i p_i - \beta' \sum_i p_i \varepsilon_i$, with respect to p, α', and β'. α' and β' are Lagrange multipliers for the normalization constraint, $\sum_i p_i = 1$, and (e.g., in a physics context) the "energy" constraint, $\sum_i p_i \varepsilon_i = U$, where U represents the "internal energy." ε_i are "energy" states and at the extremal value, $\beta' = \hat{\beta}$ is the inverse temperature. One can also see $\beta' = \hat{\beta} = \beta$ as a parameter β of the extremal principle, and think of U as a function of $\beta\varepsilon$.

The classic MC principle can be obtained from maximizing the functional $\psi(p|q) = \frac{1}{N} \log P_{\text{multinomial}}(Np|q)$ (in the large N limit), under the constraint, $\sum_i p_i = 1$. Asymptotically, one obtains $\psi = H(p) - H(p|q)$ where $H(p|q) = -\sum_i p_i \log q_i$ is the equilibrium cross-entropy. Using the parametrization $q_i(\beta\varepsilon) = \exp(-\alpha - \beta\varepsilon_i)$ for the prior probabilities in the variables $\beta\varepsilon_i$, one obtains $\psi = H(p) - \alpha \sum_i p_i - \beta \sum_i p_i \varepsilon_i$. Maximizing $\psi - \alpha'' \sum_i p_i$ with Lagrange multiplier α'' allows us to absorb the term $\alpha \sum_i p_i$ into the normalization constraint by identifying $\alpha' = \alpha + \alpha''$. Since the distribution q is normalized, $\alpha = \alpha(\beta\varepsilon)$ is a function of $\beta\varepsilon$, and one obtains the MC solution, $\hat{p} = q(\beta\varepsilon) = \frac{1}{Z} \exp(-\beta\varepsilon_i)$. Here $Z = \exp(\alpha)$ is the normalization constant. Defining $\tilde{H}(\beta\varepsilon) = H(\hat{p})$, we obtain in the maximum configuration that $\hat{\psi}(\beta\varepsilon) = \tilde{H} - \alpha - \beta U$, with $U = \sum_i \hat{p}_i \varepsilon_i$. Moreover, $\hat{\psi}$ can be interpreted as the *Legendre transformation* of $H(p) - \alpha \sum_i p_i$, where the term $-\sum_i p_i \beta\varepsilon_i$ transforms the variables p_i into the adjoined variables, $\beta\varepsilon_i$. Since $P_{\text{multinomial}}(N\hat{p}|q)$ asymptotically approaches 1, we obtain $\hat{\psi} = 0$ and therefore $U - \tilde{H}/\beta = -\alpha/\beta$, where $F = -\alpha/\beta$ is the *Helmholtz free energy*. Since $\tilde{H}(\beta\varepsilon)$ represents thermodynamic entropy in equilibrium thermodynamics, one might ask whether a similar transformation exists for driven systems.

Let us parametrize the weights q_i and λ_i as

$$
\begin{aligned}
q_i &= \exp(-\alpha - \beta\varepsilon_i) \\
1 - \lambda_i &= \exp(-\beta^* \varepsilon_i^*)
\end{aligned}
\tag{23}
$$

Note that $1 - \lambda_i$ need not sum to 1, so $\exp(-\beta^* \varepsilon_i^*)$ does not need a normalization factor. Moreover, from $\lambda_1 = 0$, it follows that $\varepsilon_1^* = 0$. Inserting these definitions in the cross-entropy in Equation (20), we obtain

$$
S_{\text{cross}}(p^*, p, \bar{\ell}|q, \lambda) = S_D + \bar{\ell}\alpha \sum_{i=1}^{W} p_i + \beta^* \sum_{i=1}^{W} p_i^* \varepsilon_i^* + \bar{\ell}\beta \sum_{i=1}^{W} p_i \varepsilon_i,
\tag{24}
$$

with—what we call—the *driving entropy*

$$S_{\mathrm{D}} = -\sum_{j=2}^{W} \left(\bar{\ell} p_j - p_j^* \right) \Delta s_j. \tag{25}$$

Here $\Delta s_j = \log \lambda_j - \log Q_{j-1}$, which, in the given parametrization, is a function of $\beta^* \varepsilon^*$ and $\bar{\ell} \beta \varepsilon$. In the maximum configuration, the terms $\bar{\ell} \beta \sum_i p_i \varepsilon_i$ and $\beta^* \sum_i p_i^* \varepsilon_i^*$ again are the Legendre transformation from variables p and p^* to the adjoined variables, $\hat{\ell} \beta \varepsilon$ and $\beta^* \varepsilon^*$. $\hat{\ell}$ is the MC of $\bar{\ell}$ and $\psi = S_{\mathrm{MC}} - S_{\mathrm{cross}}$. By defining now the "thermodynamic entropy," $\tilde{S}_{\mathrm{MC}}(\hat{\ell} \beta \varepsilon, \beta^* \varepsilon^*) = \hat{S}_{\mathrm{MC}} / \hat{\ell}$, where $\hat{S}_{\mathrm{MC}} = S_{\mathrm{MC}}(\hat{p}, \hat{p}^*, \hat{\ell})$, and, analogously, the "thermodynamic driving entropy," $\tilde{S}_{\mathrm{D}}(\hat{\ell} \beta \varepsilon, \beta^* \varepsilon^*) = \hat{S}_{\mathrm{D}} / \hat{\ell}$, one obtains $\hat{\psi}(\hat{\ell} \beta \varepsilon, \beta^* \varepsilon^*) / \hat{\ell} = \tilde{S}_{\mathrm{MC}} - \tilde{S}_{\mathrm{D}} - \alpha - \beta^* U^* / \hat{\ell} - \beta U$. Here $U^* = \sum_{i=1}^{W} \hat{p}_i^* \varepsilon_i^*$, and $U = \sum_{i=1}^{W} \hat{p}_i \varepsilon_i^*$ is the "internal energy." Since asymptotically (for large N) we have $\hat{\psi} = 0$ again, it follows that

$$U - \tilde{S}_{\mathrm{MC}} / \beta = -\alpha / \beta - \beta^* U^* / \beta - \tilde{S}_{\mathrm{D}} / \beta. \tag{26}$$

In contrast to the maximum entropy principle for i.i.d. processes, where the inverse temperature and the normalization term of the distribution function are one-to-one related to the Lagrangian multipliers, this is no longer true for driven systems. The parameters α, β, and β^* are entangled in S_{D}! However, this fact does not—as one might expect—destroy the overall Legendre structure of the statistical theory of driven processes. It has, however, a important consequence: we can no longer expect notions such as the *Helmholtz free energy* (which has a clear interpretation for reversible processes) to have the same meaning for driven systems. It is, however, exactly this kind of question that the theory that we outlined in this paper may help us tackle. For example, in a driven system, we have to decide whether we call $U - \tilde{S}_{\mathrm{MC}} / \beta$ the Helmholtz free energy, F, or if it is $-\alpha / \beta$. If we identify $F = -\alpha / \beta$ and define the "*driven* Helmholtz free energy" as, $F_{\mathrm{D}} = U - \tilde{S}_{\mathrm{MC}} / \beta$, then we obtain $F_{\mathrm{D}} = F - \beta^* U^* / \beta - \tilde{S}_{\mathrm{D}} / \beta$. This means that, for driven systems, it remains possible to switch between thermodynamic potentials using Legendre transformations, e.g., by switching from U to F_{D}. Most importantly, the landscape of thermodynamic potentials is now enriched with two additional terms, $W_{\mathrm{D}} = \beta^* U^* / \beta$ and $Q_{\mathrm{D}} = \tilde{S}_{\mathrm{D}} / \beta$, which incorporate the driving process and its interactions with the system,

$$U - \tilde{S}_{\mathrm{MC}} / \beta = F - W_{\mathrm{D}} - Q_{\mathrm{D}}. \tag{27}$$

It is tempting to interpret W_{D} and Q_{D} as irreversible work and dissipated heat, respectively. However, a clear interpretation of those terms can only be achieved within the concrete context of a given application.

5. Conclusions

We have shown that it is possible to derive a consistent *maximum configuration* framework for driven dissipative systems that are a combination of a state-dependent driving process and a relaxation process that can always can be modeled as an SSR process. The considered processes must exhibit stationary distributions. The presented framework is a statistical theory that not only allows us to compute visiting distribution functions p from an appropriate entropy functional but also allows us to predict properties, such as its state-dependent driving rates, p^*, or the average relaxation times, $\bar{\ell}$. We find that the visiting distributions are in exact agreement with alternatively computable analytical solutions.

Remarkably, the maximum configuration entropy functional that emerges within the context of driven processes decouples completely from the detailed constraints posed by the interactions with the driving system. In this sense, it represents a property that is universally shared by processes within the considered class of driven processes. As a consequence, the maximum configuration mantra: maximize *entropy + constraint terms* to predict the maximum configuration, remains valid also for

driven systems. One only needs to keep in mind that the "reasonable" constraint terms that specify the interactions of the system with the driving process may have a more complicated structure than we are used to from i.i.d. processes. The presented theory further allows us to compute a large number of expectation values that correspond to physically observable macro variables of the system. This includes the possibility of studying fluctuations of the system and the driving process together in the same framework.

The structure of the maximum configuration entropy and the corresponding maximum entropy principle is interesting by itself. By introducing the relative frequencies p^* of the driving process and the average free relaxation time, $\bar{\ell}$, as independent quantities, the entropy decouples from the actual driving process and from how the system is linked to it. At the same time, this is only possible when the distributions p^* and p become functionally entangled. $\bar{\ell}$ effectively quantifies the strength of this entanglement. For slowly driven systems, it is the ground state that is entangled with all other states. For fast driving, all states become "hierarchically" entangled. Remarkably, however complicated this entanglement may be, we still can compute all of the involved distribution functions. As for i.i.d. processes, the maximum configuration asymptotically (for large systems) dominates the system behavior. Therefore, the maximum configuration entropy, at its maximum configuration (determined by the macro state of the driven system), can be given a thermodynamic meaning.

Finally, we discussed the thermodynamic structure of the presented statistical theory. The type of processes we can consistently describe with the presented framework include (kinetic) energy dissipations in elastic collision processes (see [7]). It is therefore not only conceivable, but in fact highly probable, that for a wide class of driven processes, a consistent thermodynamics can be derived from the statistical theory. Despite the more complicated relations between distribution functions and system parameters, the Legendre structure of equilibrium thermodynamics survives in the statistical theory of the driven processes presented here. The landscape of macro variables ($U, U^*, \beta, \beta^*, \dots$) and potentials ($F, F_{\mathrm{D}}, Q_{\mathrm{D}}, \dots$) becomes richer, and now includes terms that describe the driving process and how it couples to the relaxing system. Most importantly, however, the underlying statistical theory allows us to analyze all the differential relations between macro variables and potentials that we might consider. These may depend on the parametrization of q and λ. The presented work makes it conceivable that statistical theories can be constructed for situations, where multiple, not just two, driving and relaxing processes are interacting with one another. This results in more complicated functional expressions for MC entropy, cross-entropy, additional macro state variables, and thermodynamic potentials but does not introduce fundamental or conceptual problems.

Author Contributions: RH and ST conceptualized the work and wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Thurner, S.; Corominas-Murtra, B.; Hanel, R. The three faces of entropy for complex systems—Information, thermodynamics and the maxent principle. *Phys. Rev. E* **2017**, *96*, 032124. [CrossRef] [PubMed]
2. Thurner, S.; Hanel, R.; Klimek, P. *Introduction to the Theory of Complex Systems*; Oxford University Press: Oxford, UK, 2018.
3. Hanel, R.; Corominas-Murtra, B.; Thurner S. Analytical computation of frequency distributions of path-dependent processes by means of a non-multinomial maximum entropy approach. *New J. Phys.* **2017**, *19*, 033008. [CrossRef]
4. Corominas-Murtra, B.; Hanel, R.; Thurner, S. Understanding scaling through history-dependent processes with collapsing sample space. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 5348–5353. [CrossRef] [PubMed]
5. Corominas-Murtra, B.; Hanel, R.; Thurner, S. Extreme robustness of scaling in sample space reducing processes explains Zipf's law in diffusion on directed networks. *New J. Phys.* **2016**, *18*, 093010. [CrossRef]

6. Corominas-Murtra, B.; Hanel, R.; Thurner, S. Sample space reducing cascading processes produce the full spectrum of scaling exponents. *Sci. Rep.* **2017**, *7*, 11223. [CrossRef] [PubMed]

7. Corominas-Murtra, B.; Hanel, R.; Zavojanni, L.; Thurner, S. How driving rates determine the statistics of driven non-equilibrium systems with stationary distributions. *Sci. Rep.* **2018**, *8*, 10837. [CrossRef] [PubMed]

8. Boltzmann, L. Über die Mechanische Bedeutung des Zweiten Hauptsatzes der Wärmetheorie. *Wien Ber.* **1866**, *53*, 195–220.

9. Planck, M. *Eight Lectures on Theoretical Physics*; Columbia University: New York, NY, USA, 1915.

10. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]

11. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. [CrossRef]

12. Kraft, L.G. A Device for Quantizing, Grouping, and Coding Amplitude Modulated Pulses. Master's Thesis, MIT, Cambridge, MA, USA, 1949.

13. McMillan, B. Two inequalities implied by unique decipherability. *IEEE Trans. Inf. Theory* **1956**, *2*, 115–116. [CrossRef]

14. Hanel, R.; Thurner, S. A comprehensive classification of complex statistical systems and an ab initio derivation of their entropy and distribution functions. *Europhys. Lett.* **2011**, *93*, 20006. [CrossRef]

15. Korbel, J.; Hanel, R.; Thurner, S. Classification of complex systems by their sample-space scaling exponents. *New J. Phys.* **2018**, *20*, 093007. [CrossRef]

16. Hanel, R.; Thurner, S.; Gell-Mann, M. How multiplicity of random processes determines entropy: Derivation of the maximum entropy principle for complex systems. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 6905–6910. [CrossRef] [PubMed]

17. Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **1997**, *78*, 2690. [CrossRef]