entropy

MDPI

Tutorial

An Introduction to the Non-Equilibrium Steady States of Maximum Entropy Spike Trains

Rodrigo Cofré [1,*], Leonardo Videla [1] and Fernando Rosas [2,3,4]

[1] Centro de Investigación y Modelamiento de Fenómenos Aleatorios CIMFAV-Ingemat, Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso 2340000, Chile
[2] Centre for Psychedelic Research, Department of Medicine, Imperial College London, London SW7 2DD, UK
[3] Centre for Complexity Science and Department of Mathematics, Imperial College London, London SW7 2AZ, UK
[4] Data Science Institute, Imperial College London, London SW7 2AZ, UK
* Correspondence: rodrigo.cofre@uv.cl

Received: 12 July 2019; Accepted: 7 September 2019; Published: 11 September 2019

Abstract: Although most biological processes are characterized by a strong temporal asymmetry, several popular mathematical models neglect this issue. Maximum entropy methods provide a principled way of addressing time irreversibility, which leverages powerful results and ideas from the literature of non-equilibrium statistical mechanics. This tutorial provides a comprehensive overview of these issues, with a focus in the case of spike train statistics. We provide a detailed account of the mathematical foundations and work out examples to illustrate the key concepts and results from non-equilibrium statistical mechanics.

Keywords: non-equilibrium steady states; maximum entropy principle; spike train statistics; entropy production

1. Introduction

Being the brain one of the most complex systems within the observable universe, it is not surprising that there is still a large number of unanswered questions related to its structure and functions. With the aim of developing new ways of addressing such questions, there is an increasing consensus among neuroscientists in that interdisciplinary approaches are promising. As a prominent example of this, computational neuroscience has been greatly enriched during the last decades by tools, ideas and methods coming from statistical physics [1,2]. Moreover, these methods are recently being revisited with renewed interest due to the arrival of experimental techniques that generate huge volumes of data. In particular, neuroscientists have become progressively aware of the powerful computational techniques used by statistical physicists to analyze experimental data and large scale simulations.

When studying the firing patterns of collections of neurons, one of the most popular principles from statistical mechanics is the *maximum entropy principle* (MEP), which builds the least structured model that is consistent with average values measured from experimental data. These average values are usually restricted to firing rates and synchronous pairwise correlations, which gives rise to models composed by time independent and identically distributed (i.i.d) random variables, i.e., stochastic processes without temporal structure [3–5]. Needless to say, there exists strong evidence in favour of memory effects playing a major role in spike train statistics, and biological process in general [6–9]. Following this evidence, over the last years the study of complex biological systems has started to consider time-dependent processes where the past has an influence on future behavior [10–12]. The corresponding asymmetry between past and future is called the "arrow of time",

which is the unique direction associated with the irreversible flow of time that is noticeable in most biological systems.

Interestingly, the statistical physics literature has a fertile toolkit for studying time asymmetric processes [13]. First, one introduces the distinction between steady states that imply thermal equilibrium, and steady states that still carry fluxes—being called non-equilibrium steady states (NESS). Additionally, the extent to which a steady-state is not in equilibrium (i.e., the strength of its associated currents) can be quantified by the *entropy-production rate* [14], which is associated with the degree of time-irreversibility in the corresponding process [14]. Several studies have pointed out that being out-of-equilibrium is an important characteristic of biological systems [15–17]. Therefore, statistical characterizations consistent with the out-of-equilibrium condition should reproduce some degree of time irreversibility. One popular method that is suitable for studying these issues is Markov chain modeling [11,18–22].

Despite the potential of interdisciplinary pollination related to these fascinating issues, many scientists find it hard to explore these topics because of the major entry barriers, including differences in jargon, conventions, and notations across the various fields. To bridge this gap, this tutorial intends to provide an accessible introduction to the non-equilibrium properties of maximum entropy Markov chains, with an emphasis in spike train statistics. While not introducing novel material, the main added value of this tutorial is to present results of the field of non-equilibrium statistical mechanics in a pedagogical manner based on examples. These results have direct application to maximum entropy Markov chains, and may shed new light on the study of spike train statistics. This tutorial is suitable for researchers in the fields of physics or mathematics who are curious about the interesting questions and possibilities that computational neurosciences offers. The focus on this community is motivated by the growing community of mathematical physicists interested in computational neuroscience.

The rest of this tutorial is structured as follows. First, Section 2 introduces basic concepts of neural spike trains and Markov processes. Then, Section 3 introduces the notion of observable, and explores their fundamental properties. Section 4 introduces the core ideas of MEP, proposing the formal question and exploring methods for solving it. Section 5 studies various properties of interest of MEP models, including fluctuation-dissipation relationships, and their entropy production. Finally, Section 6 summarizes our conclusions.

2. Preliminary Considerations

This section introduces definitions, notations, and conventions that are used throughout the tutorial in order to give the necessary toolkit of ideas and notions to the unfamiliar reader.

2.1. Binning and Spike Trains

Consider a network of N spiking neurons, where time has been binned (i.e., discretized) in such a way that each neuron can exhibit no more than one action potential within one time bin Δt_b. Action potentials, or "spikes", are "all-or-none" events, and hence, spike data can be encoded using sequences of zeros and ones. A spiking state is denoted by $x_t^k = 1$, and corresponds to the event in which the k-th neuron spikes during the t-th time bin, while $x_t^k = 0$ implies that it remains silent.

A *spike pattern* is defined as the spike-state of all neurons at time bin t, and is denoted by $\boldsymbol{x}_t := \left[x_t^k\right]_{k=1}^N$. A *spike block* is a consecutive sequence of spike patterns, denoted by $\boldsymbol{x}_{t,r} := \left[\boldsymbol{x}_s\right]_{s=t}^r$ (see Figure 1). While the length of the spike block $\boldsymbol{x}_{t,r}$ is $r - t + 1$, it is useful to consider spike blocks of infinite length starting from time $t = 0$, which are denoted by $\boldsymbol{x}$. Finally, in this tutorial we consider that a *spike train* is an infinite sequence of spiking patterns. This assumption turns out to be useful because it allows us to put our analysis in the framework of stochastic processes, and because it also allows us to characterize asymptotic statistical properties.

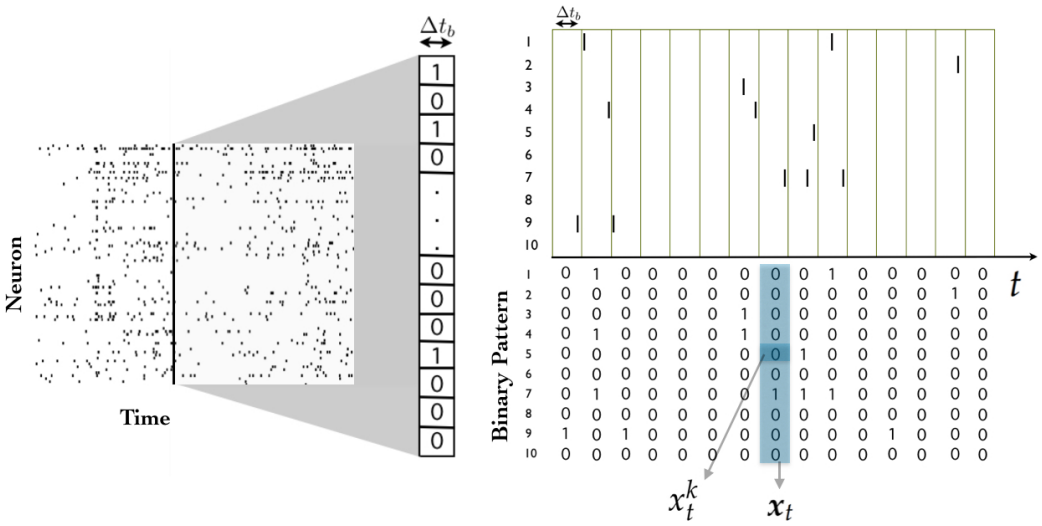

Figure 1. Illustration of a spike train, a spiking state and spike pattern. The time bin size Δt_b determine the binary patterns.

The set of all possible spike patters (or state space) in a network of N neurons is denoted by $\mathbb{S}$, and the set of all spike blocks of length R in a network of N neurons is denoted by $\mathbb{S}^R$.

Even at a single neuron level, for repetitions of the same stimulus, neurons respond randomly, but with a certain statistical structure. This is the main reason to look for statistical characterizations of spike trains. When trying to find a statistical representation considering a whole population of neurons responding simultaneously to a given stimulus, the problem is the following. Consider an experimental spike train from a network of N neurons where sequences of spike patterns are considered time-independent. The spike patterns can take 2^N values (state space). For $N > 10$ is not possible to observe all possible states in real experimental data nor computer simulations (2 h of recordings binned at 20 ms produce less than 2^{19} spike patterns). For $N = 100$ the state space is 2^{100}, therefore the frequentist approach is useless to estimate the invariant measure. Can we learn something about the statistics of spike patterns from data having access only to a very small fraction of the state space? The maximum entropy principle provides an answer to this question. This principle has been used in the context of spike train statistics mainly considering firing rates and synchronous pairwise correlations, which gives rise to trivial stochastic processes composed by (i.i.d) random variables [3–5]. However, as mentioned in the introduction, there exists strong evidence in favour of past events playing a role in spike train statistics, and the biological process in general [6–9,11]. This principle can be generalized considering non-synchronous correlations, affording to build Markov chains from data. This approach opens the way to a richer modeling framework that can afford to model time irreversibility (highly expected in biological systems) and to a remarkable mathematical machinery based on non-equilibrium statistical mechanics which can be used to characterize collective behavior and to explore the capabilities of the system. We focus our tutorial on non-equilibrium steady states in the context of maximum entropy spike trains. In the next section of this tutorial we present the elementary properties of Markov chains (our main object of analysis) which will be used in the next chapters to extract relevant information about the underlying neuronal network generating the data.

2.2. Elementary Properties of Markov Chains

A stochastic process is a collection of random variables $X_t \in \mathbb{S}$ indexed by $t \in T$ that often refers to time. The set $\mathbb{S}$ represents the phase-space of the process; in the case of stochastic processes representing spike trains, one usually takes $\mathbb{S} = \{0,1\}^N$. Moreover, considering the temporal binning

discussed in Section 2.1, usually $T = \mathbb{N}$ (the set of natural numbers) corresponds to the so-called discrete-parameter stochastic processes.

While spike trains can be characterized by stochastic processes dependent on an infinite past [23,24], Markov chains are particularly well-suited for modeling data sequences with finite temporal dependencies. In the next paragraphs we give the precise definition of a Markov process.

A stochastic process $(X_t : t \in \mathbb{N})$ defined on a measure space Ω is said to be a $\mathbb{P}$–Markov chain if it satisfies the *Markov property* (with respect to the probability measure $\mathbb{P}$): if, for every $t \in \mathbb{N}$ and for each sequence of states $x_0, x_1, \ldots, x_{t+1} \in \mathbb{S}$, the following relationship holds:

$$\mathbb{P}(X_{t+1} = x_{t+1} | X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = x_t) = \mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t) \,. \tag{1}$$

This property is usually paraphrased as: the conditional distribution of the future given the current state and all past events depends exclusively on the current state of the process. It is direct to show that the Markov property is equivalent to the following condition: for every increasing sequence of indices $(i_1 < i_2 < \ldots < i_n)$ in $\mathbb{N}$, and for arbitrary states $x_{i_1}, x_{i_2}, \ldots, x_{i_n}$ in $\mathbb{S}$, we have:

$$\mathbb{P}(X_{i_n} = x_{i_n} | X_{i_{n-1}} = x_{i_{n-1}}, \ldots, X_{i_1} = x_{i_1}) = \mathbb{P}(X_{i_n} = x_{i_n} | X_{i_{n-1}} = x_{i_{n-1}}) \,.$$

To characterize the transition probabilities, define a $\mathbb{S}$-*indexed stochastic matrix* to be a doubly indexed array of non-negative real numbers $P = (p(i,j) : i, j \in \mathbb{S})$ such that $\sum_{j \in \mathbb{S}} p(i,j) = 1$ for every $i \in \mathbb{S}$. It can be shown that a Markov chain is well-defined if the following is provided:

(i) An initial probability distribution, encoded by a vector $\mu := (\mu_i : i \in \mathbb{S})$.
(ii) A collection of $\mathbb{S}$-indexed stochastic matrices $\{P_t := (p_t(i,j))_{i,j \in \mathbb{S}} : t \in \mathbb{N}\}$.

Using these two elements, one can build probability measures P^n on $\mathbb{S}^n$ as follows,

$$P^n(i_0, i_1, \ldots, i_{n-1}) = \mu(i_0) \prod_{j=0}^{n-2} P_j(i_j, i_{j+1}) \,.$$

Furthermore, the Kolmogorov extension theorem [25] guarantees the existence of a unique probability measure $\mathbb{P}_\mu$ on $\mathbb{S}^{\mathbb{N}}$ such that the coordinate process satisfies:

$$\mathbb{P}_\mu(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = P^n(i_0, i_1, \ldots, i_n) \,,$$

and with respect to which $(X_t : t \in T)$ is a Markov chain. In this case $\mathbb{P}_\mu$ is said to be the probability law of the Markov chain $(X_t : t \in \mathbb{N})$. This notation also remarks that $\mathbb{P}_\mu$ is the law with initial distribution μ.

2.3. Homogeneity, Ergodicity and Stationarity

A Markov chain is said to be *homogeneous* if the transition matrices do not depend on the time parameter t, i.e., if there exists a $\mathbb{S}$-indexed stochastic matrix P such that $P_t = P$ for every $t \in T$. Note that if $(X_t : t \in T)$ is a $\mathbb{P}$–homogeneous Markov chain, then for every $t \in T$:

$$\mathbb{P}(X_{t+1} = j | X_t = i) = p(i,j) := p_{ij} \,. \tag{2}$$

In the rest of this paper we focus exclusively on homogeneous Markov chains, since this is the model assumed in the maximum entropy framework.

Consider now a homogeneous Markov chain $(X_t : t \in T)$ with initial distribution μ and transition matrix P. Moreover, consider $p_{ij}^{(m)}$ to be the (i,j)-th entry of the product matrix $P^m = P \cdot P \cdot \ldots \cdot P$. These quantities correspond to the m–steps transition probabilities. Equation (2) can be generalized to

$$\mathbb{P}(X_{t+m} = j | X_t = i) = p_{ij}^{(m)} \,.$$

A stochastic matrix P is said to be *ergodic* if there exists $k \in \mathbb{N}$ such that all the k–step transition probabilities are positive—i.e., there is a non-zero probability to go between any two states in k steps. A homogeneous Markov chain is ergodic if it can be defined by an initial distribution μ and an ergodic matrix.

Finally, a probability distribution π on $\mathbb{S}$ is called a *stationary distribution* for the Markov chain specified by P if

$$\pi P = \pi \, . \tag{3}$$

Equivalently, π is stationary for P if π is a left eigenvector of the transition matrix corresponding to the eigenvalue $\lambda = 1$, and is a probability distribution on $\mathbb{S}$. While it is true that 1 is always an eigenvalue of P, it may be the case that no eigenvector associated to it can be normalized to a probability distribution. Further conditions for existence and uniqueness will be given in the next paragraph. Finally, if a $\mathbb{S}$–indexed stochastic matrix P admits a stationary probability distribution π and $(X_t : t \in \mathbb{N})$ is a Markov chain with initial distribution π and transition matrix P, then for every $t \in \mathbb{N}$ and $i \in \mathbb{S}$:

$$\mathbb{P}_\pi(X_t = i) = \pi_i \, .$$

In this case $(X_t : t \in \mathbb{N})$ is said to be a stationary Markov chain, or that the Markov chain is started from stationarity.

The notion of homogeneous ergodic Markov chains is relevant in the context of spike train statistics because of the *Ergodic Theorem for finite-state Markov Chains*, which state that for all finite-state, homogeneous, ergodic Markov chains $(X_t : t \geq 0)$ with transition matrix P the following hold:

(a) There exists a unique stationary distribution π for P that satisfies that $\pi_i > 0$ for every $i \in \mathbb{S}$.
(b) For every $j \in \mathbb{S}$,

$$\lim_{m \to +\infty} p_{ij}^{(m)} = \pi_j \, .$$

Equivalently, for every distribution ν, $\lim_{t\to\infty} \mathbb{P}_\nu(X_t = j) = \pi_j$. This property guarantees the uniqueness of the maximum entropy Markov chain.

2.4. The Reversed Markov Chain

Given a discrete ergodic Markov chain, it is mathematically possible to define its associated time reversed Markov chain. Some Markov chains in the steady-state yield the same Markov chain (in distribution) if the time course is inverted and others do not. It has been argued multiple times that those Markov chains that are different from their time inverted version are better suited to represent biological stochastic processes [6,7,9,11,12].

Let $\overrightarrow{P}$ be a stochastic matrix, and assume that it admits a stationary probability measure π. Assume too that $\pi_i > 0$ for every $i \in \mathbb{S}$ (according to (a) in the Ergodic Theorem from the previous section, this is the case when $\overrightarrow{P}$ is ergodic). Define the $\mathbb{S}$–indexed matrix $\overleftarrow{P}$ with entries:

$$\overleftarrow{P}_{ij} = \frac{\pi_j}{\pi_i} \overrightarrow{P}_{ji} \, .$$

A direct calculation shows that $\overleftarrow{P}$ is also a stochastic matrix. Moreover, if π is stationary for $\overrightarrow{P}$, then it is for $\overleftarrow{P}$ as well.

Using the above facts, let $\mathbb{P}_\pi^{\rightarrow}$ and $\mathbb{P}_\pi^{\leftarrow}$ be the laws of two stationary Markov chains, denoted by X_t and Y_t, whose stationary distribution is π and transition probabilities are $\overrightarrow{P}$ and $\overleftarrow{P}$, respectively. The following holds

$$\begin{aligned}\mathbb{P}_\pi^{\leftarrow}(Y_0 = i_0, Y_1 = i_1, \ldots, Y_n = i_n) &= \pi_{i_0} \overleftarrow{P}_{i_0 i_1} \overleftarrow{P}_{i_1 i_2} \cdots \overleftarrow{P}_{i_{n-1} i_n} \\ &= \pi_{i_0} \frac{\pi_{i_1}}{\pi_{i_0}} \overrightarrow{P}_{i_1 i_0} \frac{\pi_{i_2}}{\pi_{i_1}} \overrightarrow{P}_{i_2 i_2} \cdots \frac{\pi_{i_n}}{\pi_{i_{n-1}}} \overrightarrow{P}_{i_n i_{n-1}} \\ &= \pi_{i_n} \overrightarrow{P}_{i_n i_{n-1}} \overrightarrow{P}_{i_{n-1} i_{n-2}} \cdots \overrightarrow{P}_{i_1 i_0} \\ &= \mathbb{P}_\pi^{\rightarrow}(X_0 = i_n, X_1 = i_{n-1} \ldots, X_n = i_0)\,.\end{aligned}$$

By virtue of this result, it is natural to call the chain $(Y_t : t \geq 0)$ the *reversed chain* associated to $(X_t : t \geq 0)$.

2.5. Reversibility and Detailed Balance

A transition matrix P is *reversible* with respect to π if the associated Markov chain started from π has the same law as the reversed chain started from the same distribution. The reversibility of P with respect to π is equivalent to the condition of *detailed balance*, given by

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \forall i, j \in \mathbb{S}\,. \tag{4}$$

Note that any probability measure π that satisfies detailed balance with respect to P is necessarily stationary, since

$$\sum_{i \in \mathbb{S}} \pi_i P_{ij} = \sum_{i \in \mathbb{S}} \pi_j P_{ji} = \pi_j \sum_{i \in \mathbb{S}} P_{ji} = \pi_j \quad \text{for every } j \in \mathbb{S}\,.$$

The converse is, however, not true in general: a stationary distribution may not satisfy Equation (4).

Intuitively, Equation (4) states that, in the stationary state, the fluxes between each pair of states balance each other. In contrast, detailed balance is broken when there is a cycle of three or more states in the state space supporting a net probability current—even in the steady state. Detailed balance is also interpreted as "time reversibility", as one could not distinguish the steady state dynamics of the system when going forward or backward in time. Certainly, this property is not expected in stochastic processes generated by biological systems. Several disciplines use the term "equilibrium" to refer to long-term behaviour, i.e., what is not transient. In this tutorial we use the term *equilibrium state* exclusively to refer to probability vectors that satisfy the detailed balance conditions—given in Equation (4). Markov chains that satisfy the detailed balance condition are referred as equilibrium steady states, and conversely, steady states that do not satisfy the detailed balance conditions are called *Non-Equilibrium Steady States* (NESS).

How to characterize (finite state, homogeneous) reversible Markov chains? Following [26], consider any finite graph $(\mathbb{S}, (c_{ij})_{i,j \in \mathbb{S}})$, with vertex set $\mathbb{S}$ and with the edge between vertices i and j labelled by the non-negative edge $c_{ij} = c_{ji}$. The graph can be visualized as a system of points labelled by $\mathbb{S}$, and with a line segment between points whenever the corresponding conductance is positive. Define $c_i = \sum_{j \in \mathbb{S}} c_{ij}$ and the $\mathbb{S}$–indexed stochastic matrix given by

$$p_{ij} = \frac{c_{ij}}{c_i}\,,$$

Now define $C = \sum_{i\in\mathbb{S}} c_i$. It is straightforward to prove that P is reversible with respect to the probability measure given by

$$\pi_i = \frac{c_i}{C}\,,$$

and thus it is stationary for P. The unique Markov chain started from π and transition matrix P is called the stationary random walk on the network $(\mathbb{S}, (c_{ij})_{i,j\in\mathbb{S}})$. Conversely, any reversible $\mathbb{S}$–valued Markov chain can be identified with the random walk on the graph with vertex set $\mathbb{S}$ and edges given by $c_{ij} = c_{ji} = \pi_i p_{ij}$.

2.6. Law of Large Numbers for Ergodic Markov Chains

The Law of Large Numbers (LLN) that applies to independent and identically distributed random variables (i.i.d.) can be extended to the realm of ergodic Markov chains. In effect, for a given ergodic Markov chain $(X_t : t \geq 0)$ with stationary distribution π and transition matrix P, define the random variables $N_i^{(T)}$ equal to the number of occurrences of the state i up to time $T-1$, i.e.,

$$N_i^{(T)} = \sum_{t=0}^{T-1} \mathbf{1}_{\{X_t=i\}},$$

where $\mathbf{1}_{\{.\}}$ is an indicator function. Similarly, define the random variables $N_{ij}^{(T)}$ as the number of occurrences of the consecutive pair of states $(i,j) \in \mathbb{S}^2$—in that order—up to time $T-1$, i.e.,

$$N_{ij}^{(T)} = \sum_{t=1}^{T-1} \mathbf{1}_{\{X_{t-1}=i, X_t=j\}}\,.$$

With this, the *Strong Law of Large Numbers for Markov chains* can be stated as follows: if $(X_t : t \geq 0)$ is ergodic and π is its unique stationary distribution, then

$$\mathbb{P}_\mu\left(\lim_{T\to+\infty} \frac{N_i^{(T)}}{T} = \pi_i\right) = 1 \quad \text{and} \quad \mathbb{P}_\mu\left(\lim_{T\to+\infty} \frac{N_{ij}^{(T)}}{T} = \pi_i p_{ij}\right) = 1,$$

holds for any initial distribution μ. The result, in turn, implies the *Weak Law of Large Numbers for Markov chains*, which state that, following the above notation, for every $\varepsilon > 0$ and for every starting distribution μ:

$$\lim_{T\to+\infty} \mathbb{P}_\mu\left(\left|\frac{N_i^{(T)}}{T} - \pi_i\right| > \varepsilon\right) = 0 \quad \text{and} \quad \lim_{T\to+\infty} \mathbb{P}_\mu\left(\left|\frac{N_{ij}^{(T)}}{T} - \pi_i p_{ij}\right| > \varepsilon\right) = 0\,.$$

Let's denote by $C(\mathbb{S})$ the space of real-valued functions on $\mathbb{S}$. Clearly, any function of $C(\mathbb{S})$ can be written as $f(x) = \sum_{i\in\mathbb{S}} a_i \mathbf{1}_i(x)$ for certain constants a_i, $i \in \mathbb{S}$. Then, the above result generalizes as: for every $f \in C(\mathbb{S})$, ergodic chain X_t, and probability distribution μ, the following holds:

$$\mathbb{P}_\mu\left(\lim_{T\to+\infty} \frac{1}{T}\sum_{t=0}^{T-1} f(X_t) = \mathbb{E}_\pi(f(X_0))\right) = 1\,.$$

This corresponds to a particular form of the Birkhoff Ergodic Theorem, which is briefly outlined in the next section and is relevant to characterize spike trains of observables as from data it is possible to accurately measure average values of firing rates and correlations. For an ergodic stationary Markov chain with a state space relatively small with respect to the sample size, this theorem guarantees that

from a large sample the transition probabilities and the invariant measure can be recovered. This is not the case in spike train statistics at the population level as only a very small fraction of the state space is sampled in experimental spike trains. However, some features of the spike trains can be sampled very accurately from experimental data. In the next section, we present the basic elements to build from these characteristics a statistical model of the entire population.

3. Observables of Markov Chains and Their Properties

The notion of observable plays a central role in the study of maximum entropy spike trains. This section discusses their nature and fundamental properties.

3.1. Observables and Their Empirical Averages

Suppose a spiking neuronal network of N neurons is provided. Suppose too that measurements of spike patterns for T time bins have been performed. The observables are real-valued functions over the possible spike blocks, denoted here by $\mathbb{B} := \mathbb{S}^T$. Let $C(\mathbb{B})$ be the space of such observables, i.e., the linear space of real-valued functions $f : \mathbb{B} \mapsto \mathbb{R}$. Recall the space $C(\mathbb{S})$ of observables of range 1, discussed at the end of the above section. This space can be naturally embedded into $C(\mathbb{B})$; thus, it can be considered as a linear subspace of the latter. More generally, the space of observables of range R for $R \leq T$, denoted $C(\mathbb{S}^R)$, is just the space of real-valued functions on $\mathbb{S}^R$, that we identify with its image through the natural embedding into $C(\mathbb{B})$.

We are interested in the average of observables with respect to several probability measures. If μ is a probability measure on $\mathbb{B}$ (i.e., $\mu(\omega) \geq 0$ and $\sum_{\omega\in\mathbb{B}} \mu(\omega) = 1$) and f an observable of range $R \leq T$ i.e., $f \in C(\mathbb{S}^R)$, we define its expectation with respect to μ as

$$\mu(f) = \mathbb{E}_\mu\{f\} := \sum_{\omega\in\mathbb{B}} f(\omega)\mu(\omega).$$

Since the space of blocks of length T is finite, the above sum is always finite, and thus our definition makes sense for every probability measure on $\mathbb{B}$.

In the context of spike-trains, an important class of observable is made up of $\{0,1\}$-valued functions. It can be proved that any finite-range binary observable can be written as a finite sum of finite products of functions of the form $\mathbf{1}_{\{X_i^{(j)}=1\}}$ that represents the event that the j-th neuron fires during the i-th bin. The average value if this observable is known as the firing rate of neuron j. This quantity has been proposed as one of the major neural coding strategies used by the brain [27].

Consider a spike block $\boldsymbol{x}_{0,T-1}$, where T is the sample length. Although in most cases the probability measure μ that characterizes the spiking activity is not known, it is meaningful to use the experimental data to estimate the mean values of specific observables. The range of the validity of this procedure is usually based on prior assumptions about the nature of the source that originates the sample. For example, it can be assumed that the sample is a short piece of an infinite path that comes running from the far past, and so it can be assumed that this piece exhibits a behavior that is close to the stationary distribution. In this case, one can consider for any number $R \leq T$ the quantity:

$$Q(y_0, y_1, \ldots, y_{R-1}) = \sum_{j=0}^{T-R} \mathbf{1}_{\{\boldsymbol{x}_{j,j+R-1}=(y_0,y_1,\ldots,y_{R-1})\}},$$

that counts the number of appearances of the sequence $(y_0, \ldots, y_{R-1})$ as a consecutive sub-sequence of $\boldsymbol{x}_{0,T-1}$. Now, for any set $A \subseteq \mathbb{S}^R$, define:

$$\mu_{\boldsymbol{x}_{0,T-1}}(A) := \frac{1}{T-R+1} \sum_{\boldsymbol{y}\in A} Q(y),$$

$$\mu_{\boldsymbol{x}_{0,T-1}}(f) = A_T(f) = \frac{1}{T-R+1}\sum_{i=0}^{T-R} f(\boldsymbol{x}_{i,R-1+i}).$$

When the empirical distribution is not explicitly stated, it is customary to write $\langle f \rangle$ to denote the average of the observable f with respect to this probability measure.

3.2. Moments and Cumulants

Observables are random variables whose average values can be determined from experimental data or from the explicit representation of the underlying measure characterizing the stochastic process generating the data. Important statistical properties of random variables are encoded in the *cumulants*. We will use the cumulants in Section 5 of this tutorial to characterize and infer properties of maximum entropy Markov chains. Let us now introduce them.

The moment of order r of a real-valued random variable X is given by $m_r = \mathbb{E}(X^r)$, for $r \in \mathbb{N}$ (here we freely use the notation $\mathbb{E}$ to denote the expectation with respect to a probability measure that should be inferred from the context). The moment generating function (or Laplace transform) of a random variable is defined by:

$$M(t) = \mathbb{E}(e^{tX}),$$

and provided it is a function of t with continuous derivatives of arbitrary order at 0, we have that:

$$m_r = \left(\frac{d^r}{dt^t} M\right)_{t=0}.$$

The cumulants κ_r are the coefficients in the Taylor expansion of the cumulant generating function. The cumulants are defined as the logarithm of the moment generating function, namely,

$$\ln M(t) = \sum_r \kappa_r t^r / r!\,.$$

The relation between the moments and cumulants is obtained extracting coefficients from the Taylor expansion, i.e.,

$$\kappa_r = \left(\frac{d^r}{dt^r} \ln(M(t))\right)_{t=0} \tag{5}$$

which yields the first values:

$$\begin{aligned}
\kappa_1 &= m_1\,,\\
\kappa_2 &= m_2 - m_1^2\,,\\
\kappa_3 &= m_3 - 3m_2m_1 + 2m_1^3\,,\\
\kappa_4 &= m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - 6m_1^4,
\end{aligned}$$

and so on. In particular, the first four cumulants are the mean, the variance, the skewness and the kurtosis.

3.3. Observables and Ergodicity

Let $\theta : \Omega \mapsto \Omega$ be the shift operator that acts on a sequence $\omega \in \Omega$ as:

$$(\theta(\omega))_i = \omega_{i+1},$$

i.e., θ shifts the sequence one position to the left. Now, assume that the Markov chain $(X_t : t \geq 0)$ is ergodic. Let π be its unique stationary probability distribution. The Birkhoff Ergodic Theorem states that under the above assumptions, for every $f \in C(\mathbb{B})$:

$$\mathbb{P}_\mu \left(\lim_{N \to +\infty} \frac{1}{N} \sum_{n=0}^{N-1} f \circ \theta^n = \mathbb{E}_\pi(f) \right) = 1,$$

for every initial measure μ. This equation means that under the ergodic hypothesis, the temporal averages converge to the spatial averages. The importance of this fundamental result should not be underestimated since this result supports the practice of regarding averages of (hopefully) large samples of experimental data as faithful approximations of the *true* values of the expectations of the observables.

3.4. Central Limit Theorem for Observables

Consider an arbitrarily large sequence of spike patterns of N neurons. Consider $t \in \mathbb{N}$ and let $\boldsymbol{x}_{0,t-1}$ be the spike-block of length t. Also, let f be an arbitrary observable of fixed range R. The asymptotic properties of $A_t(f)$ are established in the following context: the finite sample is drawn from an ergodic Markov chain, i.e., $\boldsymbol{x} \sim \mathbb{P}_\nu$, where $\mathbb{P}_\nu$ is the Markov probability measure of an ergodic chain $(X_t : t \geq 0)$ started from an arbitrary initial distribution. Let π be the unique stationary measure for the Markov chain. Observe that by virtue of the ergodic assumption, the empirical averages of observables become more accurate as the sampling size grows, i.e.,

$$\mathbb{P}_\nu \left(A_t(f) \to \mathbb{E}_\pi\{f\} \right) = 1.$$

for any starting condition ν. However, the above result does not clarify the rate at which the accuracy improves. The central limit theorem (CLT) for ergodic Markov chains provides a result to approach this issue (for datails see [28]).

Theorem 1 (Central limit theorem for ergodic Markov chains). *Under the above assumptions, and keeping notation, define:*

$$\sigma = \sqrt{\left(\mathbb{E}_\pi((f(X_0, \ldots, X_{R-1}) - \mathbb{E}_\pi(f(X_0, \ldots, X_{R-1}))^2) \right)} \, .$$

Let L_t be the law of the random variable $\frac{\sqrt{t}}{\sigma}\left[A_t(f) - \mathbb{E}_\pi\{f\}\right]$ under the measure $\mathbb{P}_\nu$ of an ergodic Markov chain started from an arbitrary distribution. Let L be the law of a standard normal random variable. Then $L_t \to L$ in the sense of weak convergence of convergence in distribution. This is usually written as:

$$\mathbb{P}_\nu \left\{ \frac{\sqrt{t}}{\sigma}\left[A_t(f) - \mathbb{E}_\pi\{f\}\right] \leq x \right\} \to \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{s^2}{2\sigma}} ds.$$

This theorem implies that "typical" fluctuations of $A_t(f)$ around its long term average $\mathbb{E}_\pi\{f\}$ are of the order of $\sigma/\sqrt{t}$. For spike trains, this theorem quantifies the expected Gaussian fluctuations of observables in terms of the sample size of the experimental data.

3.5. Large Deviations of Average Values of Observables

Although the CLT for ergodic Markov chains is precise in describing the typical fluctuations around the mean, it does not characterize the probabilities of large fluctuations. While it is clear that the probability of large fluctuations of average values vanish as the sample size increases, it is sometimes relevant to characterize the decrease rate of this probability. That is what the large deviation principle (LDP) does.

Let f be a function of finite range defined on the space of sequences. In many situations, f will be a $\{0,1\}$–valued function. Let $\mathbb{P}_\pi$ be the probability measure on the space of sequences induced by an ergodic Markov chain with stationary probability π. The empirical average $A_t(f)$ satisfies a large deviation principle (LDP) with rate function I_f, defined as

$$I_f(s) := -\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}_\pi\{A_t(f) > s\}, \tag{6}$$

if the above limit exists. The above condition implies for large t that $\mathbb{P}_\pi\{A_t(f) > s\} \approx e^{-tI_f(s)}$. In particular, if $s > \mathbb{E}_\pi\{f\}$ the Law of Large Numbers (LLN) ensure that $\mathbb{P}_\pi\{A_t(f) > s\}$ goes to zero as t increases, but the rate function quantifies the speed at which this occurs.

Calculating I_f using the definition (Equation (6)), is usually impractical. However, the Gärtner-Ellis theorem provides a clever alternative to circumvent this problem [29]. Let us introduce the *scaled cumulant generating function* (SCGF) associated to the random variable f by

$$\lambda_f(k) =: \lim_{t\to\infty} \frac{1}{t} \ln \mathbb{E}_\pi \left[e^{tkA_t(f)} \right], \quad k \in \mathbb{R}, \tag{7}$$

when the limit exists (details about cumulant generating functions are found in [30]). While the empirical average $A_t(f)$ is taken over a sample (empirical measure), the expectation in (7) is computed over the probability distribution given by $\mathbb{P}_\pi\{\cdot\}$.

Theorem 2 (**Gärtner-Ellis theorem**). *If λ_f is differentiable, then the average $A_t(f)$ satisfies a LDP with rate function given by the Legendre transform of λ_f, that is*

$$I_f(s) = \max_{k\in\mathbb{R}} \{ks - \lambda_f(k)\}. \tag{8}$$

Therefore, the large deviations of empirical averages $A_t(f)$ can be characterized by first computing their SCGF and then finding their Legendre transform.

A useful application of the LDP is to estimate the likelihood that the empirical average $A_t(f)$ takes a value far from its expected value. Let us assume that $I_f(s)$ is a positive differentiable convex function. Then, $\lambda_f(k)$ is also differentiable [31] (for a comprehensive discussion about the differentiability of $\lambda_f(k)$ see [30].) Then, as $I_f(s)$ is convex it has a unique global minimum. Denoting this minimum by s^*, from the differentiability of $I_f(s)$ it follows that $I_f(s^*) = 0$. Additionally, it follows from properties of the Legendre transform that $s^* = \lambda_f'(0) = \mathbb{E}_p\{f\}$, which is the LLN that says that $A_t(f)$ concentrates around s^*. Consider $s \neq s^*$ and that $I_f(s)$ admits a Taylor expansion around s^*

$$I_f(s) = I_f(s^*) + I_f'(s^*)(s-s^*) + \frac{I_f''(s^*)(s-s^*)^2}{2} + O(s-s^*)^3 .$$

As s^* is zero and a minimum of $I(s)$, the first two terms of this expansion are zero, and as $I(s)$ is convex $I''(s) > 0$. For large t, it follows from (6) that

$$\begin{aligned} p\{A_t(f) > s\} &\approx e^{-tI_f(s)} \\ &\approx e^{-t\left(\frac{I_f''(s^*)(s-s^*)^2}{2}\right)}, \end{aligned} \tag{9}$$

so the "small deviations" (we are using Taylor expansion) of $A_t(f)$ around s^* are Gaussian (in Equation (9) $1/I_f''(s^*) = \lambda_f''(0) = \sigma^2$). In this sense, the LDP can be considered as an extension of the CLT as it goes beyond the small deviations around s^* (Gaussian), but additionally the large deviations (not Gaussian) of $A_t(f)$.

4. Building Maximum Entropy Temporal Models

This section presents the main concepts behind the construction of maximum entropy models for temporal data. The next Section 4.1, introduces the concept of entropy, and then Section 4.2 formulates the problem of maximizing the entropy rate. Methods for solving this problem are discussed in Section 4.3, which are then illustrated in an example presented in Section 4.4.

4.1. The Entropy Rate of a Temporal Model

4.1.1. Basic Definitions

In order to give mathematical meaning to the rather vague notion of uncertainty, a natural approach is to employ the well-established notion of *Shannon entropy*. For any probability measure p defined over the state space E (not necessarily $\mathbb{S}$), the Shannon entropy of p is given by

$$S[p] := -\sum_{x \in E} p(x) \log p(x) \, .$$

Note that this definition can be used for measures on the spaces of infinite sequences $E^{\mathbb{N}}$. However, as in most cases of interest, the value saturates in infinite. A better suited notion in this context is given by the *entropy rate*, which plays a crucial role in the rest of this tutorial.

Definition 1 (entropy rate). *Let μ be a probability measure on the space of sequences $\mathbb{S}^{\mathbb{N}}$. For $n \geq 1$ let μ_n be the probability measure induced by μ on the initial n coordinates, i.e., μ_n is the probability distribution on E^n given by:*

$$\mu_n(x_0, x_1, \ldots, x_n - 1) = \mu\left(\omega \in \mathbb{S}^{\mathbb{N}} : X_i = x_i \, for \, i = 0, 1, \ldots, n-1\right).$$

The entropy rate of the measure μ is defined by:

$$\mathcal{S}[\mu] = \lim_{n \to \infty} \frac{1}{n} S[\mu_n]. \tag{10}$$

The above definition applies to any probability distribution on the space of sequences. Intuitively, the entropy rate correspond to the entropy per time unit, and represents how much "uncertainty" is created by the process as time moves forward.

4.1.2. The Entropy Rate of I.I.D. and Markov Models

Let us consider first a null model of spike activity, where there is complete statistical independence between two consecutive spike patters. For this, first recall that $\mathbb{S} = \{0,1\}^N$, where N is the fixed number of neurons. Without loss of generality, we can enumerate the elements of $\mathbb{S}$ as $s_1, s_2, \ldots, s_{2^N}$. Let $\nu = (\nu_1, \nu_2, \ldots, \nu_{2^N})$ be a probability measure on $\mathbb{S}$ such that:

$$\nu(s_k) = \nu_k$$

For a T–block $x = (x_0, x_1, \ldots, x_{T-1}) \in \mathbb{S}^T$ and for every $s \in \mathbb{S}$, we set:

$$N_s^T(x) = \sum_{i=0}^{T-1} \mathbf{1}_{\{x_i = s\}}.$$

On the space of infinite spike trains $\mathbb{S}^{\mathbb{N}}$ we consider the probability $\mu = \nu^{\otimes \mathbb{N}}$, i.e., the product measure on the space of spike trains. Observe that the induced measure is given by:

$$\mu_n(x_0, x_1, \ldots, x_{T-1}) = \prod_{k=1}^{2^N} \nu_k^{N_s^t(x_0,\ldots,x_{T-1})} .$$

With this, a straightforward calculation shows that

$$\mathcal{S}[\mu] = S[\nu] = -\sum_{k=1}^{2^N} \nu_k \ln(\nu_k) ,$$

and in this case we observe that the entropy rate is equal to the entropy of the probability distribution induced by each coordinate map.

A reasonable next step in the hierarchy of models is to weaken the independence hypothesis and assume instead that the spike activity keeps some bounded memory of the past. For this, following the considerations of Section 2, let us consider an ergodic discrete Markov chain with transition matrix P and invariant distribution π taking values in $\mathbb{S}$. Let $\mu = \mu(P, \pi)$ be the measure induced by this chain on the space $\mathbb{S}^{\mathbb{N}}$. Observe that, with the above notation:

$$\mu_n(x_0, x_1, \ldots, x_{n-1}) = \pi_{x_0} \prod_{j=1}^{n-1} P_{x_{j-1}x_j}.$$

A direct computation shows that

$$\begin{aligned} S[\mu_1] &= -\sum_{(x_0,x_1)\in\mathbb{S}^2} \pi_{x_0} P_{x_0x_1} \ln(\pi_{x_0} P_{x_0x_1}) \\ &= -\sum_{x\in\mathbb{S}} \pi_x \ln(\pi_x) - \sum_{(x_0,x_1)\in\mathbb{S}^2} \pi_{x_0} P_{x_0x_1} \ln(P_{x_0x_1}), \end{aligned}$$

and induction shows that:

$$S[\mu_n] = -\sum_{x\in\mathbb{S}} \pi_x \ln(\pi_x) - n \sum_{(x_0,x_1)\in\mathbb{S}^2} \pi_{x_0} P_{x_0x_1} \ln(P_{x_0x_1}).$$

Thus dividing by n and taking the limit in Equation (10), one finds that

$$\mathcal{S}[\mu] = -\sum_{(x_0,x_1)\in\mathbb{S}^2} \pi_{x_0} P_{x_0x_1} \ln(P_{x_0x_1}).$$

4.2. Entropy Rate Maximization under Constraints

Now we introduce the central problem of this tutorial. Assume we have empirical data from spiking activity. Consider the empirical averages of K observables, $\langle f_k \rangle$, for $f_k, k = 1, \ldots, K$. We need to characterize the Markov chains that are consistent with these average values. Except for trivial and uninteresting situations, there is no finite set of empirical averages that uniquely determines a distribution μ on $\mathbb{S}^{\mathbb{N}}$ that fits the averages, in the sense that

$$\mu(f_k) = \langle f_k \rangle \text{ for } k = 1, \ldots, K.$$

Consequently, we need to impose further restrictions in order to guarantee uniqueness. A useful and meaningful approach is the so-called Maximum Entropy Markov Chain model (MEMC), which fit the unique probability measure μ among all the stationary Markov measures ν on $\mathbb{S}^{\mathbb{N}}$ that match the

expected values of a given set of observables and that maximizes the entropy rate. Mathematically, it is written in the following form:

$$\begin{aligned} &\max_{\nu \in \mathcal{M}_{inv}} && \mathcal{S}[\nu] \\ &\text{subject to} && \nu(f_k) = \langle f_k \rangle_e = C_k, \quad \forall k \in \{1, \ldots, K\}, \end{aligned}$$

where $\mathcal{M}_{inv}$ is a shorthand for the sets of stationary Markov measures on $\mathbb{S}^{\mathbb{N}}$. Formally:

$$\mathcal{M}_{inv} := \{(\pi, P) : \pi \text{ is a probability on } \mathbb{S}, P \text{ is stochastic, } \pi P = \pi\}.$$

It is to be noted that the maximum entropy principle can be derived in some scenarios from more general principles based on large deviation theory [30]. In this framework, entropy maximization corresponds to Kullback-Leiber divergence minimization. This approach can be useful for accounting additional information that is not in the form of functional constraints, but as a Bayesian prior. A major drawback of this approach to be applied to spike trains, is that it assumes stationarity in the data. While this condition is not to be naturally expected in biological systems, controlled experiments can be carried out in the context of spike train analysis in order to maintain these conditions [3,8,32,33]. The maximum entropy principle as presented here is useful only in the stationary case. However, some extensions have been proposed [34,35]. Note also that there are alternative variational principles which can be used to find distributions that extremize the value of quantities such as the maximum entropy production principle [36–38], or the Prigogine minimum entropy production principle [39,40]. To the best of our knowledge, these alternatives have not yet been explored in the context of spike train statistics.

4.3. Solving the Optimization Problem

We now discuss techniques for finding models that maximize the entropy rate.

4.3.1. Lagrange Multipliers and the Variational Principle

To solve the above optimization problem, let us introduce the set of Lagrange multipliers $h_k \in \mathbb{R}$ and an *energy* function $\mathcal{H} = \sum_{k=1}^{K} h_k f_k$, which is a linear combination of observables. Consider the following unconstrained optimization problem, which can be framed in the context of the *variational principle* of the thermodynamic formalism [41]:

$$\mathcal{F}[\mathcal{H}] = \sup_{\nu \in \mathcal{M}_{inv}} \Big\{ \mathcal{S}[\nu] + \nu(\mathcal{H}) \Big\} = \mathcal{S}[\mu] + \mu(\mathcal{H}), \tag{11}$$

where $\mathcal{F}[\mathcal{H}]$ is called the *free energy* and $\nu(\mathcal{H}) = \sum_{k=1}^{K} h_k \nu(f_k)$ is the average value of $\mathcal{H}$ with respect to the measure ν. The following holds:

$$\frac{\partial \mathcal{F}[\mathcal{H}_{\boldsymbol{h}}]}{\partial h_k} = \mathbb{E}_p\{f_k\} = C_k, \quad \forall k \in \{1, ..., K\},$$

where $\mathbb{E}_p\{f\}$ is the average of f_k with respect to p (maximum entropy measure), which is equal (by restriction) to the average value of f_k with respect to the empirical measure from the data.

The maximum-entropy (ME) principle [42] has been successfully applied to spike data from the cortex and the retina [3,8,9,11,12,43]. The approach starts by fixing the set of constraints determined by the empirical average of observables measured from spiking data. Maximizing the entropy (concave functional) under constraints, gives a unique distribution. The choice of observables to measure in the empirical data (constraints) determines the statistical model. The approach of Lagrange multipliers may not be practical when trying to fit a MEMC. In the next section we introduce an alternative optimization based on spectral properties.

4.3.2. Transfer Matrix Method

In order to illustrate the transfer matrix method, we start with a classical example that allow us to introduce a fundamental definition. Let A be a adjacency matrix i.e., a $\{0,1\}$-valued square matrix with rows and columns indexed by the elements of $\mathbb{S}$. If there exists an $n \geq 0$ such that

$$A_{ij}^n > 0$$

for every $i, j \in \mathbb{S}$, we say that A is *primitive*. The next well-known theorem of Linear Algebra is crucial [44] for the uniqueness of the MEMC.

Theorem 3 (**Perron-Frobenius theorem**). *Let A be a primitive matrix. Then,*

- *There is a positive maximal eigenvalue $\rho > 0$ such that all other eigenvalues satisfy $\mid \rho' \mid < \rho$. Moreover ρ is simple;*
- *There are positive left- and right-eigenvectors $u = (u_1, \ldots, u_k), v = (v_1, \ldots, v_k)$ s.t. $uA = \rho\, u, Av = \rho v$.*

Apply the above theorem to a primitive matrix A, and define:

$$P_{ij} = \frac{A_{ij} v_j}{\rho v_i}; \quad \pi_i = \frac{u_i v_i}{\langle u, v \rangle},$$

where $\langle u, v \rangle$ is the standard inner product in $\mathbb{R}^{2^N}$ (we refer the reader to [44] for details). The matrix P built above is stochastic. Moreover, π is its unique stationary measure. Define the Parry measure to be the Markov measure:

$$\mu(i_0, i_1, \ldots, i_n) = \pi_{i_0} P_{i_0 i_1}, \ldots, P_{i_{n-1} i_n}.$$

It is well known that the Parry measure is the unique measure of maximal entropy consistent with the adjacency matrix A [45,46].

Inspired by this result, we consider now the general case. Consider constraints given by a set of empirical averages of observables, as explained in the previous section. The above example certainly fits this setting: just consider binary observables associated to each pair of states (i, j) that evaluates to 1 when a transition from state i to state j has been observed in the data. In our general setting, we assume that the chosen observables have a finite maximum range R. From these observables the energy function $\mathcal{H}$ of finite range R is built as a linear combination of these observables. Using this energy function we build a matrix denoted by $\mathcal{L}_\mathcal{H}$, so that for every $y, w \in \mathbb{S}^R$ its entries are given as follows:

$$\mathcal{L}_\mathcal{H}(y, w) = \begin{cases} e^{\mathcal{H}(y_1 w_{1,R-1})} & \text{if } y_{1,R-1} = w_{0,R-2} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

where $y_1 w_{R-1}$ is the concatenated block built from y_1 and $w_{1,R-1}$. For observables of range one, the matrix above is defined as $\mathcal{L}_\mathcal{H}(y, w) = e^{\mathcal{H}(y)}$. Assuming $\mathcal{H} > -\infty$, the elements of the matrix $\mathcal{L}_\mathcal{H}$ are non-negative. Furthermore, in every non trivial case, the matrix is primitive and satisfies the Perron-Frobenius theorem [44]. Denote by ρ the unique largest eigenvalue of $\mathcal{L}_\mathcal{H}$. Just as above, we denote by $\mathbf{u}$ and $\mathbf{v}$ the left and right eigenvectors of $\mathcal{L}_\mathcal{H}$ associated to ρ. Notice that $u_i > 0$ and $v_i > 0$, for all $i \in \mathbb{S}$. The *free energy* associated to a transfer matrix is the logarithm of the unique maximum eigenvalue.

The matrix $\mathcal{L}_\mathcal{H}$ can be turned into a Markov matrix of maximum entropy. For a primitive matrix M with spectral radius ρ, and positive right eigenvector $\mathbf{v}$ associated to ρ, the stochastic matrix built from M is computed as follows:

$$\mathfrak{S}(M) = \frac{1}{\rho} D^{-1} M D \,,$$

where D is the diagonal matrix with entries $D_{ii} = v_i$. The MEMC transition matrix P and unique stationary probability measure π are explicitly given by

$$P = \mathfrak{S}(\mathcal{L}_{\mathcal{H}}); \quad \pi_i := \frac{u_i \, v_i}{\langle u, v \rangle}, \quad \forall i \in \mathbb{S}. \tag{13}$$

Note that when $\mathcal{H} = 0$, the MEMC is characterized by the Markov transition matrix with components [47]:

$$P_{ij} = \frac{A_{ij} v_j}{\rho v_i},$$

where A is the adjacency matrix.

4.3.3. Finite Range Gibbs Measures

For a fixed energy function $\mathcal{H}$ of range $R \geq 2$, there is a unique stationary Markov measure μ for which there exist a constant $\gamma \geq 1$ such that [46],

$$\gamma^{-1} \leq \frac{\mu[\boldsymbol{x}_{1,n}]}{\exp(\sum_{k=1}^{n-R+1} \mathcal{H}(\boldsymbol{x}_{k,k+R-1}) - (n+R-1)\mathcal{F}[\mathcal{H}])} \leq \gamma, \tag{14}$$

that attains the supremum (11). The measure μ, as defined by (14), is known in the symbolic dynamics literature as *Gibbs measure in the sense of Bowen* [48]. All MEMCs belong to this class of measures. Moreover, the classical Gibbs measures in statistical mechanics are particular cases of (14), when $\gamma = 1$, $\mathcal{F}[\mathcal{H}] = \log Z$ and $\mathcal{H}$ is an energy function of range one, leading to an i.i.d stochastic process characterized by the product measure μ. In this case the following holds:

$$\mu(\boldsymbol{x}) = \frac{e^{\mathcal{H}(\boldsymbol{x})}}{Z} \quad \forall \boldsymbol{x} \in \mathbb{S}; \quad Z = \sum_{\boldsymbol{x} \in \mathbb{S}} e^{\mathcal{H}(\boldsymbol{x})}.$$

The free energy that is defined here has a deep relationship with the free energy in thermodynamics. Consider a thermodynamic system in equilibrium. The Helmholtz free energy derived from the partition function as follows:

$$F(\beta) = -\beta^{-1} \log Z$$

where $\beta = 1/(kT)$ and k is Boltzmann's constant and T is the temperature.

This quantity is related to the cumulant generating function for the energy. In the context of the maximum entropy principle, the physical temperature and the Boltzmann's constant play no role, so usually both are considered equal to 1. From the free energy, all of the thermodynamic properties of the system can be obtained via its derivatives, examples are the internal energy, specific heat, and entropy. It is to be noted that the definition used in this tutorial for the free energy (11) follows from the conventions used in the field of thermodynamic formalism [41,45,46] and changes its sign with the usual convention in the field of statistical mechanics.

4.4. Example

We present here the toy example that we will use to explore statistical properties of spike trains using the non-equilibrium statistical physics approach. We present the transfer matrix technique to compute the Markov transition matrix, its invariant measure and free energy from a potential $\mathcal{H}$.

Consider a range-2 potential with two neurons ($N = 2$). We use the notation introduced in Section 2.1:

$$\mathcal{H}(\mathbf{x}^{0,1}) = h_1 x_0^1 x_1^2 + h_2 x_0^2 x_1^1.$$

The state space of this problem is given by:

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The transfer matrix (12) associated to $\mathcal{H}$ is, in this case, a 4×4 matrix

$$\mathcal{L}_{\mathbf{x}\mathbf{x}'} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & e^{h_2} & e^{h_2} \\ 1 & e^{h_1} & 1 & e^{h_1} \\ 1 & e^{h_1} & e^{h_2} & e^{h_1+h_2} \end{pmatrix}.$$

This matrix satisfies the hypothesis of the Perron-Frobenius theorem. The maximum eigenvalue is

$$\rho = \frac{1}{2}(3 + e^{(h_1+h_2)} + \sqrt{5 + 4e^{h_1} + 4e^{h_2} + 2e^{(h_1+h_2)} + e^{(2h_1+2h_2)}}),$$

and the free energy

$$\mathcal{F}[\mathcal{H}] = \log(\rho). \tag{15}$$

5. Statistical Properties of Markov Maximum Entropy Measures

The procedure of finding a maximum entropy model gives us a full statistical model of the system of interest. In this section we discuss the added value that having such a model can provide.

5.1. Cumulants from Free Energy

All the statistical properties of the observables and their correlations can be obtained by taking the successive derivatives of the free energy with respect to the Lagrange Multipliers. This property explains the important role played by the free energy in the framework of MEMC. In general,

$$\frac{\partial^n \mathcal{F}[\mathcal{H}]}{\partial h_k^n} = \kappa_n \quad \forall k \in \{1, ..., K\},$$

where κ_n is the cumulant of order n (Equation (5)). In particular, taking the first derivative:

$$\frac{\partial \mathcal{F}[\mathcal{H}]}{\partial h_k} = \mathbb{E}_p\{f_k\} \quad \forall k \in \{1, ..., K\}, \tag{16}$$

where $\mathbb{E}_p\{f_k\}$ is the average with respect to the maximum entropy distribution p, which is equal to the average value of f_k with respect to the empirical measure. With Equation (16) the parameters of the MEMC can be fitted to be consistent with fixed average values of observables.

Suppose that we compute from data the average values of the following observables $\langle x_0^1 x_1^2 \rangle = 0.1$ and $\langle x_0^2 x_1^1 \rangle = 0.3$. We solve (16) (two equations and two unknowns) and obtain $h_1 = -1.98306$ and $h_2 = 1.48406$. With these parameters, the following Markov transition matrix and invariant measure are obtained from (13):

$$P_{\mathbf{x}\,\mathbf{x}'} = \begin{pmatrix} 0.232971 & 0.469441 & 0.0987018 & 0.198886 \\ 0.115617 & 0.232971 & 0.216056 & 0.435357 \\ 0.549892 & 0.15252 & 0.232971 & 0.0646176 \\ 0.272896 & 0.0756914 & 0.509966 & 0.141446 \end{pmatrix} \quad \pi(\mathbf{x}) = \begin{pmatrix} 0.29102 \\ 0.248443 \\ 0.248443 \\ 0.212095 \end{pmatrix}.$$

5.2. Fluctuation-Dissipation Relations

For a first-order stationary Markov chain, since each $X_n, n \geq 1$ depends on its predecessor, this induces a non-zero time-correlation between X_n and X_{n+r}, even when the distance r is greater than 1. This correlation, and more generally, time correlations between observables can be directly derived from the free energy. This relationship is usually referred to as Fluctuation-dissipation, and is also related to the linear response function that is presented in Section 5.7.

Let P be an ergodic matrix and indexed by the states in some finite set E, and π be its unique stationary measure. In this general context, for two real-valued functions that depend on a fixed finite number of components, we define the n–step correlation as

$$C_{f,g}(n) = \mathbb{E}_\pi(f(X_0)g(X_n)) - \mathbb{E}_\pi(f(X_0))\mathbb{E}_\pi(g(X_0)) \,.$$

In the particular case of MEMC with potentials of range $R > 1$ there is a positive time correlation between pairs of observables $f(x_n)$ and $g(x_{n+r})$. Suppose the correlations decay fast enough so that (at least)

$$\sum_{n=0}^{\infty} |C_{f,g}(n)| < \infty \,.$$

Then the following sum (known as the Green-Kubo formula [49]) converges and is non-negative:

$$\sigma^2_{f_k,f_j} = C_{f_k,f_j}(0) + \sum_{r=1}^{\infty} C_{f_k,f_j}(r) + \sum_{r=1}^{\infty} C_{f_j,f_k}(r). \tag{17}$$

Additionally, it can be shown that the energy function and the free energy depends smoothly upon maximum entropy parameters. Moreover, the correlations between observables can be obtained from the free energy through:

$$\sigma^2_{f_k,f_j} = \frac{\partial^2 \mathcal{F}[\mathcal{H}]}{\partial h_k \, \partial h_j} = \frac{\partial \mu(f_j)}{\partial h_k}.$$

The relationship between a correlation and a derivative of the free energy is called the fluctuation-dissipation theorem [50]. For a MEMC characterized by $\mu(P, \pi)$, the fluctuation-dissipation relationships can be obtained explicitly:

$$\begin{aligned} \frac{\partial^2 \mathcal{F}[\mathcal{H}]}{\partial h_k \, \partial h_j} =& \mathbb{E}_\mu[f_k f_j] - \mathbb{E}_\mu[f_k]\mathbb{E}_\mu[f_j] + \sum_{r=1}^{\infty} \sum_{\mathbf{x},\mathbf{x'} \in \mathbb{S}} \left(f_k(\mathbf{x}) f_j(\mathbf{x'}) \pi_{\mathbf{x}} P^r_{\mathbf{x}\mathbf{x'}} - \mathbb{E}_\mu[f_k]\mathbb{E}_\mu[f_j] \right) \\ &+ \sum_{r=1}^{\infty} \sum_{\mathbf{x},\mathbf{x'} \in \mathbb{S}} \left(f_j(\mathbf{x}) f_k(\mathbf{x'}) \pi_{\mathbf{x}} P^r_{\mathbf{x}\mathbf{x'}} - \mathbb{E}_\mu[f_k]\mathbb{E}_\mu[f_j] \right) . \end{aligned} \tag{18}$$

For MEMC built from K observables, the correlations can be conveniently arranged in a $K \times K$ symmetric matrix denoted by χ (the symmetry refers to the Onsager reciprocity relations [51]).

$$\chi_{jk} = \frac{\partial^2 \mathcal{F}[\mathcal{H}]}{\partial h_k \, \partial h_j} = \frac{\partial \mu(f_j)}{\partial h_k} = \frac{\partial \mu(f_k)}{\partial h_j} = \chi_{kj}. \tag{19}$$

For the example Section 4.4, we obtain the matrix χ by taking the second derivatives of (15) and evaluate at the parameters found previously,

$$\chi_{kj} = \begin{pmatrix} 0.0971481 & 0.0606071 \\ 0.0606071 & 0.127964 \end{pmatrix} .$$

In Figure 2, we plot the right hand side of Equation (18) for the MEMC built from the example Section 4.4 consistent with constraints considered in the example of Section 5.1, for the auto-correlation of the observable $x_0^2 x_1^1$.

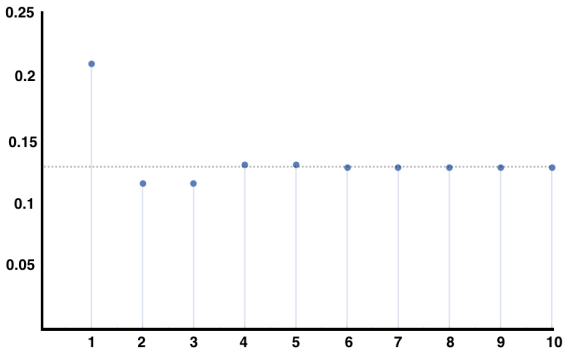

Figure 2. Plot of the auto-correlation of the observable $x_0^2 x_1^1$ with respect to the MEMC consistent with constraints $\langle x_0^1 x_1^2 \rangle = 0.1$ and $\langle x_0^2 x_1^1 \rangle = 0.3$. The plot show the sum of Equation (18) from $r = 1$ up to the number in the abscissa. Note the fast convergence towards χ_{22}.

5.3. Resonances and Decay of Correlations

We now turn back to the general setting of an arbitrary ergodic matrix P with stationary measure π associated to a Markov chain taking values on a finite state space (not necessarily the space of spike-patterns). Without loss of generality, assume that P is indexed by the states in $E = \{1, 2, \ldots, M\}$. It can be proved that in this case there exists $(l_i : i = 1, 2, \ldots, M)$ and $(r_i : i = 1, 2, \ldots, M)$, sets of left and right eigenvectors respectively, associated to the eigenvalues $(\rho_i : i = 1, \ldots, M)$. We can assume that the eigenvectors and left and right eigenvalues have been sorted and normalized in such a way that $\rho_1 = 1$, l_1 is the unique P–stationary probability vector π, $r_1 = (111 \ldots 1)^T$, and

$$\langle l_i | r_j \rangle = \delta_{i,j},$$

where $\delta_{i,j}$ is the Kronecker delta, and $\langle uv \rangle = \langle u, v \rangle$ corresponds to the Dirac's bra-ket, $|u\rangle\langle v| = uv^T$. With the same notation, the spectral decomposition of P is written:

$$P = \sum_{i=1}^{M} \rho_i |r_i\,]\,[\,l_i\,|\,.$$

Hence:

$$P^n = \sum_{i=1}^{M} \rho_i^n |r_i\,]\,[\,l_i\,|\,. \tag{20}$$

Given two functions $f : E \mapsto \mathbb{R}$ and $g : E \mapsto \mathbb{R}$ the following holds,

$$\begin{aligned} C_{f,g}(n) &:= \mathbb{E}_\pi(f(X_0)g(X_n)) - \mathbb{E}_\pi(f(X_0))\mathbb{E}_\pi(g(X_0)) \\ &= \langle \pi f \circ P^n g \rangle - \langle \pi f \rangle \langle \pi g \rangle. \end{aligned} \tag{21}$$

Recall the discussion in previous sections regarding the reverse chain Section 2.4. Writing $\mathbb{E}_\pi^{\leftarrow}$ for the expectation operator associated to the reverse Markov measure, i.e., to the measure $\mu = \mu(\pi, \overleftarrow{P})$, one can see that

$$\mathbb{E}_\pi(f(X_0)g(X_n)) = \mathbb{E}_\pi^{\leftarrow}(f(X_n)g(X_0))\,,$$

and hence (21) becomes

$$\langle \pi | g \circ \overleftarrow{P}^n f \rangle - \langle \pi | f \rangle \langle \pi | g \rangle.$$

From (20)

$$f \circ P^n g = \sum_{i=1}^{M} \langle l_i | g \rangle | f \circ r_i \rangle$$

and thus (21) becomes

$$\begin{aligned} C_{f,g}(n) &= \sum_{i=1}^{M} \rho_i^n \langle l_i | g \rangle \langle \pi | f \circ r_i \rangle - \langle \pi | f \rangle \langle \pi | g \rangle \\ &= \sum_{i=2}^{M} \rho_i^n \langle l_i | g \rangle \langle \pi | f \circ r_i \rangle. \end{aligned} \tag{22}$$

In Figure 3, we show the auto-correlations of the same observable considered in Figure 1, for the same MEMC. We observe modulations in the decay of the auto-correlations due to the complex eigenvalues in Equation (22), which arise in the non-symmetric transition matrix induced by the irreversibility of the MEMC.

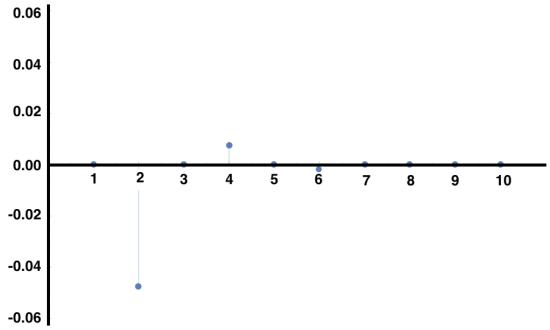

Figure 3. Auto-correlations of the observable $x_0^2 x_1^1$ for the MEMC with the same parameters as in Figure 1. Modulations in the decay of correlations are due to the complex eigenvalues of the MEMC.

We have found in Equation (22) an explicit expression for the decay of correlation for observables from the set of eigenvalues and eigenvectors of the transition matrix P. This is relevant in the context of spike train statistics because as the matrix P characterizing the spike trains is not expected to be symmetric, its eigenvalues are not necessarily real and modulations in the decay of correlations are expected (resonances). When measuring correlations between observables from data, one may observe this oscillatory situation that resembles resonances. This may be a symptom of a non-equilibrium situation.

5.4. Large Deviations for Average Values of Observables in MEMC

Obtaining the probability of "rare" average values of firing rates, pairwise correlations, triplets or non-synchronous observables is relevant in spike train statistics as these observables are likely to play an important role in neuronal information processing, and rare values may convey crucial information or be a symptom that the system in not working properly.

Here, we build from a previous article [52] where it is shown that the SCGF (7) can be obtained directly from the inferred Markov transition matrix P through the Gärtner-Ellis theorem (8). Consider a MEMC with transition matrix P. Let f be an observable of finite range and $k \in \mathbb{R}$. We introduce the *tilted transition matrix by* f of P, parametrized by k and denoted by $\widetilde{P}^{(f)}(k)$ [53] as follows:

$$\widetilde{P}_{ij}^{(f)}(k) = P_{ij}e^{kf(ij)} \quad i,j \in \mathbb{S}. \tag{23}$$

The tilted transition matrix can be directly obtained from the spectral properties of the transfer matrix (12),

$$\begin{aligned}\widetilde{P}_{ij}^{(f)}(k) &= \frac{e^{\mathcal{H}_{ij}}v_j}{v_i\,\rho}e^{kf(ij)} \\ &= \frac{e^{\left[\mathcal{H}_{ij}+kf(ij)\right]}v_j}{v_i\,\rho} \quad i,j \in \mathbb{S}.\end{aligned}$$

Recall that $\mathbf{v}$ is the right eigenvector associated to its maximum eigenvalue ρ of the transfer matrix $\mathcal{L}$. Here we use the notation $\mathcal{H}_{ij}$ to specify that the energy function is built from the elements of the state space i and j. Remarkably, this result is valid not only for the observables in the energy function, i.e., from here the LDP of more general observables can be computed.

To obtain an explicit expressions for the SCGF $\lambda_f(k)$, it is possible to take advantage of the structure of the underlying stochastic process. For instance, for i.i.d. random process X_t where $X_i \sim X$ from Definition 7, one can obtain that

$$\lambda(k) = \lim_{t\to\infty}\frac{1}{t}\ln\mathbb{E}[e^{tkA_t(f)}]^t = \ln\mathbb{E}[e^{kf(X)}],$$

which is the case of range one observables. Using the Equation (23), we obtain that the maximum eigenvalue of the tilted matrix $\rho(\widetilde{P}_f(k))$ is,

$$\rho\left(\widetilde{P}_f(k)\right) = \sum_j \pi_j e^{kf(j)} \quad j \in \mathbb{S}.$$

As $\widetilde{P}_f$ is a primitive matrix, the uniqueness of $\rho\left(\widetilde{P}_f(k)\right)$ is ensured from the Perron-Frobenius theorem. For additive observables of ergodic Markov chains, a direct calculation (see [54]) leads us to

$$\lambda_f(k) = \ln\left(\rho\left(\widetilde{P}^{(f)}\right)\right).$$

It can also be proved that $\lambda_f(k)$, in this case, is differentiable [54], setting up the scene to use the Gärtner-Ellis theorem to obtain $I_f(s)$ as shown in Figure 4.

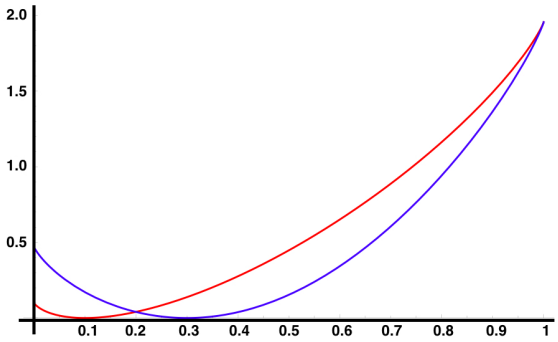

Figure 4. Rate functions of observables $x_0^1x_1^2$ in red, and $x_0^2x_1^1$ in blue for the MEMC consistent with constraints $\langle x_0^1x_1^2\rangle = 0.1$ and $\langle x_0^2x_1^1\rangle = 0.3$. The minimum value of both functions coincide with their expected values with respect to the MEMC. Around the minimum Gaussian fluctuations are expected (9). Far from the expected values are the large deviations.

5.5. Information Entropy Production

Given a Markov chain $(X_t : t \geq 0)$ on a general finite state space E with transition matrix P started from the distribution ν, denoted $\nu^{(n)}$ the distribution of X_n, namely, for $i \in E$:

$$\nu^{(n)}(i) = \mathbb{P}_\nu(X_n = i)\,.$$

Obviously, $\nu^{(0)} = \nu$, and

$$\nu_j^{(n+1)} = \sum_{i \in E} \nu_i^{(n)} P_{ij}\,.$$

The information-theoretic entropy of the probability distribution ν at time n is given by

$$\mathcal{S}_n(\nu) := -\sum_{i \in E} \nu_i^{(n)} \log \nu_i^{(n)},$$

and the *change of entropy* over one time step is defined as

$$\Delta\mathcal{S}_n := \mathcal{S}_{n+1}(\nu) - \mathcal{S}_n(\nu).$$

A bit of algebra yields

$$\Delta\mathcal{S}_n = -\sum_{i,j \in E} \nu_j^{(n)} P_{ji} \log \frac{\nu_j^{(n+1)} P_{ji}}{\nu_i^{(n)} P_{ij}} + \frac{1}{2} \sum_{i,j \in E} \left[\nu_j^{(n)} P_{ji} - \nu_i^{(n)} P_{ij}\right] \log \frac{\nu_j^{(n)} P_{ji}}{\nu_i^{(n)} P_{ij}}.$$

The first term on the right hand side above is called *information entropy flow* and the second term *information entropy production* [12].

In the stationary case, i.e., when P admits a stationary measure π and the chain is started from that distribution, one has that $\nu^{(n)} = \pi$ for every $n \geq 0$; thus, in this case, the change of entropy rate is zero, i.e., for stationary chains, the information entropy flow equals (minus) information entropy production. This case is the focus of this work. The chain is associated to spike train activity for transitions between L–blocks. Starting from stationarity the entropy production rate is explicitly given by

$$IEP(P, \pi) := \frac{1}{2} \sum_{\mathbf{x}, \mathbf{x}' \in \mathbb{S}^L} \left[\pi(\mathbf{x}') P_{\mathbf{x}'\mathbf{x}} - \pi(\mathbf{x}) P_{\mathbf{x}\mathbf{x}'}\right] \log \frac{\pi(\mathbf{x}') P_{\mathbf{x}'\mathbf{x}}}{\pi(\mathbf{x}) P_{\mathbf{x}\mathbf{x}'}} \geq 0. \tag{24}$$

The non-negativity implies that information entropy is positive as long as the process violates the detailed balance conditions (4). This is analogous to the second law of thermodynamics [55]. From this equation it is easy to realize that if the Markov chain satisfies the detailed balance condition, the information entropy production is zero.

In Figure 5, we compute the information entropy production from Equation (24), for the MEMC of the example (Section 4.4) for different values of the parameters h_1, h_2.

It may seem contradictory that in stationary state the entropy is constant, while there is a positive "production" of entropy. The information entropy production in stationary state *always* compensate the information entropy flow, which leaves the information entropy rate constant. In this case we refer to non-equilibrium steady states (NESS).

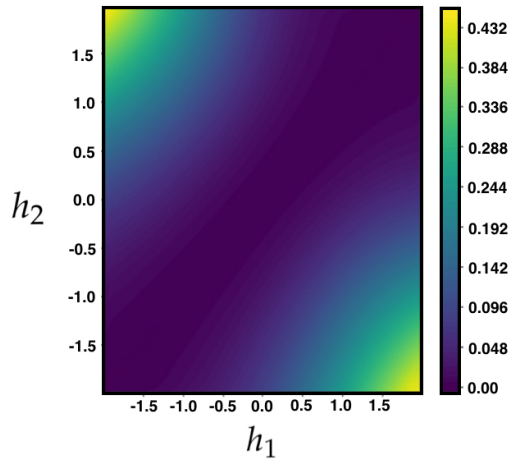

Figure 5. IEP for the MEMC of the example (Section 4.4) for different values of parameters h_1, h_2. Observe that $IEP(P, \pi) = 0$ when $h_1 = h_2$ and that increases as the parameters become more different (more asymmetry in P).

5.6. Gallavotti-Cohen Fluctuation Theorem

To characterize the fluctuations of the IEP, consider the MEMC $\mu(P, \pi)$ and the following observable:

$$W_n(\mathbf{x}_{0,n}) = \frac{1}{n} \ln \left(\frac{\mu(\mathbf{x}_{0,n})}{\mu(\mathbf{x}_{n,0})} \right),$$

It can be shown that $\lim \frac{1}{n} W_n \to IEP(\pi, P)$. The Gallavotti-Cohen fluctuation theorem is as a statement about properties of the SCGF and rate function of the IEP [14].

$$\lambda_W(k) = \lambda_W(-k-1), \quad I_W(s) = I_W(-s) - s. \tag{25}$$

This symmetry holds for a general class of stochastic processes including NESS from Markov chains [56], and is a *universal property* of the IEP, i.e., it is independent of the parameters of the MEMC. To compute $\lambda_W(k)$ and $I_W(s)$, define $A(k)_{ij} = P_{ij} \left[\frac{\pi_i P_{ij}}{\pi_j P_{ji}} \right]^k$. If $\rho(k)$ is the largest eigenvalue of $A(k)$, then $\lim_{n\to\infty} \ln \mathbb{E}(e^{n\lambda W_n}) = \ln \rho(k)$.

In Figure 6, we illustrate the Gallavotti-Cohen symmetry property of the large deviation functions associated to the IEP (Equation (25)).

These properties are relevant to the large deviations of the averaged entropy production denoted $\frac{W_t}{t}$ over a trajectory $\boldsymbol{x}_{0,t-1}$ of the Markov chain $p(\pi, P)$. The following relationship holds,

$$\frac{p\left\{\frac{W_t}{t} \approx s\right\}}{p\left\{\frac{W_t}{t} \approx -s\right\}} \asymp e^{ts}.$$

This means that the positive fluctuations of $\frac{W_t}{t}$ are exponentially more probable than negative fluctuations of equal magnitude.

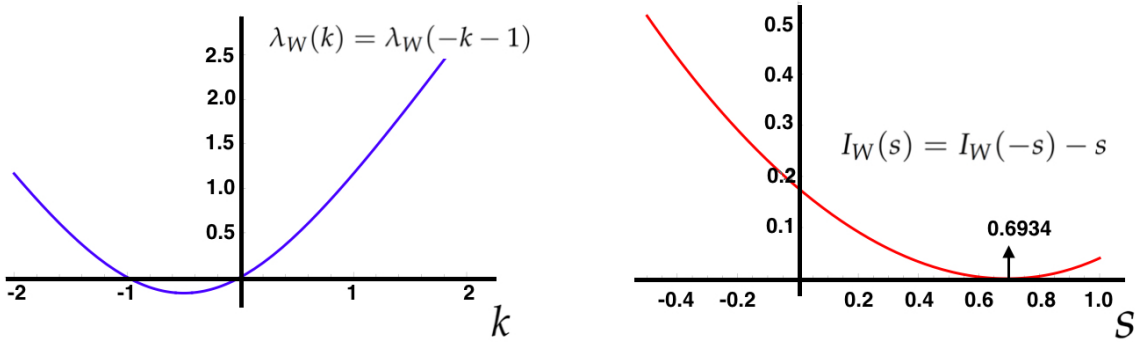

Figure 6. Gallavotti-Cohen symmetry property for the SCGF and rate function of the IEP (Equation (25)). Left: SCGF of the IEP of the MEMC with the same parameters considered in the previous examples. Right: Rate function of the observable W, the minimum is attained at the expected value of IEP.

5.7. Linear Response

The linear response serves to quantify how a small perturbation $\delta \boldsymbol{h}$ of a set of the maximum entropy parameters affects the average values of observables in terms of the unperturbed measure. This is relevant in the context of spike trains statistics to identify stiff and sloppy directions in the space of parameters. A small change in a sloppy parameter produces very little impact in the statistical model. In contrast, a small change in a stiff parameter produces a significant change. For a MEMC characterized by $\mu = (P, \pi)$ corresponding to an energy function with fixed parameters $\mathbf{h}$ denoted by $\mathcal{H}_{\mathbf{h}}$, one can obtain the average value of a given observable f_k from (16).

Now, consider a perturbed energy denoted by $\tilde{\mathcal{H}} = \mathcal{H}_{\boldsymbol{h}+\delta\boldsymbol{h}}$. Using a Taylor expansion, the average value of an arbitrary observable f_k with respect to the MEMC can be obtained $\tilde{\mu} = (\tilde{P}, \tilde{\pi})$ associated to the perturbed energy. Considering the Taylor expansion of $\mathcal{F}[\mathcal{H}_{\boldsymbol{h}+\delta\boldsymbol{h}}]$ about $\mathcal{H}_{\boldsymbol{h}}$

$$\frac{\partial \mathcal{F}[\mathcal{H}_{\boldsymbol{h}+\delta\boldsymbol{h}}]}{\partial h_k} = \frac{\partial \mathcal{F}[\mathcal{H}_{\boldsymbol{h}}]}{h_k} + \sum_j \frac{\partial^2 \mathcal{F}[\mathcal{H}_{\boldsymbol{h}}]}{\partial h_k h_j} \delta h_j + O(\delta h_j)^2 \,, \tag{26}$$

$$\mathbb{E}_{\tilde{\mu}}[f_k] = \mathbb{E}_{\mu}[f_k] + \sum_j \frac{\partial^2 \mathcal{F}[\mathcal{H}_{\boldsymbol{h}}]}{\partial h_k h_j} \delta h_j + O(\delta h_j)^2 \,, \tag{27}$$

$$\Delta \mathbb{E}[f] \approx \chi \cdot \delta \mathbf{h} \,. \tag{28}$$

We use (16) to go from (26) to (27). Observe from (27) that a small perturbation of a parameter h_j influences the average value of all other observables in the energy function (as f_k is arbitrary). The perturbation is modulated by the second derivatives of the free energy corresponding to the unperturbed regime $\mathcal{F}[\mathcal{H}_{\boldsymbol{h}}]$ (see Figure 7).

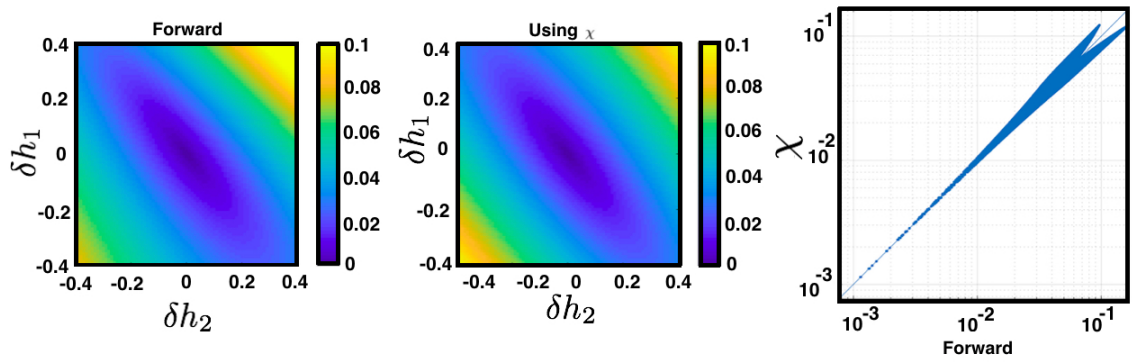

Figure 7. Linear response for the MEMC of the example (Section 4.4) for different values of perturbations δh_1 and δh_2. The colors represent $\|\mathbb{E}_{\tilde{\mu}}[f_k] - \mathbb{E}_{\mu}[f_k]\|$ computed using two methods. The "forward" method consists in computing $\mathbb{E}_{\tilde{\mu}}[f_k]$ from $\tilde{\mu}$ and $\mathbb{E}_{\mu}[f_k]$ from μ. The figure in the middle is obtained by computing $\|\mathbb{E}_{\tilde{\mu}}[f_k] - \mathbb{E}_{\mu}[f_k]\|$ from χ using Equation (28). (**Right**) The difference between both methods illustrated in a scatter plot in logarithmic scale.

6. Discussion and Future Work

This tutorial explores how one can use maximum entropy methods to capture asymmetric temporal aspects of spike trains from experimental data. In particular, we showed how spatio-temporal constraints can produce homogeneous irreducible Markov chains whose unique steady state is, in general, non-equilibrium (NESS)—thus, detailed balance condition is not satisfied causing strictly positive entropy production. This fact highlights that only non-synchronous maximum entropy models induce time irreversible processes, which is one of the key hallmarks of biological systems.

We have presented a survey of diverse techniques from mathematics and statistical mechanics to study these NESS, which correspond to a rich toolkit that can be employed to study unexplored aspects of spike train statistics. We emphasise that many of these concepts, including entropy production and fluctuation-dissipation relationships, have not been explored much in the context of spike train analysis. However, the fact that time irreversibility is such an important feature of living systems suggest that these notions might play an important role in neural dynamics.

Possible extensions include measuring the entropy production for different choices of spatio-temporal constraints using the maximum entropy method on biological spike train recordings. A more ambitious extension is to explore the relationship between entropy production computed from experimental data obtained from different physiological processes and relate them to features such as adaptation or learning. Concerning time-dependent neuronal network models, future studies might lead to a better understanding of the impact of particular synaptic topologies of neuronal network models on the corresponding entropy production, decay of correlations, resonances and other sophisticated statistical properties.

Other possible extensions are related to the drawbacks of current approaches. This can include limitations of the maximum entropy method related to the requirement of stationarity in the data, which is not a natural condition for some biological scenarios. However, several of the techniques presented in this tutorial naturally extend to the non-stationary case, including the information entropy production, which can still be defined along non-stationary trajectories [14]. Related to this issue is that the approach presented in this tutorial does not make any reference to the stimulus. While this issue has been addressed in the synchronous framework [34], there is still an open field to explore the Markovian extension of these ideas. Another interesting topic to explore in future studies is the inclusion of the non-stationary approach such as the state space analysis proposed in [35]. Also, another open problem is related to the efficient implementation of the transfer matrix technique, which currently requires an important computational effort in the case of large neural networks. Recently, some improvements of this approach have been proposed based in Monte Carlo methods [57].

In summary, we believe that these topics are fertile ground for multi-disciplinary exploration by teams composed of mathematicians, physicists, and neuroscientists. It is our hope that this work may foster future collaborative research among disciplines, which might bring new breakthroughs to advance our fundamental understanding of how the brain works.

Author Contributions: The three authors conceived the main ideas and concepts, wrote and revised the manuscript. All authors have read and approved the final manuscript.

Funding: L.V. was supported by CONICYT-Beca de Doctorado No. 21170406 Convocatoria 2017. F.R. was supported by the Ad Astra Chandaria Foundation. R.C. was supported by CONICYT-PAI Inserción 79160120, Proyecto REDES ETAPA INICIAL, Convocatoria 2017 REDI170457, Fondecyt Iniciación 2018 Proyecto 11181072.

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

Abbreviations

The following abbreviations are used in this manuscript:

MEP	Maximum entropy principle
MEMC	Maximum entropy Markov chain
SCGF	Scaled cumulant generating function
CLT	Central limit theorem
LLN	Law of large numbers
LDP	Large deviation principle
IEP	Information entropy production
KSE	Kolmogorov-Sinai entropy
NESS	Non-equilibrium steady states

Symbol list

$\mathbb{S}$	$\{0,1\}^N$ the state space of spike patterns of N neuron
Ω	The set of infinite sequences of spike patterns
x_n^k	Spiking state of neuron k at time n
$\boldsymbol{x}_n$	Spike pattern at time n
$\boldsymbol{x}_{t_1,t_2}$	Spike block from time t_1 to t_2
$\nu(f)$	Expectation of the observable f w.r.t. the probability measure ν
$A_T(f)$	Empirical Average value of the observable f considering T spike patterns
$\mathbb{S}^R$	Space of spike blocks of N neurons and length R
$\mathcal{S}[\mu]$	Entropy of the probability measure μ
$\mathcal{H}$	Energy function
$\mathcal{F}[\mathcal{H}]$	Free energy

References

1. Rieke, F.; Warland, D.; de Ruyter van Steveninck, R.; Bialek, W. *Spikes, Exploring the Neural Code*; M.I.T. Press: Cambridge, MA, USA, 1996.
2. Bialek, W. *Biophysics: Searching for Principles*; Princeton University Press: Princeton, NJ, USA, 2012.
3. Schneidman, E.; Berry, M.J.; Segev, R.; Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **2006**, *440*, 1007–1012. [CrossRef] [PubMed]
4. Ganmor, E.; Segev, R.; Schneidman, E. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proc Natl. Acad. Sci. USA* **2011**, *108*, 9679–9684. [CrossRef] [PubMed]
5. Tkačik, G.; Marre, O.; Amodei, D.; Schneidman, E.; Bialek, W.; Berry, M.J. Searching for collective behavior in a large network of sensory neurons. *PLoS Comput. Biol.* **2014**, *10*, e1003408. [CrossRef] [PubMed]
6. Palsso, B. *Systems Biology: Properties of Reconstructed Networks*; Cambridge University Press: Cambridge, UK, 2006.
7. Tang, A.; Jackson, D.; Hobbs, J.; Chen, W.; Smith, J.; Patel, H.; Prieto, A.; Petrusca, D.; Grivich, M.; Sher, A.; et al. A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J. Neurosci.* **2008**, *28*, 505–518. [CrossRef] [PubMed]

8. Marre, O.; El Boustani, S.; Frégnac, Y.; Destexhe, A. Prediction of spatiotemporal patterns of neural activity from pairwise correlations. *Phys. Rev. Lett.* **2009**, *102*, 138101. [CrossRef]
9. Vasquez, J.; Palacios, A.; Marre, O.; Berry, M., II; Cessac, B. Gibbs distribution analysis of temporal correlation structure on multicell spike trains from retina ganglion cells. *J. Physiol. Paris* **2012**, *106*, 120–127. [CrossRef] [PubMed]
10. Mora, T.; Deny, S.; Marre, O. Dynamical criticality in the collective activity of a population of retinal neurons. *Phys. Rev. Lett.* **2015**, *114*, 078105. [CrossRef]
11. Cofré, R.; Cessac, B. Exact computation of the maximum entropy potential of spiking neural networks models. *Phys. Rev. E* **2014**, *107*, 368–368. [CrossRef]
12. Cofré, R.; Maldonado, C. Information entropy production of maximum entropy Markov chains from spike trains. *Entropy* **2018**, *20*, 34. [CrossRef]
13. Schulman, L.S. *Time's Arrows and Quantum Measurement*; Cambridge University Press: Cambridge, UK, 1997.
14. Jiang, D.Q.; Qian, M.; Qian, M.P. *Mathematical Theory of Non-Equilibrium Steady States*; Springer: Berlin/Heidelberg, Germany, 2004.
15. Schrödinger, E. *What Is Life? The Physical Aspect of the Living Cell*; Cambridge University Press: Cambridge, UK, 1944.
16. Prigogine, I. *Nonequilibrium Statistical Mechanics*; Monographs in Statistical Physics; Interscience publishers, John Wiley & Sons: Hoboken, NJ, USA, 1962.
17. Deem, M. Mathematical adventures in biology. *Phys. Today* **2007**, *60*, 42–47. [CrossRef]
18. Filyukov, A.; Karpov, V. Description of steady transport processes by the method of the most probable path of evolution. *Inzhenerno-Fizicheskii Zhurnal* **1967**, *13*, 624–630. [CrossRef]
19. Filyukov, A.; Karpov, V. Method of the most probable path of evolution in the theory of stationary irreversible processes. *Inzhenerno-Fizicheskii Zhurnal* **1967**, *13*, 798–804. [CrossRef]
20. Favretti, M. The maximum entropy rate description of a thermodynamic system in a stationary non-equilibrium state. *Entropy* **2009**, *4*, 675–687. [CrossRef]
21. Monthus, C. Non-equilibrium steady states: maximization of the Shannon entropy associated with the distribution of dynamical trajectories in the presence of constraints. *J. Stat. Mech. Theor. Exp.* **2011**, *3*, P03008. [CrossRef]
22. Shi, P.; Qian, H. *Frontiers in Computational and Systems Biology*; Feng, J., Fu, W.; Sun, F., Eds.; Springer: London, UK, 2010; chapter Irreversible Stochastic Processes, Coupled Diffusions and Systems Biochemistry; pp. 175–201.
23. Galves, A.; Löcherbach, E. Infinite systems of interacting chains with memory of variable length-A stochastic model for biological neural nets. *J. Stat. Phys.* **2013**, *151*, 896–921. [CrossRef]
24. Cofré, R.; Cessac, B. Dynamics and spike trains statistics in conductance-based Integrate-and-Fire neural networks with chemical and electric synapses. *Chaos Solitons Fractals* **2013**, *50*, 13–31. [CrossRef]
25. Halmos, P.R. *Measure Theory*; Graduate Texts in Mathematics; Springer: New York, NY, USA, 1974.
26. Levin, D.; Peres, Y. *Markov Chains and Mixing Times*, 2nd ed.; American Mathematical Society: Providence, RI, USA, 2017.
27. Gerstner, W.; Kistler, W. *Spiking Neuron Models*; Cambridge University Press: Cambridge, UK, 2002.
28. Jones, G.L. On the Markov chain central limit theorem. *Probab. Surv.* **2004**, *1*, 299–320. [CrossRef]
29. Ellis, R. *Entropy, Large Deviations and Statistical Mechanics*; Springer: Berlin, Germany, 1985.
30. Touchette, H. The large deviation approach to statistical mechanics. *Phys. Rep.* **2009**, *478*, 1–69. [CrossRef]
31. Dembo, A.; Zeitouni, O. Large deviations techniques and applications. In *Stochastic Modelling and Applied Probability*; Springer: Berlin, Germany, 2010; Volume 38. [CrossRef]
32. Marre, O.; Amodei, D.; Deshmukh, N.; Sadeghi, K.; Soo, F.; Holy, T.; Berry, M., II. Mapping a complete neural population in the Retina. *J. Neurosci.* **2012**, *43*, 14859–14873. [CrossRef]
33. Tkačik, G.; Mora, T.; Marre, O.; Amodei, D.; Berry, M., II; Bialek, W. Thermodynamics for a network of neurons: Signatures of criticality. *Proc Natl. Acad. Sci. USA* **2015**, *112*, 11508–11513. [CrossRef]
34. Granot-Atedgi, E.; Tkačik, G.; Segev, R.; Schneidman, E. Stimulus-dependent maximum entropy models of neural population codes. *PLoS Comput. Biol.* **2013**, *9*, e1002922. [CrossRef]
35. Shimazaki, H.; Amari, S.; Brown, E.N.; Grün, S. State-space analysis of time-varying higher-order spike correlation for multiple neural spike train data. *PLoS Comput. Biol.* **2012**, *8*, e1002385. [CrossRef]

36. Dewar, R. Information theory explanation of the fluctuation theorem, maximum entropy production and self-organized criticality in non-equilibrium stationary states. *J. Phys. A Math. Gen.* **2003**, *36*, 631. [CrossRef]
37. Dewar, R. Maximum entropy production and the fluctuation theorem. *J. Phys. A Math. Gen.* **2005**, *38*, L371. [CrossRef]
38. Martyushev, L.; Seleznev, V. Maximum entropy production principle in physics, chemistry and biology. *Phys. Rep.* **2006**, *426*, 1–45. [CrossRef]
39. Jaynes, E. The minimum entropy production principle. *Ann. Rev. Phys. Chem.* **1980**, *31*, 579–601. [CrossRef]
40. Pressé, S.; Ghosh, K.; Lee, J.; Dill, K.A. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.* **2013**, *85*, 1115. [CrossRef]
41. Ruelle, D. *Thermodynamic Formalism*; Addison-Wesley: Reading, MA, USA, 1978.
42. Jaynes, E. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620. [CrossRef]
43. Tkačik, G.; Marre, O.; Mora, T.; Amodei, D.; Berry, M., II; Bialek, W. The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech.* **2013**, *2013*, P03011. [CrossRef]
44. Seneta, E. *Non-Negative Matrices and Markov Chains*; Springer: New York, NY, USA, 2006.
45. Walters, P. Ruelle's operator theorem and g-measures. *Trans. Am. Math. Soc.* **1975**, *214*, 375–387.
46. Bowen, R. Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms. In *Lecture Notes in Mathematics*, revised ed.; Springer: Berlin, Germany, 2008; Volume 470.
47. Parry, W.; Pollicott, M. Zeta functions and the periodic orbit structure of hyperbolic dynamics. *Astérisque, Société mathématique de France* **1990**, 187–188. Available online: http://www.numdam.org/issue/AST_1990__187-188__1_0.pdf (accessed on 11 September 2019).
48. Chazottes, J. Fluctuations of observables in dynamical systems: From limit theorems to concentration inequalities. In *Nonlinear Dynamics New Directions*; González-Aguilar, H., Ugalde, E., Eds.; Springer: Cham, Switzerland, 2015; pp. 47–85.
49. Gaspard, P. *Chaos, Scattering and Statistical Mechanics*; Non-Linear Science series; Cambridge University Press: Cambridge, UK, 1998; Volume 9.
50. Bettolo, U.M.; Puglisi, A.; Rondoni, L.; Vulpiani, A. Fluctuation–dissipation: Response theory in statistical physics. *Phys. Rep.* **2008**, *461*, 111–195.
51. Gaspard, P. Random paths and current fluctuations in nonequilibrium statistical mechanics. *J. Math. Phys.* **2014**, *55*, 075208. [CrossRef]
52. Cofré, R.; Maldonado, C.; Rosas, F. Large deviations properties of maximum entropy Markov shains from spike trains. *Entropy* **2018**, *20*, 573. [CrossRef]
53. Touchette, H. A Basic Introduction to Large Deviations: Theory, Applications, Simulations. *arXiv* **2012**, arxiv:1106.4146v3.
54. Ellis, R.S. The theory of large deviations and applications to statistical mechanics. In *Long-Range Interacting Systems*; Oxford University Press: Oxford, UK, 2010.
55. Nicolis, G.; Nicolis, C. *Foundations of Complex Systems: Emergence, Information and Prediction*; World Scientific: Singapore, 2012.
56. Maes, C. The fluctuation theorem as a Gibbs property. *J. Stat. Phys.* **1999**, *95*, 367–392. [CrossRef]
57. Nasser, H.; Cessac, B. Parameter estimation for spatio-temporal maximum entropy distributions: Application to neural spike trains. *Entropy* **2014**, *16*, 2244–2277. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).