entropy

MDPI

Article

Large Deviations Properties of Maximum Entropy Markov Chains from Spike Trains

Rodrigo Cofré [1,*], Cesar Maldonado [2] and Fernando Rosas [3,4]

[1] Centro de Investigación y Modelamiento de Fenómenos Aleatorios, Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso 2340000, Chile
[2] IPICYT/División de Matemáticas Aplicadas, Instituto Potosino de Investigación Científica y Tecnológica, San Luis Potosí 78216, Mexico; cesar.maldonado@ipicyt.edu.mx
[3] Centre of Complexity Science and Department of Mathematics, Imperial College London, London SW7 2AZ, UK; f.rosas@imperial.ac.uk
[4] Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK
* Correspondence: rodrigo.cofre@uv.cl; Tel.: +56-32-2603625

Received: 6 June 2018; Accepted: 11 July 2018; Published: 3 August 2018

Abstract: We consider the maximum entropy Markov chain inference approach to characterize the collective statistics of neuronal spike trains, focusing on the statistical properties of the inferred model. To find the maximum entropy Markov chain, we use the thermodynamic formalism, which provides insightful connections with statistical physics and thermodynamics from which large deviations properties arise naturally. We provide an accessible introduction to the maximum entropy Markov chain inference problem and large deviations theory to the community of computational neuroscience, avoiding some technicalities while preserving the core ideas and intuitions. We review large deviations techniques useful in spike train statistics to describe properties of accuracy and convergence in terms of sampling size. We use these results to study the statistical fluctuation of correlations, distinguishability, and irreversibility of maximum entropy Markov chains. We illustrate these applications using simple examples where the large deviation rate function is explicitly obtained for maximum entropy models of relevance in this field.

Keywords: computational neuroscience; spike train statistics; maximum entropy principle; large deviation theory; out-of-equilibrium statistical mechanics; thermodynamic formalism; entropy production

1. Introduction

Spiking neuronal networks are perhaps the most sophisticated information processing devices that are available for scientific inquiry. There exists already an understanding of their basic mechanisms and functionality: they are composed by interconnected neurons which fire action potentials (also known as "spikes") collectively in order to accomplish specific tasks, e.g., sensory information processing or motor control [1]. However, the interdependencies in the spiking activity of populations of neurons can be extremely complex. In effect, these interdependencies can involve neighboring or also distant cells, being established either via structural connections, i.e., physical mediums such as synapses, or by functional connections reflected through spike correlations [2].

Understanding the way in which neuronal networks process information requires disentangling structural and functional connections while clarifying their interplay, which is a challenging but critical issue [3,4]. For this aim, networks of spiking neurons are usually measured using in vitro or in vivo multi-electrode-arrays, which connect neurons to electronic sensors specially designed for spike detection. Recent progress in acquisition techniques allows the simultaneous measurement of

the spiking activity from increasingly large populations of neurons, enabling the collection of large amounts of experimental data [5]. Prominent examples of spike train recordings have been obtained from vertebrate retina (salamander, rabbit, and degus) [6–8] and cat cortex [9].

However, despite the progress in multi-electrode and neuroimaging recording techniques, modeling the collective spike train statistics is still one of the key open challenges in computational neuroscience. Analysis over recorded data has shown that, although the neuronal activity is highly variable (even when presented repeatedly the same stimulus), the statistics of the response is highly structured [10,11]. Therefore, it seems that much of the inner dynamics of neuronal networks is encoded in the statistical structure of the spikes. Unfortunately, traditional methods of estimation, inference, and model selection are not well-suited for this scenario since the number of possible binary patterns that a neuronal network can adopt grows exponentially with the size of the population. In fact, even long experimental recordings usually contain only a small subset of the possible spiking patterns, which makes the empirical frequencies poor estimators for the underlying probability distribution. For practical purposes, this induces dramatic limitations, as standard inference tools become unreliable as soon as the number of considered neurons grows beyond 10 [6].

Given the binary nature of the spiking data, it is natural to relate neuronal networks and digital communication system via Shannon's information theory. A possibly more subtle way of establishing this link is provided by the physics literature that studies stochastic spins systems. In fact, a succession of research efforts has helped develop a framework to study the spike train statistics based on tools of statistical physics, namely the maximum entropy principle (MEP), which provides an intuitive and tractable procedure to build a statistical model for the whole neuronal network. In 2006, Schneidman et al. [6] and Pillow et al. [12] used the MEP to characterize the spike train statistics of the vertebrate retina responding to natural stimuli, constraining only range one features, namely firing rates and instantaneous pairwise interactions. Since then, the MEP approach has become a standard tool to build probability measures in the field of spike train statistics [6,8,12,13]. This approach has triggered fruitful analyses of the neural code, including works about criticality [14], redundancy and error correction [7] among other intriguing and promising topics.

Although relatively successful, this approach for linking neuronal populations and statistical mechanics is based on assumptions that go against fundamental biological knowledge. Firstly, these works assume that the spike patterns are statistically independent of past and future activities of the network. In fact, and not surprisingly, there exists strong evidence supporting the fact that memory effects play a major role in spike train statistics [8,9,15]. Secondly, most studies that apply statistical mechanics to analyze neuronal data use tools that assume that the underlying system is in thermodynamic equilibrium. However, it has been recognized that being out-of-equilibrium is one of the distinctive properties of living systems [16–18]. Consequently, any statistical description that is consistent with the out-of-equilibrium condition of living neuronal networks should reflect some degree of time asymmetry (i.e., time irreversibility), which can be characterized using Markov chains [19–24].

As a way of addressing the above observations, some recent publications study maximum entropy Markov chains (MEMC) based on a variational principle from the thermodynamic formalism of dynamical systems (see for instance [8,24,25]). This framework is an extension of the classic approach based on the MEP that considers correlation of spikes among neurons simultaneously and with different time delays as constraints, being able in this way to account for various memory effects.

Most of the literature in spike train statistics via the MEP pays little attention to the fact that model estimation is done based on finite data (errors due to statistical fluctuations are likely to occur in this context). As the MEP can be seen as a statistical inference procedure, it is natural to inquire about the uncertainty (i.e., fluctuations and convergence properties) related to the inferred MEMC, or, in other words, ask for the robustness of the inference as a function of the sampling size of the underlying data set. Quantifying this error is particularly relevant in the light of recent results that suggest that the parameters inferred by the MEP approach in the context of experimental biological recordings are

sharply poised at criticality [7,26]. On the other hand, once the MEMC has been inferred, it is also important to quantify how well a sample of the MEMC reproduce the average values of features of interest and how likely is that a sample of the MEMC produce a "rare" or unlikely event.

To provide some first steps in addressing the above issues, this paper studies the MEMC framework using tools from large deviation theory (LDT) [27,28]. We exploit the fact that the average values of features obtained from samples of the MEMC satisfy a large deviation property, and use LDT techniques to estimate their fluctuations in terms of the sampling size. We also show how to compute the rate functions using the tilted transition matrix technique and the Gärtner–Ellis theorem. It is to be noted that there is a large body of theoretical work linking the maximum entropy principle and large deviations [27,29]. However, these techniques have been scarcely used in spike train analysis (only to study the i.i.d case [30–33]), most likely because of the lack of a suitable introduction of these concepts within the neuroscientific literature. It is our hope that these applications might trigger the interest of the computational neuroscience community into the large deviation literature. Consequently, another goal of this paper is to provide an accessible introduction of the MEMC and LDT formalisms to the community of computational neuroscience, avoiding some technicalities while preserving the core ideas and intuitions. To the best of our knowledge, this manuscript presents the first attempt to bring these two topics together in the context of spike train statistics. This article is part of a more ambitious program that attempts to build a more unified theoretical structure and a complete toolbox helpful to approach spike train statistics using the thermodynamic formalism [24,25].

The rest of this paper is organized as follows. Section 2 presents the basic definitions and tools needed to apply large deviations techniques further in the paper. In particular, we present the maximum entropy principle framed in the thermodynamic formalism as a variational principle. In Section 3, we introduce some basic aspects of the theory of large deviations. In Section 4, we focus on the empirical averages of features. We present some examples of relevance in spike train statistics, where we are able to compute explicitly the rate function for each feature in the maximum entropy potential. In Section 5, we present further applications of the theory of large deviations in this field with a list of illustrative examples and finally we present our conclusions in Section 6.

2. Preliminaries

This section introduces the general definitions, notations and conventions that are used throughout the paper, providing in turn the necessary background for the unfamiliar reader.

2.1. Data Binarization and Spike Trains

Let us consider a set of measurements from a network of N interacting neurons. The "raw data" consist of N continuous signals containing the extra-cellular potential (electrical potential measured outside of the cell) of each of the neurons recorded over the length of the experiment. These data are processed by spike sorting algorithms [34,35], which are signal processing techniques designed to extract the spiking activity of each neuron.

Neurons have a minimal characteristic time in which no two spikes can occur, called "refractory period" [36], which provides a natural time-scale that can be used for "binning" (i.e., for discretizing) the time index of the measurements, denoted by Δt_b (When binning, sometimes it can be useful to go beyond the refractory period. In those cases, two spikes may occur within the same time bin. The convention is to consider this event equivalent to just one spike.). Denoting the time index by the integer variable t, one can say that $x_t^k = 1$ whenever the k-th neuron emits a spike during the t-th time bin, while $x_t^k = 0$ otherwise. This standard procedure transforms experimental data into sequences of binary patterns (see Figure 1).

A *spike pattern* is the spike-state of all the measured neurons at time bin t, which 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 $\boldsymbol{x}_{t,r} := \left[\boldsymbol{x}_s\right]_{s=t}^r$. While the length of the spike block $\boldsymbol{x}_{t,r}$ is $r - t + 1$, is also useful to consider spike blocks of infinite length starting from time $t = 0$, which are denoted by $\boldsymbol{x}$. Finally, is this paper, we consider that a *spike*

train is either a spike block of finite length or an infinite sequence of spiking patterns, which will be useful later when discussing asymptotic properties. The set of all possible spike blocks of length R corresponding to a network of N neurons is denoted by $\mathcal{A}_R^N$. The set of all spike blocks of infinite length is denoted by $\Omega \equiv \mathcal{A}_{\mathbb{N}}^N$, which is a useful mathematical object as clarified below. Let us define $proj_R : \Omega \to \mathcal{A}_R^N$ the natural projection given by $proj_R(\boldsymbol{x}) = \boldsymbol{x}_{0,R-1}$.

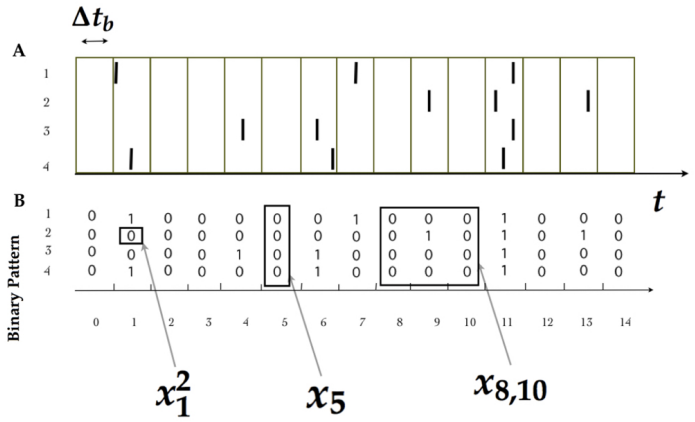

Figure 1. (**A**) Each bar indicates a spike of a neuron indexed from 1 to 4 in continuous time. (**B**) After binning Δt_b, the spiking activity is transformed into binary patterns in discrete time. We illustrate the notation used throughout this paper.

2.2. Features

Following the machine-learning nomenclature, a *feature* is a function that extracts a property of interest from the data. Formally, is defined as a function $f : \Omega \to \mathbb{R}$ that associates a real number to each $\boldsymbol{x} \in \Omega$. The feature f is said to have a temporal range or simply a range R if for every $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ such that $\boldsymbol{x} \neq \boldsymbol{y}$, one has that $f(\boldsymbol{x}) = f(\boldsymbol{y})$ if and only if $\boldsymbol{x}_{0,R-1} = \boldsymbol{y}_{0,R-1}$, that is, if f only depends on the first R spike patterns of the spike-train. A special class of features, over which this work is focused on, are binary functions consisting of finite products of spike states, i.e.,

$$f_l(\boldsymbol{x}) = \prod_{k=1}^{q} x_{t_k}^{i_k}.$$

Above, l is a shorthand notation for the set $\{t_k, i_k\}_{k=1}^q$, where $[t_k]_{k=1}^q$ and $[i_k]_{k=1}^q$ are collections of time and neuron indexes, respectively, and q is the number of spikes considered by the feature. Correspondingly, for a given index l, one has $f_l(\boldsymbol{x}) = 1$ if and only if the i_k-th neuron spikes at time t_k, for all $k \in \{1, \ldots, q\}$ in the spike-train $\boldsymbol{x}$, while $f_l(\boldsymbol{x}) = 0$ otherwise. Note that, when considering features of range $R \geq 1$, the firing times t_k are constrained within the interval $\{\, 0, \ldots, R-1 \,\}$. We define the reduced feature $\tilde{f} : \mathcal{A}_{\mathbb{N}}^R \to \mathbb{R}$ such that

$$\tilde{f}(\boldsymbol{x}_{0,R-1}) = \tilde{f}(proj_R(\boldsymbol{x})) = f(\boldsymbol{x}).$$

2.3. Statistical Structure

For a given spiking neuronal network involved in a particular experimental protocol, the measured activity usually contains a significant amount of stochasticity that is characteristic of measurements at this spatiotemporal scale. This randomness is caused mostly by:

(i) the random variation in the ionic flux of charges crossing the cellular membrane per unit time at the post synaptic button due to the binding of neurotransmitter;

(ii) the fluctuations in the current resulting from the large number of opening and closing of ion channels [37,38];
(iii) noise coming from electrical synapses [39].

To capture this stochasticity within our modeling, it is natural to endow Ω with a probabilistic structure. For this, we assume there exists a probability distribution $p\{\cdot\}$ over Ω that quantifies the intrinsic randomness that is associated to the spiking phenomena. From this point of view, all $A \subset \Omega$ are events that might take place with probability $p\{A\}$. Following a standard practice in computational neuroscience, we assume that the stochastic process generating the spikes is stationary, i.e., that their statistics do not change in time. As we discuss below, this assumption is crucial for the maximum entropy inference. Although an extension of our approach to a non-stationary scenario is possible, we focus here on the stationary case as it greatly simplifies the presentation. Using the stationary assumption, given the probability distribution of the whole process $p\{\cdot\}$ one can define a unique corresponding probability distribution over $\mathcal{A}_R^N$ following the natural projection, given by:

$$p_R\{B \in \mathcal{A}_R^N\} := p\{proj_R^{-1}(B) \in \Omega\}. \tag{1}$$

As a consequence of assuming a stochastic process guiding the neuronal activity, a feature $f : \Omega \to \mathbb{R}$ becomes a random variable. Consequently, the statistics of f are defined by

$$p\{f = a\} = p\{\boldsymbol{x} \in \Omega | f(\boldsymbol{x}) = a\}.$$

In particular, considering the feature $f(\boldsymbol{x}) = x_t^k$, one can note that individual spike-states (as well as spike patterns and spike blocks) become discrete random variables. As a convention, we denote X_t^k a random spike-state that follows an implicit underlying probability distribution $p\{\cdot\}$, while lower-case expressions (e.g., x_t^k) are used for denoting concrete realization of these random variables. The mean value of a feature f with respect to the probability $p\{\cdot\}$ is given by:

$$\mathbb{E}_p\{f\} = \sum_{\boldsymbol{x} \in \Omega} f(\boldsymbol{x}) p\{\boldsymbol{x}\}.$$

For the case of features of range R, the mean value can be expressed alternatively as:

$$\mathbb{E}_p\{f\} = \sum_{\boldsymbol{x}_{0,R-1} \in \mathcal{A}_R^N} \tilde{f}(\boldsymbol{x}_{0,R-1}) p_R\{\boldsymbol{x}_{0,R-1}\} = \mathbb{E}_{p_R}\{\tilde{f}\}$$

which is a finite sum. Above, $\tilde{f}$ is the reduced feature, as defined in Section 2.2.

2.4. Empirical Averages

Let us consider spiking data of the form $\boldsymbol{x}_{0,T-1}$, where T is the sample length. Although in general the underlying probability measure $p\{\cdot\}$ that governs the spiking activity is unknown, it is useful to use the available data to estimate the mean values of specific features. If f is a feature of range R, the empirical average value of f from the sample $\boldsymbol{x}_{0,T-1}$ is

$$A_T(f) = \frac{1}{T-R+1} \sum_{i=0}^{T-R} f(\boldsymbol{x}_{i,R-1+i}). \tag{2}$$

In particular, for features of range one, the previous expression becomes $A_T(f) = \frac{1}{T} \sum_{i=0}^{T-1} f(\boldsymbol{x}_i)$.

An interesting questions is under which conditions $A_T(f) \to \mathbb{E}_p\{f\}$ as T grows. This, and other convergence issues, are explored in Section 4.

3. Inference of the Statistical Model with the MEP

Following Section 2.3, the probability measure $p\{\cdot\}$ represents the inherent stochasticity of the spiking neural population under consideration. As $p\{\cdot\}$ is unknown, one would like to infer it from data. In the sequel, we first introduce the general MEP as a method for inferring $p\{\cdot\}$. Then, we show this principle for the case where only synchronous constraints are considered. Finally, we present the case of where non-synchronous correlations are included to constrain the maximization problem.

3.1. Fundamentals of the MEP

The MEP was first proposed by E. T. Jaynes as a way for estimating probability distributions when the information for performing the inference is scarce [40]. Rooted in principles of statistical physics, this approach selects a probability measure that satisfies the evidence supported by the available information while leaving all other aspects as random as possible. For quantifying the corresponding randomness, Jaynes shows that the most natural metric is the Shannon entropy [41]. The probability measure found by this procedure is known as the *maximum entropy distribution*.

Formally, the MEP is a concave constrained maximization problem, where the constraints that define the optimization space correspond to the available information that guide the inference process. Accordingly, if additional constraints are introduced then the optimization space is reduced; this corresponds to the informative power of new information, which restricts the space of models that are consistent with it.

The inference procedure based on the MEP follows the following steps:

I. Choose K features of interest $f_1, \ldots, f_K$ (cf. Section 2.2).

II. Using the available data $\boldsymbol{x}_{0,T-1}$, compute the empirical averange of each feature $A_T(f_k) := c_k$.

III. Assuming stationarity, define the space of statistical models $\mathcal{M}(c_1, \ldots, c_K) \subset \mathcal{M}$ given by

$$\mathcal{M}(c_1, \ldots, c_K) = \{g \in \mathcal{M} |\, \mathbb{E}_g\{f_1\} = c_1, \ldots, \mathbb{E}_g\{f_K\} = c_K\},$$

where $\mathcal{M}$ is the set of probability measures and $\mathcal{M}(c_1, \ldots, c_K)$ is the family of probability measures that are consistent with the empirical mean values $c_1, \ldots, c_K$ obtained in Step II.

IV. Defining the entropy rate of the stochastic process as

$$\mathcal{S}\{p\} = \lim_{t\to\infty} \frac{1}{t} \sum_{\boldsymbol{x}_{0,t-1} \in \mathcal{A}_t^N} p_t\{\boldsymbol{x}_{0,t-1}\} \log \frac{1}{p_t\{\boldsymbol{x}_{0,t-1}\}}, \tag{3}$$

find the maximum entropy process, characterized by the probability measure

$$\hat{p} = \underset{q \in \mathcal{M}(c_1,\ldots,c_k)}{\arg\max} \; \mathcal{S}\{q\}. \tag{4}$$

Some small remarks are to be said about this procedure. One can think of this as a data-driven algorithm, whose input is the data $\boldsymbol{x}_{0,T-1}$ and output is the maximum entropy measure $\hat{p}$. The first two steps of the process are known in the machine learning literature as "feature selection" and "feature extraction", respectively (see, e.g., [42,43]). The goal of these steps is to reduce the dimensionality of the input for the subsequent stages to prevent the selected model to overfitting the data (i.e., to include in the model effects of noise and biases due to the finiteness of the data). Hence, what drives the model selection stages is not the whole data but the quantities $c_1, \ldots, c_K$, which define the space to be explored in Step IV.

Steps III and IV are known as "model selection". According to the the machine learning jargon, these steps deliver a generative model, in the sense the obtained model can later be used to generate new data. In this sense, it is interesting that, although the data are finite, the entropy rate calculated in Step IV is computed over all spike blocks of all lengths t, which is possible due to the generative

nature of the candidate models. The inputs for the model selection stages are not the entire data $\boldsymbol{x}_{0,T-1}$ but only the values $c_1, \ldots, c_K$, which represent the knowledge obtained from the data that guides the search in the space of candidate models. Moreover, as these quantities represent all the available knowledge one has about the underlying stochastic process generating the spikes, one would like to select a model that reflect that information while making no further assumptions. By recalling the work made by Claude Shannon on the analysis of information sources (cf. [44] and references therein), one can interpret the entropy rate as a measure of how hard is to predict the realization of a stochastic process. This implies, in turn, that the maximum entropy measure within $\mathcal{M}\{c_1, \ldots, c_K\}$ is the most random, i.e., unstructured, among those which satisfies the constraints $A_T(f_1) = c_1, \ldots, A_T(f_K) = c_k$. Although the framework presented above is general enough to encompass the cases considering synchronous and non-synchronous constraints [21], when considering features of range $R > 2$ the problem go beyond the realm of equilibrium statistical mechanics. We present an alternative way general enough to deal with systems assuming non-synchronous constraints. In Section 3.2, we present the method for finding the maximum entropy measure when only synchronous features are selected, leaving for Section 3.3 the more general situation including non-synchronous constraints.

3.2. Time-Independent Constraints

Assuming only synchronous interactions is the equivalent to only considering features of range one (i.e., features that consider neurons at the same time index, cf. Section 2.2), which leads to restricting the candidate models to those in where the present state is statistically independent of past and future states. Moreover, by the assumption of stationarity, this leads to modeling the collective spiking activity as a sequence of i.i.d. random variables. The challenge, in this case, is to estimate the corresponding distribution as reliably as possible. A large portion of the literature of maximum entropy spike train statistics focuses on synchronous interactions between neurons, implicitly neglecting interactions across time. Although this assumption induces a strong simplification, the resulting models have proven to be rich in structure and can provide interesting results and insights about the neural code [6,12]. In the following, we recall how this problem can be addressed using the MEP.

As a consequence of the assumptions of temporal independence and stationarity, it can be shown that Equation (4) is reduced to

$$\hat{p}_1 = \underset{q_1 \in \mathcal{M}_1(c_1,\ldots,c_k)}{\arg\max} \sum_{\boldsymbol{x}_0 \in \mathcal{A}_1^N} q_1\{\boldsymbol{x}_0\} \log \frac{1}{q_1\{\boldsymbol{x}_0\}} \tag{5}$$

where $\mathcal{M}_1(c_1, \ldots, c_k)$ corresponds to the set of distributions q_1 over $\mathcal{A}_1^N$ (cf. range one projections in Equation (1)) such that the constraints $\mathbb{E}_{q_1}\{f_k\} = c_k$ are satisfied for each $k = 1, \ldots, K$. Note that the above sum is over the 2^N possible spike patterns, being a simpler condition than Equation (4). In fact, following a simple argument based on Lagrange multipliers and the concavity of the entropy, it can be shown that the distribution $\hat{p}_1$ that solves Equation (5) is unique. Moreover, it is a Boltzmann–Gibbs distribution [41]:

$$\hat{p}_1\{\boldsymbol{x}_0\} = \frac{e^{-\mathcal{H}_{\boldsymbol{\beta}}(\boldsymbol{x}_0)}}{Z(\boldsymbol{\beta})} \quad \forall \boldsymbol{x}_0 \in \mathcal{A}_1^N; \quad Z(\boldsymbol{\beta}) = \sum_{\boldsymbol{x}_0 \in \mathcal{A}_1^N} e^{-\mathcal{H}_{\boldsymbol{\beta}}(\boldsymbol{x}_0)}, \tag{6}$$

where $\mathcal{H}_{\boldsymbol{\beta}}$ is referred as the *energy or potential* function

$$\mathcal{H}_{\boldsymbol{\beta}}(\boldsymbol{x}_0) = \sum_{k=1}^{K} \beta_k \tilde{f}_k(\boldsymbol{x}_0), \tag{7}$$

where $\boldsymbol{\beta} \in \mathbb{R}^K$ is the vector of Lagrange multipliers. Following the statistical physics literature, $Z(\boldsymbol{\beta})$ is called the *partition function*, whose logarithm is referred as *free energy*.

Conversely, from the uniqueness property of the maximum entropy distribution, one can conclude that there is only one Boltzmann–Gibbs distribution $\hat{p}_1$ that belongs to $\mathcal{M}(c_1,\ldots,c_K)$, being the only solution of Equation (5). Interestingly, this alternative approach is much easier to solve the original optimization problem (In particular, $\mathcal{M}_1\{c_1,\ldots,c_k\}$ is not easy to parameterize and hence the application of standard techniques of convex optimization (e.g., gradient decent) is not straightforward.). In effect, one only needs to find the values of the parameter vector β_k that reproduces the empirical average values $c_1,\ldots,c_K$. Moreover, it is known that, for any Boltzmann–Gibbs distribution p_1, the following holds:

$$\frac{\partial \ln Z(\boldsymbol{\beta})}{\partial \beta_k} = \mathbb{E}_{\hat{p}_1}(\tilde{f}_k). \tag{8}$$

Therefore, using Equation (8), one could find the appropriate values of $\boldsymbol{\beta}$ for which $\mathbb{E}_{\hat{p}_1}\{\tilde{f}_k\} = c_k$ are satisfied (However, for practical purposes, this problem cannot be solved for systems with $N > 20$ [8], so alternative procedures are needed. For the interested reader, we refer to [7,14,45,46]).

3.3. Non-Synchronous Constraints

A generalization of the previous approach is to include average values of features corresponding to interactions in the spiking activity across time as constraints. This is a natural assumption in biological spiking networks as it is expected that the spike of one neuron influence other subsequent spikes.

Statistical models with time-dependencies can be generated with the MEP by introducing features that involve different time indexes. In effect, selecting features of range R induces interdependencies and a corresponding "memory" in the model of length $R-1$, and thus it is natural to look for the best suited Markov chain over the state space $\mathcal{A}_N^R$. A Markov chain model would allow expressing the probability of a spike train $\boldsymbol{x}_{0,T}$ for $T > R$ as

$$p\{\boldsymbol{x}_{0,T}\} = \pi\{\boldsymbol{x}_{0,R-1}\}P\{\boldsymbol{x}_{1,R}|\boldsymbol{x}_{0,R-1}\}\cdots P\{\boldsymbol{x}_{T-R,T-1}|\boldsymbol{x}_{T-R+1,T}\},$$

where P is a transition probability matrix (note that $P\{\cdot,\cdot\}$ has a consistency requirement: for $\boldsymbol{w},\boldsymbol{y} \in \mathcal{A}_N^R$, $P\{\boldsymbol{w}|\boldsymbol{y}\} > 0$ only when $\boldsymbol{y}_{1,R-1} = \boldsymbol{w}_{0,R-2}$) and π is a corresponding invariant probability distribution (which is unique due to the ergodicity assumption, cf. Section 3.3.1) associated to P. Note that, due to the stationarity condition, the transition probabilities $P\{\cdot|\cdot\}$ are constant over the realization of the whole process (see Appendix A for more details.).

Following the MEP, as described in Section 3.1, we look for a procedure for finding a Markov transition matrix P that maximizes its entropy rate while satisfying some constrains given the empirical averages of observables $f_1,\ldots,f_K$. For ergodic Markov chains, a well-known calculation (that can be found, e.g., in [44]) shows that the entropy rate, as given by Equation (3), is equivalent to the following simple expression:

$$\mathcal{S}_{KS}(\pi,P) = -\sum_{i,j\in\mathcal{A}_N^R}\pi_i\sum_j P_{ij}\log P_{ij}. \tag{9}$$

where $\pi_i = \pi\{\boldsymbol{x}_{0,\,R-1} = i\}$ and $P_{ij} = P\{j|i\}$ for $i,j \in \mathcal{A}_R^N$. It is important to notice that Equation (9) corresponds to the *Kolmogorov–Sinai entropy (KSE)* in the dynamical systems literature [47]. In general, Equation (9) is larger when, for a fixed i, the conditional probabilities P_{ij} are closer to an uniform distribution, i.e., when knowing the transition statistics gives little certainty about the next step.

It can be shown that, if the considered features do not involve correlations across time (i.e., they are features of range one, cf. Section 2.2), then the resulting transition probabilities are such that the corresponding stochastic process is i.i.d (i.e., when $P_{ij} = \pi_j$). Interestingly, in this scenario, Equation (9) reduces to the Shannon entropy of π_i. This clarifies that this approach based on Markov chains extends the classical MEP and the results presented in Section 3.2.

Finding the MEMC raises, however, some extra technicalities with respect to the time-independent case. Recall that the goal is no longer to estimate a probability distribution, but to reconstruct from data a transition matrix P and a corresponding invariant measure π. On the one hand, the challenge is that as P and π are not independent parameters of the process (π has to be the eigenvector associated with the unitary eigenvalue of P [48]), and, on the other hand, although Lagrange multipliers method can still be applied for constraints of range two or more, the extension for non-synchronous constraints is not straightforward. For this reason, in the sequel, we explore an alternative route to build the MEMC based on the transfer matrix technique. This technique also provides further insightful connections with statistical physics and thermodynamics from which its large deviation properties arise naturally.

3.3.1. Transfer Matrix Method

To find the MEMC associated with non-synchronous constraints, we follow the same ideas presented above in the time-independent case, but using different tools. We present them here.

Let us consider the set of features chosen to constrain the maximization of entropy rate (Step I in Section 3.1). We assume that the features chosen have a finite maximum range R. From these features, one can build the energy function $\mathcal{H}_\beta$ (Equation (7)) of finite range R as a linear combination of these features. Using this energy function, we build a matrix denoted by $\mathcal{L}_{\mathcal{H}_\beta}$, so that for every $y, w \in \mathcal{A}_R^N$ its entries are given as follows:

$$\mathcal{L}_{\mathcal{H}_\beta}(y,w) = \begin{cases} e^{\mathcal{H}_\beta(yw_{R-1})} & \text{if } y_{1,R-1} = w_{0,R-2} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

By yw_{R-1} we mean the word obtained by concatenation of y_1 and $w_{1,R-1}$. In the particular case of energy functions associated to range one features, the above matrix is defined as $\mathcal{L}_{\mathcal{H}_\beta}(y,w) = e^{\mathcal{H}_\beta(y)}$. Assuming $\mathcal{H}_\beta > -\infty$, the elements of the matrix $\mathcal{L}_{\mathcal{H}_\beta}$ are non-negative, which in turn implies ergodicity. Moreover, the matrix is primitive by construction, thus it satisfies the Perron–Frobenius theorem [49]. Hereafter, $\mathcal{L}_{\mathcal{H}}$ will be referred as the Ruelle–Perron–Frobenius matrix (RPF). Let us denote be ρ the largest eigenvalue of $\mathcal{L}_{\mathcal{H}}$, which because it satisfies the Perron–Frobenius theorem is an eigenvalue of multiplicity one and strictly larger in modulus than the rest of the eigenvalues [49]. We denote by U and V left and right eigenvectors of $\mathcal{L}_{\mathcal{H}_\beta}$ corresponding to the eigenvalue ρ. Notice that $U_i > 0$ and $V_i > 0$, for all $i \in \mathcal{A}_R^N$.

The RPF matrix is not the Markov matrix we are looking for, moreover, is not a stochastic matrix, but can be converted into a stochastic matrix. We recall that for an irreducible matrix M with spectral radius λ, and positive right eigenvector $\mathbf{v}$ associated to λ, then the *stochasticization* of M is the following stochastic matrix:

$$S(M) = \frac{1}{\lambda} D^{-1} M D \,, \tag{11}$$

where D is the diagonal matrix with diagonal entries $D(i,i) = \mathbf{v_i}$. Thus, in our context, the MEMC transition matrix is given as follows:

$$P = S(\mathcal{L}_{\mathcal{H}_\beta}), \tag{12}$$

which is the main output of the MEMC approach (see Figure 2). The unique stationary probability measure π associated to P is explicitly given by

$$\pi_i := \frac{U_i V_i}{\langle U, V \rangle}, \quad \forall i \in \mathcal{A}_R^N, \tag{13}$$

where $\langle U, V \rangle$ is the standard inner product in $\mathbb{R}^{NR}$ (we refer the reader to [49] for details and proofs).

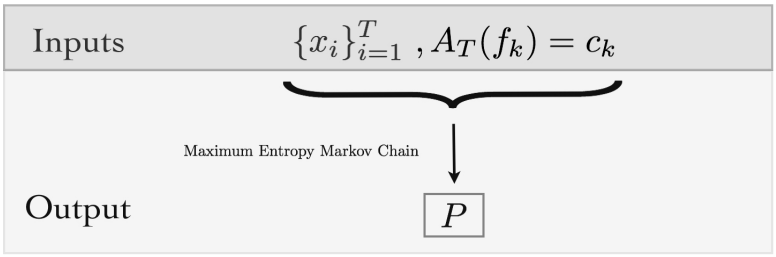

Figure 2. Algorithmic view of the maximum entropy Markov chains (MEMC): Inputs are the spike trains $\{x_i\}_{i=1}^{T}$ and the average values of a set of features. The output is the MEMC transition matrix P.

3.3.2. Thermodynamic Formalism

In the previous section, we have shown how to obtain the transition matrix and the invariant measure of a Markov chain. However, we have not yet included the constraints (we have just used the features to build the energy function); in other words, we have not yet fit the parameters of the MEMC. To fit the maximum entropy parameters, let us introduce the following quantity,

$$\mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right] = \sup_{q \in \mathcal{M}_{st}} \left\{ \mathcal{S}\{q\} + \mathbb{E}_q\left\{\mathcal{H}_{\boldsymbol{\beta}}\right\} \right\} \tag{14}$$

where $\mathcal{M}_{st}$ is the set of all stationary probability measures in $\mathcal{A}_N^R$ and $\mathbb{E}_q\left\{\mathcal{H}_\beta\right\} = \sum_{k=1}^{K} \beta_k \mathbb{E}_q\left\{f_k\right\}$ is the average value of $\mathcal{H}_\beta$ with respect to q. Solving the optimization problem in Equation (14), one gets the Markov measure we are looking for. Indeed, one knows from the thermodynamical formalism (see [50]) that for our energy function $\mathcal{H}_\beta$ of range $R \geq 2$, there exists an unique translation invariant (stationary) Markov measure p associated to $\mathcal{H}_\beta$ for which one has the constant $M > 1$ such that,

$$M^{-1} \leq \frac{p\{\boldsymbol{x}_{1,n}\}}{\exp(\sum_{k=1}^{n-R+1} \mathcal{H}(\boldsymbol{x}_{k,k+R-1}) - (n+R-1)\mathcal{P}[\mathcal{H}_\beta])} \leq M, \tag{15}$$

that attains the supremum in Equation (14), that is $\mathcal{S}\{p\} + \mathbb{E}_p\left\{\mathcal{H}_{\boldsymbol{\beta}}\right\}$.

The quantity $\mathcal{P}[\mathcal{H}_\beta]$ is called *topological pressure*, which plays the role of the free energy in the statistical mechanics. The measure p, as defined by Equation (15), is known as the Gibbs measure in the sense of Bowen. Note that one can show that MEMCs are particular cases of these measures, associated to finite-range energy functions. Moreover, Equation (6) is a particular case of Equation (15), when $M = 1$ and $\mathcal{H}_\beta$ is an energy function of range one.

The average values of the features, their correlations, as well as their higher cumulants can be obtained by taking the successive derivatives of the topological pressure with respect to their conjugate parameters $\boldsymbol{\beta}$. This explains the important role played by the topological pressure in this framework. In general,

$$\frac{\partial^n \mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]}{\partial \beta_k^n} = \kappa_n \quad \forall k \in \{1, ..., K\}, \tag{16}$$

where κ_n is the cumulant of order n (see Appendix B).

In particular, taking the first derivative:

$$\frac{\partial \mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]}{\partial \beta_k} = \mathbb{E}_p\{f_k\} = c_k, \quad \forall k \in \{1, ..., K\}, \tag{17}$$

where $\mathbb{E}_p\{f\}$ is the the average of f_k with respect to p (maximum entropy measure), which is equal (by assumption) to the average value of f_k with respect to the empirical measure from the data c_k, that constraint of the maximization problem. These equations suggest a relationship with the logarithm of the free energy or log partition function of the Boltzmann–Gibbs distribution. Indeed, for range one

potentials (time-independent Maximum entropy distributions), $\rho(\boldsymbol{\beta}) = Z(\boldsymbol{\beta})$ and $\mathcal{P}[\mathcal{H}_\beta] = \ln Z(\boldsymbol{\beta})$ which relates Equation (8) with Equation (17) (For a detailed example see Section 5.2). This problem of estimating the MEMC parameters become computationally expensive for big matrices. However, there exist efficient algorithms to estimate the parameters for the Markov maximum entropy problem in the literature [45].

4. Large Deviations and Applications in MEMC

4.1. Preliminary Considerations

This subsection reviews two elementary tools for studying the convergence of random variables while providing corresponding references. In the sequel, first the central limit theorem is introduced in Section 4.1.1, and then large deviation theory is discussed in Section 4.1.2.

4.1.1. Central Limit Theorem

Let us first assume that one can have access to arbitrarily large data sequences. Consider $t \in \mathbb{N}$ and let $x_{0,t-1}$ be the spike-block of length t (which is allowed to increase), and let $f : \Omega \to \mathbb{R}$ be an arbitrary feature (not necessarily belonging to the set of features chosen to fit the MEMC). In this section, we establish asymptotic properties of $A_t(f)$ sampled with respect to the MEMC characterized by $p\{\cdot\}$.

Throughout this work, we assume that $p\{\cdot\}$ is an ergodic Markov probability measure, meaning that every spiking block in $\mathcal{A}_R^N$ is attainable from every other block in the Markov chain within R time steps as discussed in Section 3. Thanks to the ergodic assumption, it is guaranteed that the empirical averages become statistically more accurate as the sampling size grows [51], i.e.,

$$A_t(f) \to \mathbb{E}_p\{f\}.$$

However, the above result does not clarify the rate at which the estimate accuracy improves. To answer this question, one can use the central limit theorem (CLT) for ergodic Markov chains (see [52]). This theorem states that there exists a constant $\sigma > 0$ such that for large values of t, the random variable $\frac{\sqrt{t}}{\sigma}\big[A_t(f) - \mathbb{E}\{f\}\big]$ distributes as a standard Gaussian random variable (technically, the central limit theorem says that

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

where the convergence is in distribution), with σ being the square-root of the auto-covariance function of f [52]. This, in turn, implies that "typical" fluctuations of $A_t(f)$ around its mean value $\mathbb{E}\{f\}$ are of the order of $\sigma/\sqrt{t}$.

4.1.2. Large Deviations

Although the central limit theorem for ergodic Markov chains is accurate in describing typical events (which are fluctuations around the mean value), it does not say anything about the likelihood of larger fluctuations. Even though it is clear that the probability of such large fluctuations goes to zero as the sample size increases, it is valuable to describe the corresponding decrease rate. In particular, one says that an empirical average $A_t(f)$ satisfies a large deviation principle (LDP) with rate function I_f, defined as

$$I_f(s) := -\lim_{t\to\infty} \frac{1}{t} \log p\{A_t(f) > s\}, \tag{18}$$

if the above limit exists. Intuitively, the above condition for large t implies that $p\{A_t(f) > s\} \approx e^{-tI_f(s)}$. In particular, if $s > \mathbb{E}_p\{f\}$ the Law of Large Numbers (LLN) guarantees that $p\{A_t(f) > s\}$ tends to zero as t grows; the rate function quantifies the speed at which this happens.

Calculating I_f directly, i.e., by using the definition (Equation (18)), can be a formidable task. However, the Gärtner–Ellis theorem provides a smart shortcut for avoiding this problem [27]. To this end, let us introduce the *scaled cumulant generating function* (SCGF) (the name comes from the fact that the n-th cumulant of f can be obtained by t successive differentiation operations over of $\lambda_f(k)$ with respect to k, and then evaluating the result at $k = 0$) associated to the random variable f, by

$$\lambda_f(k) =: \lim_{t\to\infty} \frac{1}{t} \ln \mathbb{E}_p \left[e^{tkA_t(f)} \right], \quad k \in \mathbb{R}, \tag{19}$$

when the limit exists (further general details about cumulant generating functions are found in Appendix B). Note that, while $A_t(f)$ is an empirical average taken over a sample, the expectation in Equation (19) is taken over the probability distribution given by the corresponding model $p\{\cdot\}$. If λ_f is differentiable, then the Gärtner-Ellis theorem ensures that the average $A_t(f)$ satisfies a LDP with rate function given by the Legendre transform of λ_f, that is

$$I_f(s) = \max_{k\in\mathbb{R}} \{ks - \lambda_f(k)\}. \tag{20}$$

Therefore, in summary, one can study the large deviations of empirical averages $A_t(f)$ by first computing their SCGF from the selected model and then finding their Legendre transform.

One of the most useful applications of the LDP is to estimate the likelihood that $A_t(f)$ adopts a value far from its expected value. To illustrate this, let us assume that $I_f(s)$ is a positive differentiable convex function (A classical result in LDP states that $I_f(s)$ is a convex function if $\lambda_f(k)$ is differentiable [28]. For a discussion about the differentiability of $\lambda_f(k)$ see [53].). Then, because of the properties of convex functions $I_f(s)$ has a unique global minimum. Denoting this minimum by s^*, it follows from the differentiability of $I_f(s)$ that $I_f(s^*) = 0$, and using properties of the Legendre transform $s^* = \lambda_f'(0) = \lim_{t\to\infty} \mathbb{E}_p(f)$. This is the LLN, i.e., $A_t(f)$ gets concentrated around s^*. Consider a value $s \neq s^*$ and assume that $I_f(s)$ admits a Taylor series around s^* given by

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

Since s^* must correspond to a zero and a minimum of $I(s)$, the first two terms in this series vanish, and as $I(s)$ is convex function $I''(s) > 0$. For large values of k, we obtain from Equation (18)

$$\begin{aligned} p\{A_t(f) > s\} &\approx e^{-tI_f(s)} \\ &\approx e^{-t\left(\frac{I_f''(s^*)(s-s^*)^2}{2}\right)} \end{aligned} \tag{21}$$

so the small deviations of $A_t(f)$ around s^* are Gaussian-distributed as for i.i.d. sums $1/I_f''(s^*) = \lambda_f''(0) = \sigma^2$. In this sense, large deviation theory can be seen as an extension of the CLT because it gives information not only about the small deviations around s^* but also about large deviations (not Gaussian) of $A_t(f)$.

4.2. Large Deviations for Features of MEMC

In this section, we focus on the statistical properties of features sampled from the inferred MEMC. For example, one may be interested in measuring the probability of obtaining "rare" average values of features such as firing rates, pairwise correlations, triplets or spatiotemporal events. This characterization is relevant as these features are likely to play an important role in neuronal information processing, and rare values may hinder the whole enterprise of conveying information. We show in this section how to obtain the large deviations rate functions of arbitrary features through the Gärtner–Ellis theorem via the SCGF. In particular, we show that the SCGF can be directly obtained from the inferred Markov transition matrix P.

Consider MEMC taking values on the state space $\mathcal{A}_R^N$ with transition matrix P. Let f be a feature of range R which consider only the block and $k \in \mathbb{R}$, we introduce $\widetilde{P}^{(f)}(k)$, the *tilted transition matrix by* f of P, parameterized by k, whose elements are given by:

$$\widetilde{P}_{ij}^{(f)}(k) = P_{ij} e^{kf(j)} \quad i, j \in \mathcal{A}_R^N. \tag{22}$$

For transition matrices P inferred from the MEP, the tilted transition matrix can be built directly from the spectral properties of the transfer matrix Equation (10) as follows,

$$\begin{aligned} \widetilde{P}_{ij}^{(f)}(k) &= \frac{e^{\mathcal{H}_\beta(i,j)} V_j}{V_i \rho} e^{kf(j)} \\ &= \frac{e^{\left[\mathcal{H}_\beta(i,j) + kf(j)\right]} V_j}{V_i \rho} \quad i, j \in \mathcal{A}_R^N. \end{aligned} \tag{23}$$

Recall that V is the right eigenvector of the transfer matrix $\mathcal{L}$. Here, we also have used the shortcut notation $\mathcal{H}_\beta(i, j)$ to indicate that the energy function takes the contributions from the blocks i and j. Remarkably, the feature f does not need to belong to the set of chosen features to fit the MEMC.

Now, we can take advantage of the structure of the given process in order to obtain more explicit expressions for the SCGF $\lambda_f(k)$, for instance, if one considers i.i.d. random variables X then, from the very definition, one can obtain that

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

which is the case of range one features. Thus, using Equation (22), we get that the maximum eigenvalue of the tilted matrix, denoted by $\rho(\widetilde{P}_f(k))$ is,

$$\rho\big(\widetilde{P}_f(k)\big) = \sum_j \pi_j e^{kf(j)} \quad j \in \mathcal{A}_1^N.$$

Since $\widetilde{P}_f$ is a positive matrix, the Perron–Frobenius theorem ensures the uniqueness of ρ .

Next, for the case of additive features, one deals with positive Markov chains, and under the assumption of ergodicity, an straightforward calculation (see, for instance, [54]) leads us to obtain that

$$\lambda_f(k) = \ln\big(\rho\big(\widetilde{P}^{(f)}\big)\big). \tag{24}$$

It also can be proven that $\lambda_f(k)$, in this case, is differentiable [54], setting up the scene to apply the Gärtner–Ellis theorem, which bypasses the direct calculation of $p\{A_T(f) > s\}$ in Equation (18), i.e., having $\lambda_f(k)$, its Legendre transform leads to the rate function of f as shown in Figure 3.

4.3. Large Deviations for the Entropy Production

A stochastic process is said to be in equilibrium if one cannot notice the effect of time on it. It is worth noticing that non-equilibrium stochastic processes are natural candidates to model spike train statistics as time plays a non-symmetrical role [25]. One of the consequences of including features of range $R > 1$ as constraints in the maximum entropy problem is that it opens the possibility to break the time-reversal symmetry present in the time-independent models. This captures the irreversible character of the underlying biological process and, thus, allows fitting more realistic statistical models from the biological point of view.

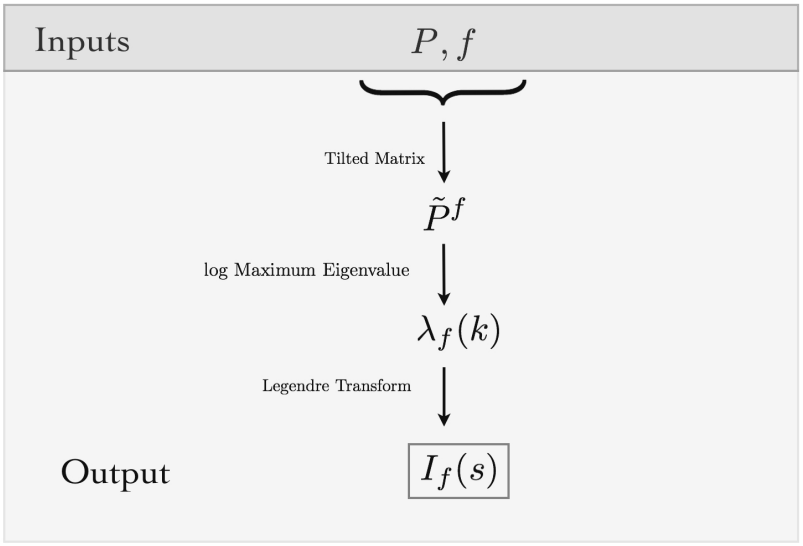

Figure 3. Algorithmic view of the method: Inputs are the maximum entropy Markov transition matrix and a feature. From the inputs, the tilted transition matrix can be built. The rate function of the feature is obtained as the Legendre transform of the log maximum eigenvalue of the tilted transition matrix. Using this framework, we can estimate the large deviations of the average values of the features.

To characterize this mathematically, we study how the distribution $p\{\cdot\}$ changes when the time order is reversed. For this aim, let us consider a spike block $\boldsymbol{x}_{0,T-1} = \boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, x_{T-1}$ containing T spike patterns, and define the time-reversed spike block $\boldsymbol{x}^{(R)}_{0,T-1}$ obtained by re-ordering the time index in reverse order, i.e., $\boldsymbol{x}^{(R)}_{0,T} = \boldsymbol{x}_{T-1}, \boldsymbol{x}_{T-2}, \ldots, \boldsymbol{x}_2, \boldsymbol{x}_0$.

A spiking network modeled by $p\{\cdot\}$ is said to be in equilibrium if $p\{\boldsymbol{x}_{0,T}\} = p\{\boldsymbol{x}^{(R)}_{0,T}\}$ for all $\boldsymbol{x}$ [55]. For a homogeneous discrete time ergodic Markov chain characterized by the Markov measure $p(\pi, P)$ taking values in $\mathcal{A}^N_R$, to be in equilibrium is equivalent to satisfy the *detailed balance conditions*, which is given by the following set of equalities:

$$\pi_i P_{ij} = \pi_j P_{ji}, \quad \forall i, j \in \mathcal{A}^N_R.$$

Conversely, when these conditions are not satisfied, the statistical model of the spiking activity is said to be a non-equilibrium system. Since non-equilibrium is expected to occur generically in neuronal network models, one would like to quantify how far from equilibrium is the inferred MEMC. For this purpose, one can define the *information entropy production* (IEP) for p, which is given by

$$IEP(p) := \lim_{t\to\infty} \frac{1}{t} \ln \left[\frac{p\{\boldsymbol{x}_{0,t-1}\}}{p\{\boldsymbol{x}^{(R)}_{0,t-1}\}} \right],$$

when the limit exists. For the maximum entropy Markov measure $p(\pi, P)$, the IEP is explicitly given by:

$$IEP(\pi, P) = \frac{1}{2} \sum_{i,j\in\mathcal{A}^N_R} \left[\pi_i P_{ij} - \pi_j P_{ji}\right] \log \frac{\pi_i P_{ij}}{\pi_j P_{ji}}, \tag{25}$$

(see [56] for the calculation). We remark that it is still possible to obtain the information entropy production rate also in the non-stationary case. Clearly, for features of range one, $IEP = 0$ always, meaning that the process is time-reversible, therefore the probabilities of every path and its corresponding time-reversal path are equal. For features of range $R > 1$, $IEP > 0$ generically (we refer the interested reader to [25] for details and examples).

However, since in practice one only have access to limited amount of data, a natural question is to ask for the entropy production of the system considered up to a finite amount of time. It turns out that this characterization can be obtained through a LDP. With this in mind one may define the following feature:

$$W_T(\boldsymbol{x}_{0,T-1}) = \frac{1}{T} \ln \left[\frac{p(\boldsymbol{x}_{0,T-1})}{p(\boldsymbol{x}^{(R)}_{0,T-1})} \right].$$

Since we have assumed that samples are produced by a stationary ergodic Markov chain characterized by $p(\pi, P)$, the ergodic theorem assures that for p-almost every sample, the quantity W_t when t goes to infinity converges, and it is by definition the IEP,

$$\lim_{t\to\infty} W_t(\boldsymbol{x}_{0,t-1})) = IEP(\pi, P).$$

Once we have the convergence for W_t, we may ask for its large deviation properties. Under the same idea above, and following [57], we introduce the following matrix:

$$F_{ij} = P_{ij} \ln \left[\frac{\pi_i P_{ij}}{\pi_j P_{ji}} \right]^k \quad i, j \in \mathcal{A}_R^N,$$

this matrix help us to build the SCGF associated to W_t, through the logarithm of the maximum eigenvalue $\rho_F(k)$. Using the Gärtner–Ellis theorem one gets the rate function $I_W(s)$ for the IEP.

4.4. Large Deviations and MEMC Distinguishability

It is clear that there exist a relationship between accuracy of the estimation and sample size. The larger the sample size the more information is available and the uncertainty diminish. In the context of maximum entropy models, this idea has been well conceptualized using tools from information geometry [30,58]. The main idea of this approach is that the maximum entropy models form a manifold of probability measures whose coordinates are the parameters $\boldsymbol{\beta}$. Consider a spike train dataset $\boldsymbol{x}_{0,T-1}$ consisting of T spike patterns obtained from a spiking neuronal network. Given a set of features $\{f_k\}_{k=1}^K$ and their empirical averages, one may infer the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ characterizing the MEMC $p(\pi, P)$. We may use the inferred MEMC to generate a sample $\boldsymbol{x'}_{0,T-1}$ of the same size as the original dataset. Considering the same set of features one may apply again the MEP to infer a new set of parameters $\boldsymbol{\beta'}$ from $\boldsymbol{x'}_{0,T-1}$, which is, in general, expected to be different from $\boldsymbol{\beta}$. These maximum entropy models will belong to a certain volume in the manifold which will decrease as the sample size increase [30]. On the other hand, increasing the sample size of $\boldsymbol{x'}_{0,T-1}$, one expects that the Markov chain $p'(\pi', P')$ specified by $\boldsymbol{\beta'}$ gets "closer" to the one characterized by $\boldsymbol{\beta}$. The idea of relating a distance in the parameter space with a distance in the space of probability measures can be rigorously formulated using large deviations techniques. Let us start by defining the relative entropy between these two MEMC (Gibbs measures in the sense of Bowen in Equation (15)), which provides a notion of "distance" (the relative entropy is not a metric as is not symmetric and do not satisfy the triangle inequality). To do that in the context of MEMCs, consider a Gibbs measure p associated to the energy function $\mathcal{H}_\beta$, and let q be another Gibbs measure. The Ruelle–Föllmer theorem gives us an expression for the relative entropy density between two Gibbs measures in terms of the pressure, the entropy rate and the expected value of the energy function with respect to q (see [29]), as follows:

$$\mathsf{d}(q \mid p) = \mathcal{P}[\mathcal{H}_\beta] - S(q) - \mathbb{E}_q(\mathcal{H}_\beta). \tag{26}$$

Observe that if $\mathsf{d}(q \mid p) = 0$, we obtain the variational characterization of Gibbs measures in Equation (14).

Consider the potential $\mathcal{H}_{\boldsymbol{\beta}} = \sum_{k=1}^{K} \beta_k f_k$ associated with a MEMC $p(\pi, P)$. Given a set of empirical averages $A_t(f_k)$ generated by a sample of $p(\pi, P)$ we obtain new maximum entropy parameters $\boldsymbol{\beta'}$. The probability that the maximum entropy parameters $\boldsymbol{\beta'}$ associated with an ergodic Markov Chain $p'(\pi', P')$ get close to $\boldsymbol{\beta}$ follows the following large deviation principle [28]:

$$\lim_{\delta \to 0} \lim_{t \to \infty} \frac{-1}{t} \ln \mathbb{P}\Big(\mid \boldsymbol{\beta} - \boldsymbol{\beta'} \mid \in \Delta\delta \Big) = \mathsf{d}(p \mid p'), \tag{27}$$

where $\Delta\delta = [-\delta, \delta]^K$. Calling and the vector $\delta\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta'}$ and choosing $\Delta\delta$ close to 0, we informally rewrite the above corollary in the form:

$$\frac{-1}{t} \ln \mathbb{P}\Big(\mid \delta\boldsymbol{\beta} \mid \in \Delta\delta \Big) \underset{t \to \infty}{\longrightarrow} \mathsf{d}(p \mid p'). \tag{28}$$

Thus, for large T, we get:

$$\mathbb{P}\Big(\mid \delta\boldsymbol{\beta} \mid \in \Delta\delta \Big) \approx e^{-t\mathsf{d}(p|p')},$$

which implies that close-by parameters are associated to close-by probability measures [30].

Consider now two MEMCs $p(\pi, P)$ and $p'(\pi', P')$ specified by $\mathcal{H}_{\boldsymbol{\beta}}$ and $\mathcal{H}_{\boldsymbol{\beta'}}$, respectively with the same family of features. We say that the MEMCs are *ϵ-indistinguishable* if:

$$-\ln \mathbb{P}\Big(\mid \delta\boldsymbol{\beta} \mid \in \Delta\delta \Big) \leq \epsilon. \tag{29}$$

As both MEMCs satisfy the variational principle (Equation (14)), the relative entropy between p and p' (Equation (26)) reads:

$$\mathsf{d}(p \mid p') = \mathcal{P}[\mathcal{H}_{\beta'}] - \mathcal{P}[\mathcal{H}_{\beta}] + p(\mathcal{H}_{\beta}) - p(\mathcal{H}_{\beta'}). \tag{30}$$

Taking the Taylor expansion of $\mathsf{d}(p \mid p')$ around $\boldsymbol{\beta'} = \boldsymbol{\beta}$ we get:

$$\mathsf{d}(p \mid p') \approx \mathsf{d}(p \mid p) + \sum_k \frac{\partial \mathsf{d}(p \mid p')}{\partial \beta'_k}\Big|_{\beta'=\beta} (\beta_k - \beta'_k) + \frac{1}{2} \sum_{k,j} \frac{\partial^2 \mathsf{d}(p \mid p')}{\partial \beta'_k \beta'_j}\Big|_{\beta'=\beta} (\beta_k - \beta'_k)(\beta_j - \beta'_j).$$

Since $\mathsf{d}(p \mid p')$ is minimized at $\beta' = \beta$, we obtain,

$$\mathsf{d}(p \mid p') \approx \frac{1}{2} \sum_{k,j} \frac{\partial^2 \mathsf{d}(p \mid p')}{\partial \beta'_k \beta'_j}\Big|_{\beta'=\beta} (\beta_k - \beta'_k)(\beta_j - \beta'_j).$$

Taking the second derivative of $\mathsf{d}(p \mid p')$ from (Equation (30)), one also has that,

$$\frac{\partial^2 \mathsf{d}(p \mid p')}{\partial \beta'_k \beta'_j} = \frac{\partial^2 \mathcal{P}[\mathcal{H}_{\beta'}]}{\partial \beta'_k \beta'_j} = L_{kj}. \tag{31}$$

The second partial derivatives of the topological pressure with respect to the parameters β'_k and β'_j can be conveniently arranged in a matrix L with components L_{kj}. Given two MEMCs specified by $\mathcal{H}_{\boldsymbol{\beta}}$ and $\mathcal{H}_{\boldsymbol{\beta'}}$, in the limit of large t they are ϵ-indistinguishable if:

$$\frac{1}{2}\Big[(\delta\boldsymbol{\beta})^{\mathsf{T}} L (\delta\boldsymbol{\beta})\Big] \leq \frac{\epsilon}{T}, \tag{32}$$

where T denotes transpose. The matrix L can be obtained from data without need to fit the parameters. Equation (32) characterize a region in the space of MEMC of indistinguishable models, whose volume can be calculated in the large t limit using spectral properties of the matrix L [30]. This result generalizes a previous result for maximum entropy distributions for range one energy functions in [31].

5. Illustrative Examples

In this section, we illustrate the presented methods in some simple scenarios. In these examples, we follow a set of steps:

1. Choose the features and build the energy function (Equation (7)).
2. Build the transfer matrix (Equation (10)).
3. Compute the free energy and find the maximum entropy parameters using (Equation (17)).
4. Build the Markov transition matrix using (Equation (12)).
5. Choose the observable to examine and build the tilted transition matrix using Equation (22).
6. Compute the Legendre transform of the log maximum eigenvalue of the tilted transition matrix to obtain the rate function (Equation (24)).

For the sake of clarity, in this section, we focus on small neuronal networks. It is clear, however, that the extension of these techniques to larger neural populations is straightforward.

5.1. First Example: Maximum Entropy Model of a Range Two Feature

Consider spiking data from two interacting neurons. We measure only the average value of a of a range two feature from the spiking data to fit a MEMC. The feature denoted by f_1 is given by $\tilde{f}_1(\boldsymbol{x}_{0,1}) = x_0^2 \cdot x_1^1$, which detects when a spike of the second neuron is followed by a spike in the first one. The system can be described with the help of an energy function $\mathcal{H}(\boldsymbol{x}_{0,1}) = \beta_1 \tilde{f}_1(\boldsymbol{x}_{0,1})$.

For a given dataset of T spike blocks of range two, the empirical average reads,

$$A_T(f_1) = c \tag{33}$$

This means that in the data one finds that this event appears c% of the time.

The transfer matrix $\mathcal{L}_{\mathcal{H}}$ (cf. Equation (10)) is primitive by construction (cf. Equation (10)) and satisfies the hypothesis of the Perron–Frobenius theorem. In fact, its unique maximum eigenvalue is $\rho(\beta_1) = e^{\beta_1} + 3$. Given the restriction in Equation (33), using Equation (17) we obtain the following relationship between the parameter β_1 and the value of the restriction c:

$$\frac{\partial \mathcal{P}\left[\mathcal{H}\right]}{\partial \beta_1} = \frac{\partial \log(e^{\beta_1} + 3)}{\partial \beta_1} = \frac{e^{\beta_1}}{e^{\beta_1} + 3} = c.$$

This equation can be solved numerically. Using the obtained value of β_1 in Equation (12), one can find the corresponding Markov transition matrix. Note that, among all the Markov chains that match exactly the restriction, the selected one maximizes the KSE. Moreover, it is direct to check that the variational principle in Equation (14) is satisfied. Examples of values of β_1 for different values of c and IEP (25) for each value of β_1 are given in the following table:

Having the MEMC, we are now interested in analyzing the statistical fluctuations of the feature f_1. Using Equation (22), we obtain the tilted transition matrix $\widetilde{P}_{ij}^{(f_1)}(k)$ for each of the values in the Table 1. In Figure 4, we compute for each value of β_1 we compute the SCGF $\lambda_{f_1}(k)$ (24) and the Legendre transform (rate function) associated to the feature $I_{f_1}(s)$.

Table 1. Set of values used in Figures 4 and 5.

c	β_1	*IEP*
0.043	−2	0.176
0.11	−1	0.056
0.25	0	0
0.475	1	0.0525
0.711	2	0.1184

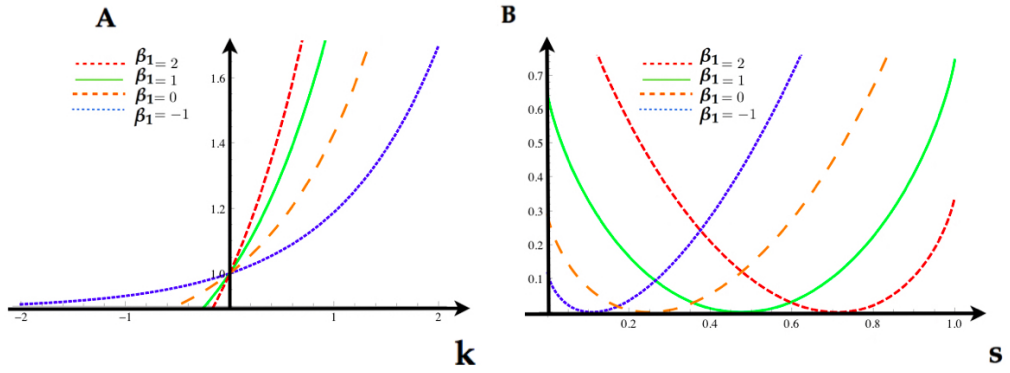

Figure 4. (**A**) Scaled cumulant generating function (SCGF) (Equation (24)) for the feature f_1 of the first example computed at the values provided by the table above. (**B**) Rate function for the same feature computed at the same parameter values as the SCGF. Each of this functions are obtained taking the Lagrange transform of the respective SCGF in (**A**).

In Figure 5, we compute for each value of IEP in the table the rate function and illustrate for this example the symmetry relationship (Equation (A9)).

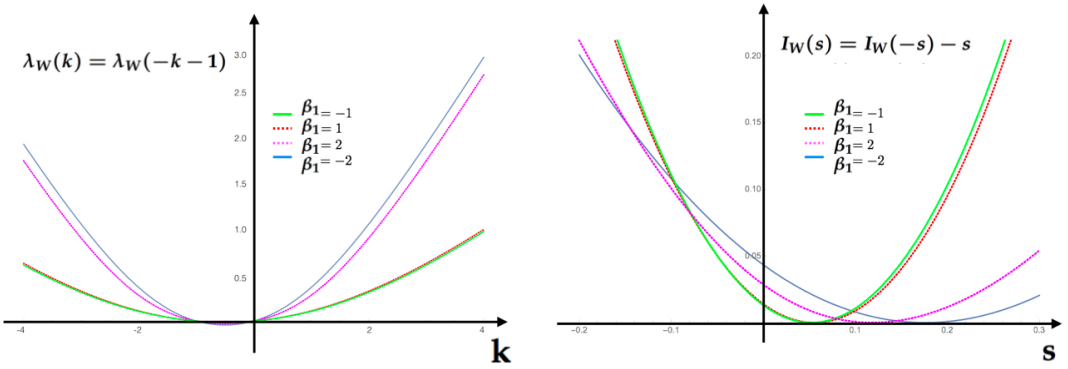

Figure 5. Gallavotti–Cohen symmetry relationship for the information entropy production IEP for values in Table 1. Left SCGF $\lambda_W(k)$. Right rate function of the IEP feature W, $I_W(s)$.

5.2. Second Example: Maximum Entropy Model With Only Synchronous Constraints

Let us now consider a network of three neurons. We focus here on range one features. In this example, we consider features related to the firing rates and synchronous pairwise correlations (Ising model [6,7]). Specifically, we consider the following energy function:

$$\mathcal{H}(\boldsymbol{x}_0) = \beta_1 x_0^1 + \beta_2 x_0^2 + \beta_3 x_0^3 + \beta_4 x_0^1 \cdot x_0^2 + \beta_5 x_0^1 \cdot x_0^3 + \beta_6 x_0^2 \cdot x_0^3,$$

with the six parameters $\beta_1, \ldots, \beta_6$. Following (Equation (10)), the transfer matrix $\mathcal{L}_{\mathcal{H}}$ indexed by the states of $\mathcal{A}_1^3$ is the following:

$$\mathcal{L}_{\mathcal{H}} = \begin{pmatrix} 1 & e^{\beta_1} & e^{\beta_2} & e^{\beta_1+\beta_2+\beta_4} & e^{\beta_3} & e^{\beta_1+\beta_3+\beta_5} & e^{\beta_2+\beta_3+\beta_6} & e^{\beta_1+\beta_2+\beta_3+\beta_4+\beta_5+\beta_6} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & e^{\beta_1} & e^{\beta_2} & e^{\beta_1+\beta_2+\beta_4} & e^{\beta_3} & e^{\beta_1+\beta_3+\beta_5} & e^{\beta_2+\beta_3+\beta_6} & e^{\beta_1+\beta_2+\beta_3+\beta_4+\beta_5+\beta_6} \end{pmatrix}.$$

This matrix is primitive, and the unique maximum eigenvalue is

$$\rho(\beta) = 1 + e^{\beta_1} + e^{\beta_2} + e^{\beta_1+\beta_2+\beta_4} + e^{\beta_3} + e^{\beta_1+\beta_3+\beta_5} + e^{\beta_2+\beta_3+\beta_6} + e^{\beta_1+\beta_2+\beta_3+\beta_4+\beta_5+\beta_6}.$$

The right eigenvector associated to this eigenvalue has all the components equal to 1. We obtain the topological pressure $\mathcal{P}[\mathcal{H}] = \log\rho(\beta)$. To find the MEMC parameters, we solve this set of equations:

$$\frac{\partial\mathcal{P}[\mathcal{H}]}{\partial\beta_1} = A_T(f_k) = c_k. \tag{34}$$

From Equation (34), provided some constraints on the average value of the features, we can solve the maximum entropy problem (see Table 2).

Table 2. Set of values used in Figure 6.

$A_T(f_k)$	c_k	β_k	$\delta\beta_k$	$\tilde{c}_k$
$A_T(x^1)$	0.3	−1.0436	0	0.30350016
$A_T(x^2)$	0.2	−1.6727	0	0.20127414
$A_T(x^3)$	0.1	−2.8163	0	0.10450018
$A_T(x^1x^2)$	0.08	0.4590	0	0.08187418
$A_T(x^1x^3)$	0.05	0.8604	0.1	0.05475019
$A_T(x^2x^3)$	0.04	1.0325	0	0.04207419

From Equation (12), one can find the Markov transition matrix. To compute the rate function of each feature in this model, we take the logarithm of the maximum eigenvalue of the tilted matrix, and obtain the tilted cumulant generating function $\lambda_f(k)$. In Figure 6, we illustrate the rate functions for each feature in the model.

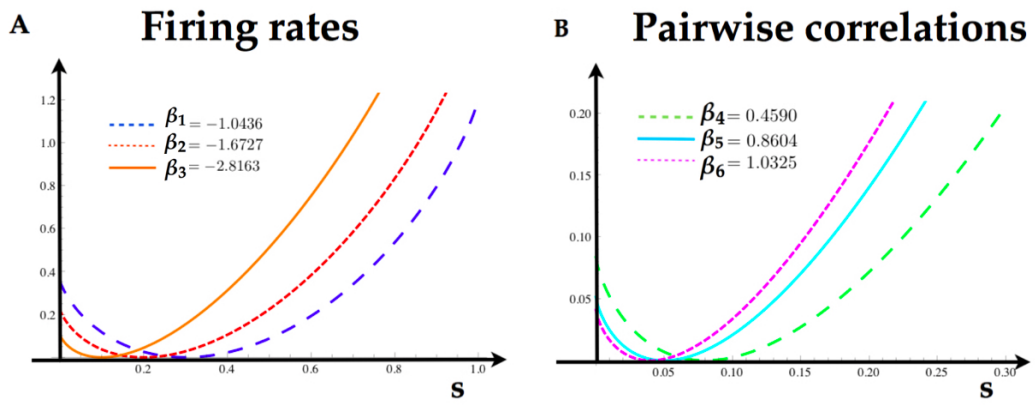

Figure 6. (**A**) Rate functions for the firing rates of each neuron of the Ising model. The minimum of the rate functions coincide with the expected value of the firing rates in Table 2. (**B**) Rate functions for the pairwise interactions computed from the parameters in Table 2.

5.3. Third Example: Past Independent and Markov Maximum Entropy Measures

To illustrate the difference between synchronous and non-synchronous maximum entropy models, we study a simple model composed of two interacting neurons:

$$\mathcal{H}(\boldsymbol{x}_{0,1}) = \beta_1 x_0^1 \cdot x_1^2 + \beta_2 x_0^2 \cdot x_1^1 + \beta_3 x_0^1 \cdot x_0^2. \tag{35}$$

We build a Markov chain by fixing the parameters of $\mathcal{H}$ at $\beta_1 = -3, \beta_2 = 3, \beta_3 = 0.5$ in the state space $\mathcal{A}_1^2$, given by

$$\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}.$$

whose corresponding transition matrix is given by

$$P = \begin{pmatrix} 0.13026 & 0.02580 & 0.65762 & 0.18632 \\ 0.65763 & 0.13026 & 0.16529 & 0.04682 \\ 0.02580 & 0.10266 & 0.13026 & 0.74128 \\ 0.15015 & 0.59735 & 0.03774 & 0.21476 \end{pmatrix}.$$

We focus on the synchronous feature $f = x_0^1 \cdot x_0^2$, whose average value with respect to the Markov measure p fixed by the parameters $\beta_1, \beta_2, \beta_3$ is $\mathbb{E}_p\{x_0^1 \cdot x_0^2\} = 0.292611$.

Using this particular Markov chain, we generate a sample of size $T = 20,000$. Then, we consider these data as a spike train of two neurons from which we have no other information. Starting from this data, we find the maximum entropy distribution that only considers the empirical average of the synchronous feature $A_T(x_0^1 \cdot x_0^2) = 0.2926$ as constraint. Therefore, we build a second model that uses the following energy function:

$$\tilde{\mathcal{H}}(\boldsymbol{x}_0) = \tilde{\beta} x_0^1 \cdot x_0^2. \tag{36}$$

Using the constraint, we obtain from Equation (8), $\tilde{\beta} = 0.215874$ fixing the maximum entropy distribution $\tilde{p}$. Note that by construction $\mathbb{E}_{\tilde{p}}\{x_0^1 \cdot x_0^2\} = A_T(x_0^1 \cdot x_0^2) = 0.2926$.

For both energy functions (Equations (35) and (36)) with the parameters mentioned before, we compute the rate functions of the synchronous feature. Additionally, from the sample of the Markov chain we compute the empirical averages of the synchronous feature using sliding windows of 50 samples. As expected, these empirical averages fluctuate around the overall average, as shown in Figure 7.

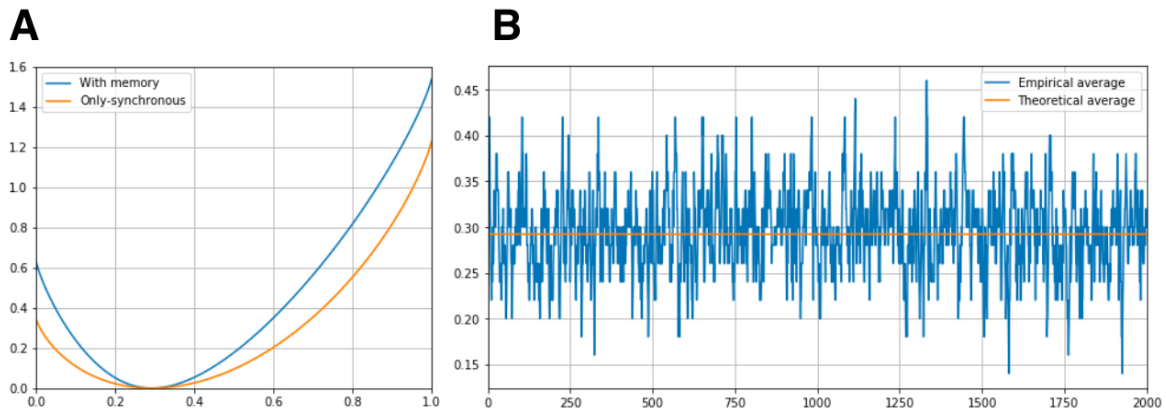

Figure 7. (**A**) Rate functions of the synchronous feature $x_0^1 x_0^2$ for both energy functions. (**B**) Moving averages computed from a sample of length 20,000 of the Markov Chain with transition matrix P.

To test the relevance of including memory into the model (and assess the performance of memoryless features), we compared the statistics of the fluctuations seen in the empirical average with the prediction by the rate functions of the two models. Figure 8 shows the histogram of empirical fluctuations, and plots the theoretical estimations of the fluctuations given by $K \exp\{-W I_f(s)\}$, where $W = 50$ is the window size, s is the fluctuation size, $I_f(s)$ is the rate function, and K is a constant that is adapted for visualization purposes.

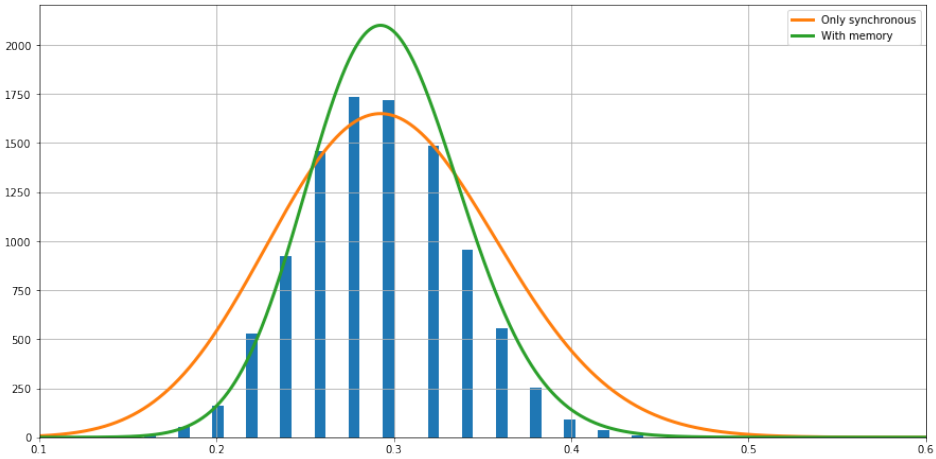

Figure 8. The fluctuations of the synchronous feature around the mean computed from the sample of the Markov chain are indicated with the bars. The Gaussian fluctuations around the mean predicted by the large deviations rate functions of both models are plotted. The curve predicted by the Markov model obtained from $\mathcal{H}$ is in green and the curve predicted by the model obtained from $\tilde{\mathcal{H}}$ is in orange.

Results show that the rate function of the model with memory fits accurately the relative frequencies of the empirical large fluctuations. In contrast, the model with no memory overestimates the fluctuation frequencies, having a much larger variance than the data, and underestimates near the expected value.

6. Conclusions

In the past few years, new experimental techniques combined with clever ideas from statistical mechanics have made it possible to infer maximum entropy models of spike trains directly from experimental recordings. However, a very important issue, namely quantifying the accuracy of the estimation obtained from a finite empirical sample, is usually ignored in this field. This is probably because the maximum entropy approach has a dual nature; one side is a convex optimization problem, which provides a unique solution independent of the sampling size, and on the other side is a Bayesian inference procedure, from which it is more natural to ask this question. As we have discussed in the Section 1 this characterization is relevant in the field of computational neuroscience as, in practice, experimental recordings are performed during a finite amount of time which causes fluctuations over the estimated quantities.

A fundamental goal of spike train analysis over networks of sensory neurons involves building accurate statistical models that predict the response of the network to a stimulus of interest. In particular, the aim of statistical inference of spiking neurons using the MEP, is that the fitted parameters shed light on some aspects of the neuronal code, therefore it is extremely important to quantify the accuracy of the statistical procedure. Additionally, one may be interested in measuring some properties of the inferred statistical model characterizing the spiking neuronal network, for example, the convergence rate of a sample or to quantify the probability of rare events of features such as firing rates, pairwise correlations, triplets or spatiotemporal events, mainly because these features are likely to play an important role in neuronal information processing. It is possible that rare and unlikely events have been generated by internal states of the neuronal tissue and not driven by the external stimulus. The events that are unlikely to occur deserve a better understanding as may carry important information about the network internal structure and may play a role in organizing a coherent dynamic to convey sensory information to the cerebral cortex.

The present contribution addressed this issue using tools from large deviations theory in the context of the MEMC. In particular, we showed that the transfer matrix technique used to build the MEMC is well adapted to compute large deviation rate functions using the Gärtner–Ellis theorem.

We also provide tools to investigate how sharply determined are the parameters of a MEMC with respect to the amount of empirical data using the concept of ϵ distinguishability. Additionally, we present a non-trivial relation between the distance in the parameter space and the distance in the manifold of maximum entropy probability measures using a LDP.

We have illustrated our method using simple examples. However, these examples might give a false impression that large deviations rate functions can always be calculated explicitly. In fact, exact and explicit expressions can be found only in small simple cases; fortunately, there exist numerical methods to evaluate rate functions [53].

Here, we have focused our attention on large deviations properties on maximum entropy models arising from spike train statistics, however, these results can be used in other fields of applications of maximum entropy models.

Author Contributions: All authors conceived the algorithm, and wrote and revised the manuscript. All authors have read and approved the final manuscript. We thank J.-P.E., A.P. and R.H. for helpful discussions.

Funding: R.C. was supported by an ERC advanced grant "Bridges", CONICYT-PAI Inserción 79160120 and Proyecto REDES ETAPA INICIAL, Convocatoria 2017 REDI170457. C.M. was at the early stage of this project, supported by the CONICYT-FONDECYT Postdoctoral Grant No. 3140572. F.R. acknowledges the support of the European Union's H2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 702981.

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

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.
$A_T(f)$	Empirical Average value of the feature f considering T spike patterns.
$\mathcal{A}_R^N$	Set of spike blocks of N neurons and length R.
$\mathcal{S}[p]$	Entropy of the probability measure p.
$\mathcal{H}$	Energy function.
$\mathcal{P}[\mathcal{H}]$	Free energy or topological pressure.
$\lambda_f(k)$	Scaled cumulant generating function of f.
$I_f(s)$	Rate function of f.

Appendix A. Discrete-Time Markov Chains and Spike Train Statistics

Consider the random process $\{X_n : n \geq 0\}$ taking values on $\mathcal{A}_R^N$. One can assume that the spiking activity of the neuronal network can be modeled by some discrete-time Markov process whose transition probabilities are obtained by means of the maximum entropy method described in Section 3. In this setting, $\mathcal{A}_R^N$ is the state space of the Markov chain, and, thus, if $X_n = x_{n,n+R-1}$, we say that the process is in the state $x_{n,n+R-1}$ at time n. The transition probabilities are given as follows,

$$\mathbb{P}\big[X_n = x_{(n)} \mid X_{n-1} = x_{(n-1)}, \ldots, X_0 = x_{(0)}\big] = \mathbb{P}\big[X_n = x_{(n)} \mid X_{n-1} = x_{(n-1)}\big], \tag{A1}$$

where we used the short hand notation $x_{(n)} := x_{n,n+R-1}$. We emphasize that in this paper the states are spike blocks of finite length R, $x_{n,n+R-1}$. All along this paper he have only considered homogeneous Markov chains, that is, Equation (A1) is independent of n.

Since transitions are considered between blocks of the form $x_{n-R,n-1} \to x_{n-R+1,n}$, the block $x_{n-R+1,n-1}$ must be common for the transition to be possible. Consider two spike blocks $i, j \in \mathcal{A}_R^N$ of range $R \geq 2$. We say that the transition from state i to state j is *allowed* if i and j have the common sub-block $x_{1,R-1} = \tilde{x}_{0,R-2}$, where $\tilde{x}_{0,R-2}$ are the first $R-1$ spike patterns of j.

Now, we define the transition matrix $P_R : \mathcal{A}_R^N \times \mathcal{A}_R^N \to \mathbb{R}$, whose entries are given by the transition probabilities, as follows,

$$(P_R)_{ij} := \begin{cases} \mathbb{P}[j \mid i] > 0 & \text{if } i \to j \text{ is allowed} \\ 0, & \text{otherwise.} \end{cases} \tag{A2}$$

Note that P has $2^{NR} \times 2^{NR}$ entries, but it is a sparse matrix since each line has, at most, 2^N non-zero entries. A stochastic matrix P is defined from transition probabilities in Equation (A2) satisfying:

$$\mathbb{P}[j \mid i] \geq 0; \qquad \sum_{j \in \mathcal{A}_R^N} \mathbb{P}[j \mid i] = 1,$$

for all states $i, j \in \mathcal{A}_R^N$. Moreover, by construction, for any pair of states, there exists a path of maximum length R in the graph of transition probabilities going from one to the other, which means that the Markov chain is primitive.

Appendix B. Cumulant Generating Function

In general, to obtain the scale cumulant generating function (as considered in Section 4.1.2), one has to deal with the moment of order r of a real-valued random variable f, which is given by,

$$m_r = \mathbb{E}(f^r),$$

for $r \in \mathbb{N}$. Provided that it has a Taylor expansion about the origin, the moment generating function (or Fourier-Laplace transform)

$$M(k) = \mathbb{E}(e^{kf}) = \mathbb{E}(1 + kf + \cdots + k^r f^r / r! + \cdots) = \sum_{r=0}^{\infty} m_r k^r / r!$$

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

$$\log M(k) = \sum_r \kappa_r k^r / r!$$

The relationship between the first moments and cumulants can be obtained by extracting coefficients from the expansion, as follows:

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

and so on. In particular, κ_1 is the mean of f , κ_2 is the variance, κ_3 the skewness and κ_4 the kurtosis.

Appendix C. Linear Response

Within the framework we have built, we can quantify how a small perturbation of the maximum entropy parameters (associated to given features) affects the average values of other features of the MEMC. It is important to quantify this perturbation because the maximum entropy parameters are obtained with finite accuracy due to finite sample effects. Fixing $\boldsymbol{\beta}$, we can obtain the average value of a given feature f_k with respect to the MEMC without need to sample, using the Gibbs–Jaynes principle for the KSE [29], which asserts that for a translation invariant probability measure p, the entropy rate $S_{KS}(p)$ is maximal under the constraints $\mathbb{E}_p\{f_k\} = c_k$, for all $k \in \{1, \ldots, K\}$ if and only if p is a Gibbs measure associated to the energy $\mathcal{H}_\beta = \sum \beta_k f_k$, where $\mathbb{E}_p\{f_k\} = \frac{\partial \mathcal{P}[\mathcal{H}_\beta]}{\partial \beta_k} = c_k$.

Now, let us consider a perturbed version of the energy denoted by $\mathcal{H}_{\boldsymbol{\beta}+\delta\boldsymbol{\beta}}$. Using a Taylor expansion, we compute the average value of an arbitrary feature here denoted by f_k with respect to the MEMC associated to the perturbed energy in terms of the unperturbed one, that is,

$$\mathbb{E}_{p_{\boldsymbol{\beta}+\delta\boldsymbol{\beta}}}\{f_k\} = \frac{\partial \mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}+\delta\boldsymbol{\beta}}\right]}{\beta_k} \tag{A3}$$

$$= \frac{\partial \mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]}{\beta_k} + \sum_j \frac{\partial^2 \mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]}{\partial \beta_k \beta_j} \delta\beta_j + O(\delta\beta_j)^2 \tag{A4}$$

$$= \mathbb{E}_{p_\beta}\{f_k\} + \sum_j \frac{\partial^2 \mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]}{\partial \beta_k \beta_j} \delta\beta_j + O(\delta\beta_j)^2 = \tilde{c}_k. \tag{A5}$$

From Equations (A3) and (A4), there is a Taylor expansion of $\mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}+\delta\boldsymbol{\beta}}\right]$ about $\mathcal{H}_{\boldsymbol{\beta}}$. From Equations (A4) and (A5), we use the Gibbs–Jaynes principle for the KSE. We see from Equation (A5) that a small perturbation of a parameter β_j influence the average value of all other features in the energy function (as f_k is arbitrary) and the magnitude of the perturbation is controlled by the second derivatives of the topological pressure of the unperturbed energy $\mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]$.

Appendix D. Time Correlations from Topological Pressure

For a pair of finite range features f_k, f_j of a stationary Markov chain, the covariance of order r is independent of time, only depends on the lag r and is defined as:

$$C_{f_k,f_j}(r) := \mathbb{E}_p\left\{f_k(x_n) f_j(x_{n+r})\right\} - \mathbb{E}_p\left\{f_k(x_n)\right\} \mathbb{E}_p\left\{f_j(x_n)\right\},$$

where $\mathbb{E}_p$ stands for the expected value with respect to the Markov measure p.

For MEMC with potentials of range $R > 1$, there is a positive time correlation correlation between pairs of features $f(x_n)$ and $g(x_{n+r})$, which we denote $\sigma^2_{f,g}$; indeed, one can show that (Green–Kubo formula):

$$\sigma^2_{f_k,f_j} = C_{f_k,f_j}(0) + \sum_{r=1}^{\infty} C_{f_k,f_j}(r) + \sum_{r=1}^{\infty} C_{f_j,f_k}(r). \tag{A6}$$

The pairwise time correlations between features can be obtained from the topological pressure:

$$\sigma^2_{f_k,f_j} = \frac{\partial^2 \mathcal{P}[\mathcal{H}_\beta]}{\partial \beta_k \, \partial \beta_j} = \frac{\partial \mu\left[f_j\right]}{\partial \beta_k}. \tag{A7}$$

for a MEMC taking values on $\mathcal{A}_R^N$ characterized by $p(\pi, P)$:

$$\frac{\partial^2 \mathcal{P}[\mathcal{H}_\beta]}{\partial \beta_k \, \partial \beta_j} = \mathbb{E}_p\{f_k f_j\} - \mathbb{E}_p\{f_k\}\mathbb{E}_p\{f_j\} + \sum_{r=1}^{\infty} \sum_{y,w \in \mathcal{A}_R^N} f_k(y) f_j(w) \pi_y P^r_{yw} + \sum_{r=1}^{\infty} \sum_{y,w \in \mathcal{A}_R^N} f_j(y) f_k(w) \pi_y P^r_{yw}$$

for $v, w \in \mathcal{A}_R^N$. For MEMC fitted through range one energy functions, $\{f(x_n); n \geq 0\}$ is an i.i.d. process and the variance of f is simply $C_f(0)$. These terms are the linear response coefficients. For MEMC associated to energy functions formed by K features, the matrix L can be conveniently arranged in a $K \times K$ symmetric matrix (known as the Onsager reciprocity relations [59]).

$$\sigma^2_{f_k,f_j} = \frac{\partial^2 \mathcal{P}[\mathcal{H}_\beta]}{\partial \beta_k \, \partial \beta_j} = \frac{\partial \mathbb{E}_p\{f_j\}}{\partial \beta_k}. \tag{A8}$$

Appendix E. Gallavotti–Cohen Fluctuation Theorem

The Gallavotti–Cohen fluctuation theorem refers to a universal property of the IEP, i.e., it is independent of the parameters of the MEMC. It is as a statement about properties of the SCGF and rate function of the IEP [57], this is,

$$\lambda_W(k) = \lambda_W(-k-1), \quad I_W(s) = I_W(-s) - s. \tag{A9}$$

This symmetry can be seen as a generalization of Kubo formula in Equation (A6) and Onsager reciprocity relations in Equation (A8) to situations far from equilibrium. It is a relationship that holds for a general class fs stochastic processes including Markov chains [55].

These properties have an impact on the large deviations of the time-averaged entropy production rate of the sample trajectory $\boldsymbol{x}_{0,t-1}$ of the Markov chain $p(\pi, P)$ denoted $\frac{W_t}{t}$. In our framework, the following relationship always 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.

References

1. Rieke, F.; Warland, D.; de Ruyter van Steveninck, R.; Bialek, W. *Spikes, Exploring the Neural Code*; MIT Press: Cambridge, MA, USA, 1996.
2. Friston, K.J. Functional and effective connectivity: A review. *Brain Connect.* **2011**, *1*, 13–36. [CrossRef] [PubMed]
3. Okatan, M.; Wilson, M.A.; Brown, E.N. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural Comput.* **2005**, *17*, 1927–1961. [CrossRef] [PubMed]
4. Ganmor, E.; Segev, R.; Schneidman, E. The architecture of functional interaction networks in the retina. *J. Neurosci.* **2011**, *31*, 3044–3054. [CrossRef] [PubMed]
5. Ferrea, E.; Maccione, A.; Medrihan, L.; Nieus, T.; Ghezzi, D.; Baldelli, P.; Benfenati, F.; Berdondini, L. Large-scale, high-resolution electrophysiological imaging of field potentials in brain slices with microelectronic multielectrode arrays. *Front. Neural Circuits* **2012**, *6*, 80. [CrossRef] [PubMed]
6. Schneidman, E.; Berry, M.J., II; Segev, R.; Bialek, W. Weak pairwise correlations imply string correlated network states in a neural population. *Nature* **2006**, *440*, 1007–1012. [CrossRef] [PubMed]
7. Tkačik, G.; Marre, O.; Mora, T.; Amodei, D.; Berry II, M.J.; Bialek, W. The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech.* **2013**, *2013*, P03011. [CrossRef]
8. Vasquez, J.C.; Marre, O.; Palacios, G.A.; Berry II, M.J.; 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]
9. 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] [PubMed]
10. Croner, L.J.; Purpura, K.; Kaplan, E. Response variability in retinal ganglion cells of primates. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 8128–8130. [CrossRef] [PubMed]

11. Shadlen, M.N.; Newsome, W.T. The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *J. Neurosci.* **1998**, *18*, 3870–3896. [CrossRef] [PubMed]
12. Pillow, J.W.; Shlens, J.; Paninski, L.; Sher, A.; Litke, A.M.; Chichilnisky, E.J.; Simoncelli, E.P. Spatio-temporal correlations and visual signaling in a complete neuronal population. *Nature* **2008**, *454*, 995–999. [CrossRef] [PubMed]
13. Tkačik, G.; Marre, O.; Amodei, D.; Schneidman, E.; Bialek, W.; Berry, M.J., II. Searching for collective behavior in a large network of sensory neurons. *PLoS Comput. Biol.* **2013**, *10*, e1003408. [CrossRef] [PubMed]
14. Tkačik, G.; Mora, T.; Marre, O.; Amodei, D.; Berry II, M.; Bialek, W. Thermodynamics for a network of neurons: Signatures of criticality. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 11508–11513. [CrossRef] [PubMed]
15. 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]
16. Schrödinger, E. *What is Life? the Physical Aspect of the Living Cell*; Cambridge University Press: Cambridge, UK, 1983.
17. Deem, M. Mathematical adventures in biology. *Phys. Today* **2007**, *60*, 42–47. [CrossRef]
18. Prigogine, I. *Nonequilibrium Statistical Mechanics*; Monographs in Statistical Physics; Interscience Publishers: Geneva, Switzerland, 1962; Volume 1.
19. Filyukov, A.A.; Karpov, V.Y. Description of steady transport processes by the method of the most probable path of evolution. *J. Eng. Phys.* **1967**, *13*, 624–630. [CrossRef]
20. Filyukov, A.A.; Karpov, V.Y. Method of the most probable path of evolution in the theory of stationary irreversible processes. *J. Eng. Phys. Thermophys.* **1967**, *13*, 416–419. [CrossRef]
21. Favretti, M. The maximum entropy rate description of a thermodynamic system in a stationary non-equilibrium state. *Entropy* **2009**, *4*, 675–687. [CrossRef]
22. 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]
23. Shi, P.; Qian, H. Chapter Irreversible Stochastic Processes, Coupled Diffusions and Systems Biochemistry. In *Frontiers in Computational and Systems Biology*; Feng, J., Fu, W., Sun, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 175–201.
24. Cofré, R.; Cessac, B. Exact computation of the maximum entropy potential of spiking neural networks models. *Phys. Rev. E* **2014**, *89*, 052117. [CrossRef] [PubMed]
25. Cofré, R.; Maldonado, C. Information entropy production of maximum entropy markov chains from spike trains. *Entropy* **2018**, *20*, 34. [CrossRef]
26. Mora, T.; Bialek, W. Are biological systems poised at criticality? *J. Stat. Phys.* **2011**, *144*, 268–302. [CrossRef]
27. Ellis, R. *Entropy, Large Deviations and Statistical Mechanics*; Springer: Berlin/Heidelberg, Germany, 1985.
28. Dembo, A.; Zeitouni, O. *Large Deviations Techniques and Applications*; Springer: Berlin/Heidelberg, Germany, 2010.
29. Georgii, H. *Probabilistic Aspects of Entropy*; Greven, A., Keller, G., Warnecke, G., Eds.; Princeton University Press: Princeton, NJ, USA, 2003.
30. Balasubramanian, V. Statistical inference, Occam's Razor, and statistical mechanics on the space of probability distributions. *Neural Comput.* **1997**, *9*, 349–368. [CrossRef]
31. Mastromatteo, I.; Marsili, M. On the criticality of inferred models. *J. Stat. Mech.* **2011**, *2011*, P10012. [CrossRef]
32. Macke, J.; Murray, I.; Latham, P. Estimation bias in maximum entropy models. *Entropy* **2013**, *15*, 3109–3129. [CrossRef]
33. Marsili, M.; Mastromatteo, I.; Roudi, Y. On sampling and modeling complex systems. *J. Stat. Mech.* **2013**, *2013*, P09003
34. Quiroga, R.Q.; Nadasdy, Z.; Ben-Shaul, Y. Unsupervised spike sorting with wavelets and superparamagnetic clustering. *Neural Comput.* **2004**, *16*, 1661–1678. [CrossRef] [PubMed]
35. Hill, D.N.; Mehta, S.B.; Kleinfeld, D. Quality metrics to accompany spike sorting of extracellular signals. *J. Neurosci.* **2011**, *31*, 8699–8705. [CrossRef] [PubMed]
36. Gerstner, W.; Kistler, W. *Spiking Neuron Models*; Cambridge University Press: Cambridge, UK, 2002.

37. Schwalger, T.; Fisch, K.; Benda, J.; Lindner, B. How noisy adaptation of neurons shapes interspike interval histograms and correlations. *PLoS Comput. Biol.* **2010**, *6*, e1001026. [CrossRef] [PubMed]
38. Linaro, D.; Storace, M.; Giugliano, M. Accurate and fast simulation of channel noise in conductance-based model neurons by diffusion approximation. *PLoS Comput. Biol.* **2011**, *7*, e1001102. [CrossRef] [PubMed]
39. Cofré, R.; Cessac, B. Dynamics and spike trains statistics in conductance-based Integrate-and-Fire neural networks with chemical and electric synapses. *Chaos Soliton. Fract.* **2013**, *50*, 13–31. [CrossRef]
40. Jaynes, E. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
41. Jaynes, E. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
42. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern. Anal. Mach.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
43. Brown, G.; Pocock, A.; Zhao, M.; Lujn, M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **2012**, *13*, 27–66.
44. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2006.
45. Nasser, H.; Cessac, B. Parameter estimation for spatio-temporal maximum entropy distributions: Application to neural spike trains. *Entropy* **2014**, *16*, 2244–2277. [CrossRef]
46. Tkačik, G.; Prentice, J.S.; Balasubramanian, V.; Schneidman, E. Optimal population coding by noisy spiking neurons. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 14419–14424. [CrossRef] [PubMed]
47. Sinai, Y.G. Gibbs measures in ergodic theory. *Russ. Math. Surv.* **1972**, *27*, 21–69. [CrossRef]
48. Chliamovitch, G.; Dupuis, A.; Chopard, B. Maximum entropy rate reconstruction of markov dynamics. *Entropy* **2015**, *17*, 3738–3751. [CrossRef]
49. Seneta, E. *Non-Negative Matrices and Markov Chains*; Springer: Berlin/Heidelberg, Germany, 2006.
50. Bowen, R. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*; Second Revised Version; Springer: Berlin/Heidelberg, Germany, 2008.
51. Levin, D.A.; Peres, Y.; Wilmer, E.L. *Markov Chains and Mixing Times*; American Mathematical Society: Providence, RI, USA, 2009.
52. Jones, G.L. On the Markov chain central limit theorem. *Probab. Surv.* **2004**, *1*, 299–320. [CrossRef]
53. Touchette, H. The large deviation approach to statistical mechanics. *Phys. Rep.* **2009**, *478*, 1–69. [CrossRef]
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. Maes, C. The fluctuation theorem as a Gibbs property. *J. Stat. Phys.* **1999**, *95*, 367–392. [CrossRef]
56. Gaspard, P. Time-reversed dynamical entropy and irreversibility in Markovian random processes. *J. Statist. Phys.* **2004**, *117*, 599–615. [CrossRef]
57. Jiang, D.Q.; Qian, M.; Qian, M.P. *Mathematical Theory of Nonequilibrium Steady States*; Springer: Berlin/Heidelberg, Germany, 2004.
58. Amari, S. Information Geometry of Multiple Spike Trains. In *Analysis of Parallel Spike Trains*; Springer Series in Computational Neuroscience; Grün, S., Rotter, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 7, pp. 221–253.
59. Gaspard, P. Random paths and current fluctuations in nonequilibrium statistical mechanics. *J. Math. Phys.* **2014**, *55*, 075208. [CrossRef]

© 2018 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/).