Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Spatio-temporal spike train analysis for large scale networks using the maximum entropy principle and Monte Carlo method

Hassan Nasser[1], **Olivier Marre**[2] **and Bruno Cessac**[1]

[1] NeuroMathComp, INRIA, 2004 Route des Lucioles, F-06902 Sophia-Antipolis, France
[2] Institut de la Vision, 17 rue Moreau, F-75012 Paris, France
E-mail: hass.nasser@gmail.com, olivier.marre@gmail.com and bruno.cessac@inria.fr

Received 25 May 2012
Accepted 16 September 2012
Published 12 March 2013

Online at stacks.iop.org/JSTAT/2013/P03006
doi:10.1088/1742-5468/2013/03/P03006

Abstract. Understanding the dynamics of neural networks is a major challenge in experimental neuroscience. For that purpose, a modelling of the recorded activity that reproduces the main statistics of the data is required. In the first part, we present a review on recent results dealing with spike train statistics analysis using maximum entropy models (MaxEnt). Most of these studies have focused on modelling synchronous spike patterns, leaving aside the temporal dynamics of the neural activity. However, the maximum entropy principle can be generalized to the temporal case, leading to Markovian models where memory effects and time correlations in the dynamics are properly taken into account. In the second part, we present a new method based on Monte Carlo sampling which is suited for the fitting of large-scale spatio-temporal MaxEnt models. The formalism and the tools presented here will be essential to fit MaxEnt spatio-temporal models to large neural ensembles.

Keywords: sequence analysis (theory), neural code, computational neuroscience, statistical inference

1742-5468/13/P03006+30$33.00

Contents

doi:10.1088/1742-5468/2013/03/P03006

J. Stat. Mech. (2013) P03006

1. Introduction

The structure of the cortical activity and its relevance to sensory stimuli or motor planning have been the subject of long standing debate. While some studies tend to demonstrate that the majority of the information conveyed by neurons is contained in the mean firing rate [58], other works have shown evidence of the role of the higher-order neural assemblies in neural coding [64, 76, 1, 32, 23].

Many single-cell studies have reported an irregular spiking activity which seems to be very close to a Poisson process; concluding that the activity spans a very large state space. Several studies claim that some specific patterns, called 'cortical songs', appear in a recurrent fashion [26], but their existence is controversial [38, 33], suggesting that the size of the state space explored by the activity could be smaller than expected. This point requires an accurate description of the neural activity of populations of neurons [65, 41, 31, 75, 30].

These controversies partially originate from the fact that characterizing the statistics of the neural activity observed during the simultaneous recording of several neurons is challenging, since the number of possible patterns grows exponentially with the number of neurons. As a consequence, the probability of each pattern cannot be reliably measured by empirical averaging, and an underlying model is necessary to reduce the number of variables to be estimated. To infer the whole state of the neural network, some attempts have been done to build a hidden dynamical model which would underlie the cortical responses of several recorded neurons. Most of the time, this approach has been used to characterize the activity of neurons during different types of behaviour. Among others, Shenoy and colleagues [59] used a dynamical system to model the activities of multiple neurons recorded in the motor areas. Most of the time, in this approach, the number of neurons greatly exceeds the number of parameters. The assumed low dimension of the underlying dynamical system is often due to the low dimension of the behavioural context itself. For example, in a task where a monkey is asked to make a choice between a small number of options (e.g. moving towards one target amongst several), one can expect that the features of the neural activity which are relevant to this task can be described using a number of parameters which is comparable to the number of possible actions.

For more complex tasks or stimuli, the dimension of these models may have to be increased. This would be especially critical in the case of sensory networks stimulated with natural or complex stimuli. For this latter issue, a different strategy has been proposed by Schneidman *et al* [55] and Shlens *et al* [61, 62]. Their purpose was to describe the statistics of the retinal activity in response to natural stimuli. They defined a set of values (mean firing rates, correlations ...) that must be fitted, and then picked the *least structured* of the models that would satisfy these constraints. This approach, which will be described below, is based on maximum entropy models of the activity. It is interesting to point out that, while the previous approach aims at finding a useful representation of the activity with the lowest dimension, the maximum entropy approach picks the model with the highest dimension.

In this paper, we first describe the challenge of modelling the statistics of the neural activity, and review the results that were obtained using maximum entropy models. Many studies focused on modelling the synchronous patterns, putting aside the issue of modelling the temporal dynamics of the neural activity. We show why the extension

J. Stat. Mech. (2013) P03006

of maximum entropy models to the temporal case raises specific issues, such as treating correctly memory and time correlations, and how they can be solved. Section 2 reviews the maximum entropy approach, and focuses on applying it to general spatio-temporal constraints. We also include a short discussion on other spatio-temporal approaches to spike train statistics such as the generalized linear model [5, 36, 43, 74, 48, 46, 3, 47]. In section 3 we present a new method, based on Monte Carlo sampling, which is suited for the fitting of large-scale spatio-temporal models. In section 4 we provide examples and numerical tests in order to show how far we can go with the Monte Carlo method and how it performs.

2. The maximum entropy principle

In this section, we present the maximum entropy principle in a general setting. We first give a set of notations and definitions, then present a brief history of this principle in spike train analysis. Finally, we introduce a framework which allows the handling of general spatio-temporal constraints.

2.1. Notations and definitions

2.1.1. Spike trains. We consider a network of N neurons. We assume there is a minimal timescale δ, such that a neuron can fire at most one spike within a time window of size δ. To each neuron k and discrete time n, we associate a spike variable: $\omega_k(n) = 1$ if neuron k fires at time n, and $\omega_k(n) = 0$ otherwise. The state of the entire network in time bin n is thus described by a vector $\omega(n) \stackrel{\text{def}}{=} [\,\omega_k(n)\,]_{k=1}^{N}$, called a *spiking pattern*.

A *spike block*, which describes the activity of the whole network between moments of time n_1 and n_2, is a finite ordered list of such vectors, written as:

$$\omega_{n_1}^{n_2} = \{\,\omega(n)\,\}_{\{n_1 \le n \le n_2\}}\,.$$

The *range* of a block is $n_2 - n_1 + 1$, the number of time steps from n_1 to n_2. Here is an example of a spike block of range 5 with $N = 4$ neurons.

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

A *spike train* or *raster* is a spike block ω_0^{T-1} from some initial time 0 to some final time $T-1$. To simplify notation we simply write ω for a spike train. We note $\Omega = \{\,0,1\,\}^{NT}$ for the set of all possible spike trains.

2.1.2. Observables. We call an *observable* a function:

$$\mathcal{O}(\omega) = \prod_{u=1}^{r} \omega_{k_u}(n_u), \tag{1}$$

i.e. a product of binary spike events where k_u is a neuron index and n_u a time index, with $u = 1, \ldots, r$, for some integer $r > 0$. Typical choices of observables are $\omega_{k_1}(n_1)$, which

J. Stat. Mech. (2013) P03006

is 1 if neuron k_1 fires at time n_1 and which is 0 otherwise; $\omega_{k_1}(n_1)\,\omega_{k_2}(n_2)$, which is 1 if neuron k_1 fires at time n_1 and neuron k_2 fires at time n_2 and which is 0 otherwise. Another example is $\omega_{k_1}(n_1)\,(1-\omega_{k_2}(n_2))$, which is 1 if neuron k_1 fires at time n_1 *and neuron* k_2 *is silent at time* n_2. This example emphasizes that observables are able to consider events where some neurons are silent.

We say that an observable $\mathcal{O}$ has *range* R if it depends on R consecutive spike patterns, e.g. $\mathcal{O}(\omega)=\mathcal{O}(\omega_0^{R-1})$. We consider here that observables do not depend explicitly on time (*time-translation invariance of observables*). As a consequence, for any time n, $\mathcal{O}(\omega_0^{R-1})=\mathcal{O}(\omega_n^{n+R-1})$ whenever $\omega_0^{R-1}=\omega_n^{n+R-1}$.

2.1.3. Spike train statistics. It is common in the study of spike trains to attempt to detect some statistical regularity. Spike train statistics is assumed to be summarized by a hidden probability μ characterizing the probability of *spatio-temporal* spike patterns: μ is defined as soon as the probability $\mu\left[\,\omega_{n_1}^{n_2}\,\right]$ of any block $\omega_{n_1}^{n_2}$ is known. We assume that μ is time-translation invariant: for any time n, $\mu\left[\,\omega_0^{R-1}\,\right]=\mu\left[\,\omega_n^{n+R-1}\,\right]$, whenever $\omega_0^{R-1}=\omega_n^{n+R-1}$.

Equivalently, μ allows the computation of the average of the observables. We denote $\mu\left[\,\mathcal{O}\,\right]$ as the average of the observable $\mathcal{O}$ under μ. If $\mathcal{O}(\omega)=\omega_{k_1}(n_1)$, then $\mu\left[\,\mathcal{O}\,\right]$ is the firing rate of neuron k_1 (it does not depend on n_1 from the time-translation invariance hypothesis); if $\mathcal{O}=\omega_{k_1}(n_1)\,\omega_{k_2}(n_2)$, then $\mu\left[\,\mathcal{O}\,\right]$ is the probability that neurons k_1 and k_2 fire during the time span n_2-n_1. Additionally, $\mu\left[\,\omega_{k_1}(0)\,\omega_{k_2}(0)\,\right]-\mu\left[\,\omega_{k_1}(0)\,\right]\mu\left[\,\omega_{k_2}(0)\,\right]$ represents the instantaneous pairwise correlation between the neurons k_1 and k_2.

There are several methods which allow the computation or estimation of μ. In the following we shall assume that neural activity is described by a Markov process with memory depth D and positive time-translation invariant transition probabilities $P\left[\,\omega(D)\,\middle|\,\omega_0^{D-1}\,\right]>0$. From the assumption $P\left[\,\omega(D)\,\middle|\,\omega_0^{D-1}\,\right]>0$, this chain has a unique invariant probability μ such that, for any $n>D$, and any block ω_0^n:

$$\mu\left[\,\omega_0^n\,\right]=\prod_{l=0}^{n-D}P\left[\,\omega(D+l)\,\middle|\,\omega_l^{D+l-1}\,\right]\mu\left[\,\omega_0^{D-1}\,\right]. \tag{2}$$

Therefore, knowing the transition probabilities (corresponding to blocks ω_0^D of range $D+1$) and μ (which can be determined as well from the transition probabilities as exposed in section 2.3.1), the probability of larger blocks can be computed. Equation (2) makes explicit the role of memory in statistics of spike blocks, via the product of transition probabilities and the probability of the initial block $\mu\left[\,\omega_0^{D-1}\,\right]$.

In contrast, if $D=0$, the probability to have the spike pattern $\omega(D)$ does not depend on the past activity of the network (memory-less case). In this case $P\left[\,\omega(D+l)\,\middle|\,\omega_l^{D+l-1}\,\right]$ becomes $\mu\left[\,\omega(l)\,\right]$ and the probability (2) of a block becomes:

$$\mu\left[\,\omega_0^n\,\right]=\prod_{l=0}^{n}\mu\left[\,\omega(l)\,\right]. \tag{3}$$

Therefore, in the memory-less case, spikes occurring at different times are independent. This emphasizes the deep difference between the case $D=0$ and the case $D>0$.

2.1.4. Empirical average. Let us assume that we are given an experimental raster of length T, such that ω_0^{T-1}. The estimation of spike statistics has to be done on this

J. Stat. Mech. (2013) P03006

sample. In the context of the maximum entropy principle, where statistics is assumed to be time-translation invariant, statistics of events is obtained via the *time-average*. The time-average or empirical average of an observable $\mathcal{O}$ in a raster ω of length T is denoted by $\pi_\omega^{(T)}[\mathcal{O}]$. For example, if $\mathcal{O} = \omega_k(0)$ the time-average $\pi_\omega^{(T)}[\mathcal{O}] = 1/(T-1)\sum_{n=0}^{T-1}\omega_k(n)$ is the firing rate of neuron k, estimated on the experimental raster ω.

The empirical average is a random variable, depending on the raster ω as well as on the time length of the sample, and it has Gaussian fluctuations whose amplitude tends to 0 as $T \to +\infty$ like $1/\sqrt{T}$. This is the case, e.g. for the empirical averages obtained from several spike train acquired with several repetitions.

2.1.5. Complexity of the set of spike blocks. If one has N neurons and wants to consider spike block events within R time steps, one has 2^{NR} possible states. For a reasonable multielectrodes array (MEA) sample, $N = 100$, $R = 3$ (for a time lag of 30 ms with a 10 ms binning), this gives $2^{300} \sim 4 \times 10^{180}$, which is considerably more than the expected number of particles in the (visible) universe. Taking into account the huge number of states in the set of blocks, it is clear that any method requiring the extensive description of the state space will fail as NR grows. Additionally, while the accessible state space is huge, the *observed* state space (e.g. in an experimental raster) is rather small. For example, in a MEA raster for a retina experiment, the sample size T is about 10^6–10^7, which is quite a bit less than 2^{NR}. As a matter of fact, any reasonable estimation method must take this small-sample constraint into account. As we show in section 2.2, the maximum entropy principle and the related notion of Gibbs distributions allows us to take these aspects into account.

2.2. The maximum entropy principle

2.2.1. Motivations. Following (2.1) the goal is to find a probability distribution μ such that:

- μ is inferred from an empirical raster ω, by computing the empirical average of a set of ad hoc observables $\mathcal{O}_k$, $k = 1, \ldots, K$. One asks that the average of $\mathcal{O}_k$ with respect to μ satisfies:

$$\mu[\mathcal{O}_k] = \pi_\omega^{(T)}[\mathcal{O}_k], \qquad k = 1, \ldots, K. \tag{4}$$

  The mean of $\mathcal{O}_k$ predicted by μ is equal to the mean computed on the experimental raster. μ is called a 'model' in the following. The set of observables $\mathcal{O}_k$ defines the model.
- μ has to be 'as simple as possible', with the least structure and a minimum number of tunable parameters. In the maximum entropy paradigm [27] these issue are (partly) solved by seeking a probability distribution μ which maximizes the entropy under the constraints (4). The entropy is defined explicitly below (see equations (5) and (21)).
- From the knowledge of μ one can compute the probability of arbitrary blocks (e.g. via equation (2)) and the average of observables other than the $\mathcal{O}_k$.

Remark. Assume that we want to select observables $\mathcal{O}_k$ in the set of all possible observables with range R. For N neurons there are 2^{NR} possible choices. When NR increases, the number of possible observables will quickly exceed the number of samples

J. Stat. Mech. (2013) P03006

available in the recording. Including all of them in the model would overfit the data. Therefore, one has to guide the choice of observables by additional criteria. We now review some of the criteria which have been used by other authors.

2.2.2. Spatial models. In a seminal paper, Schneidman *et al* [55] aimed at unravelling the role of instantaneous pairwise correlations in retina spike trains. Although these correlations are weak, researchers investigated whether they play a more significant role in spike train statistics than firing rates.

Firing rates correspond to the average of observables of the form $\omega_i(0), i = 1, \ldots, N$ (the time index 0 comes from the assumed time-translation invariance) while instantaneous pairwise correlations correspond to averages of observables of the form $\omega_i(0)\omega_j(0), 1 \leq i < j \leq N$. Analysing the role of pairwise correlations in spike train statistics, compared to the firing rate, amounts therefore to comparing two models, defined by two different types of observables.

Note that all of these observables correspond to spatial events occurring at the same time. They give no information on how the spike patterns at a given time depend on the past activity. This situation corresponds to a memory-less model ($D = 0$ in section 2.1.3), where transition probabilities do not depend on the past. As a consequence the sought probability μ weights blocks of range 1, and the probability of blocks with larger range is given by (3): spike patterns at successive time steps are independent in spatial models.

In this case, the *entropy* of μ is given by:

$$S(\mu) = -\sum\nolimits_{\omega(0)} \mu\,[\,\omega(0)\,] \log \mu\,[\,\omega(0)\,]\,. \tag{5}$$

The natural log could be replaced by the logarithm in base 2.

Now, the maximum entropy principle of Jaynes [27] corresponds to seeking a probability μ which maximizes $S(\mu)$ under the constraints (4). It can be shown (see section 2.3 for the general statement) that this maximization problem is equivalent to the following Lagrange problem: maximizing the quantity $S(\mu) + \mu\,[\,\mathcal{H}_{\boldsymbol{\beta}}\,]$, where $\mathcal{H}_{\boldsymbol{\beta}}$, called a *potential*, is given by:

$$\mathcal{H}_{\boldsymbol{\beta}} = \sum_{k=1}^{K} \beta_k \mathcal{O}_k. \tag{6}$$

The β_k are real numbers and free parameters. $\mu\,[\,\mathcal{H}_{\boldsymbol{\beta}}\,]$ is the average of $\mathcal{H}_{\boldsymbol{\beta}}$ with respect to μ. Since $\mathcal{H}_{\boldsymbol{\beta}}$ is a linear combination of observables we have $\mu\,[\,\mathcal{H}_{\boldsymbol{\beta}}\,] = \sum_{k=1}^{K} \beta_k \mu\,[\,\mathcal{O}_k\,]$. If the $\mathcal{O}_k$ have finite range and are $> -\infty$ and if the β_k are finite then it can be shown (see section 2.3) that there is only one probability μ, depending on the β_k, which solves the maximization problem. It is called a *Gibbs distribution*.

In this context ($D = 0$) it reads:

$$\mu\,[\,\omega(0)\,] = \frac{\mathrm{e}^{\mathcal{H}_{\boldsymbol{\beta}}(\omega(0))}}{Z_{\boldsymbol{\beta}}}, \tag{7}$$

where the normalization factor

$$Z_{\boldsymbol{\beta}} = \sum\nolimits_{\omega(0)} \mathrm{e}^{\mathcal{H}_{\boldsymbol{\beta}}(\omega(0))} \tag{8}$$

is the so-called *partition function*.

J. Stat. Mech. (2013) P03006

To match (4) the parameters β_k have to be tuned. This can be done thanks to the following property of $Z_{\boldsymbol{\beta}}$:

$$\mu\left[\mathcal{O}_k\right] = \frac{\partial \log Z_{\boldsymbol{\beta}}}{\partial \beta_k}. \tag{9}$$

Thus the β_k have to be tuned so that μ matches (4) as well as (9). It turns out that $\log Z_{\boldsymbol{\beta}}$ is convex with respect to the β_k, so the problem has a unique solution.

Note that $\log Z_{\boldsymbol{\beta}}$ does not only allow us to obtain the averages of the observables, it also allows us to characterize fluctuations. If a raster is distributed according to the Gibbs distribution (7), then, as pointed out in section 2.1.4, the empirical average of an observable has fluctuations. One can show that these fluctuations are Gaussian (central limit theorem). The joint probability of $\pi_\omega^{(T)}\left[\mathcal{O}_k\right]$, $k = 1, \ldots, K$ is Gaussian, with mean $\mu\left[\mathcal{O}_k\right]$ given by (9) and covariance matrix Σ/T, where the matrix Σ has entries:

$$\Sigma_{kl} = \frac{\partial^2 \log Z_{\boldsymbol{\beta}}}{\partial \beta_k \partial \beta_l}. \tag{10}$$

Let us now discuss what this principle gives in the two cases considered by Schneidman *et al*

(i) Only firing rates are constrained. Then:

$$\mathcal{H}_{\boldsymbol{\beta}}(\omega(0)) = \sum_{k=1}^{N} \beta_k \omega_k(0).$$

It can be shown that the corresponding probability μ is:

$$\mu\left[\omega(0)\right] = \prod_{k=1}^{N} \frac{\mathrm{e}^{\beta_k\, \omega_k(0)}}{1 + \mathrm{e}^{\beta_k}}.$$

Thus, the corresponding statistical model is such that spikes emitted by distinct neurons at the same time are independent. The parameter β_k is directly related to the firing rate r_k since $r_k = \mu\left[\omega_k(0) = 1\right] = \mathrm{e}^{\beta_k}/(1 + \mathrm{e}^{\beta_k})$, so that we have:

$$\mu\left[\omega_0^n\right] = \prod_{l=0}^{n} \prod_{k=1}^{N} r_k^{\omega_k(l)} \left(1 - r_k\right)^{1-\omega_k(l)},$$

the classical probability of coin tossing with independent probabilities (*Bernoulli model*). Thus, fixing only the rates as constraints, the maximum entropy principle leads us to analyse spike statistics as if each spike were thrown randomly and independently, as with coin tossing. This is the 'most random model', which has the advantage of making as few hypotheses as possible. However, when only constrained with mean firing rates, the prediction of even small spike blocks in the retina was not successful. This was expected since this model assumes independence between neurons, an assumption that has been proven wrong in earlier studies (e.g. [50]).

(ii) Firing rates and pairwise correlations are constrained. In the second model, Schneidman *et al* constrained the maximum entropy model with both mean firing

J. Stat. Mech. (2013) P03006

rates and instantaneous pairwise correlations between neurons. In this case,

$$\mathcal{H}_{\boldsymbol{\beta}}(\omega(0)) = \sum_{k=1}^{N} \beta_k\, \omega_k(0) + \sum_{k,l=1}^{N} \beta_{kl}\, \omega_k(0)\, \omega_l(0).$$

Here the potential can be identified with the Hamiltonian of a magnetic system with binary spins. It is thus often called the 'Ising model' in the spike train analysis literature, although the original Ising model has constant and positive couplings [21]. The corresponding statistical model is the least structured model respecting these first-order and second-order pairwise instantaneous constraints. The number of parameters is of the order[3] of N^2, to be compared with the 2^N possible patterns.

Schneidman *et al* showed that the Ising model successfully predicts spatial patterns, a result which was confirmed by [61] (see [39] for a review). Other works have used the same method and found also a good prediction in cortical structure *in vitro* [69], and in the visual cortex *in vivo* [79]. Later on, several authors considered higher-order terms still corresponding to $D = 0$ [40, 55, 73, 18]. Note that these results have been obtained on relatively small subsets of neurons (usually groups of 10). An interesting challenge is to test how these results hold for larger subsets of neurons, and if other constraints have to be added [19](Tkacik *et al*, in preparation).

2.2.3. One time step spatio-temporal models and detailed balance. These models are only designed to predict the occurrence of 'spatial' patterns, lying within one time bin. The use of spatial observables naturally leads to a time independence assumption where the probability of occurrence of a spatio-temporal pattern is given by the product (3). Tang *et al* [69] tried to predict the temporal statistics of the neural activity with such a model and showed that it does not give a faithful description of temporal statistics. The idea to consider spatio-temporal observables then naturally emerges with the problem of generalizing the probability equation (7) to that case.

From the statistical mechanics point of view, a natural extension consists of considering the space of rasters Ω as a lattice where one dimension is 'space' (neurons index) and the other is time. The idea is then to consider a potential, still of the form (6), but where the observables correspond to spatio-temporal events. We assume that $\mathcal{H}_{\boldsymbol{\beta}}$ has range $R = D + 1$, $0 \leq D < +\infty$. The potential of a spike block ω_0^n, $n \geq D$ is:

$$\mathcal{H}_{\boldsymbol{\beta}}\left(\,\omega_0^n\,\right) = \sum_{l=0}^{n-D} \mathcal{H}_{\boldsymbol{\beta}}\left(\,\omega_l^{D+l}\,\right). \tag{11}$$

On this basis, restricting to the case where $D = 1$ (one time step memory depth) Marre *et al* have proposed in [35] to construct a Markov chain where transition probabilities $P\left[\,\omega(l+1)\,|\,\omega(l)\,\right]$ are proportional to $\mathrm{e}^{\mathcal{H}_{\boldsymbol{\beta}}(\omega_l^{l+1})}$. If μ is the invariant probability of that chain, the application of (2) leads to probability of blocks $\mu\left[\,\omega_0^n\,\right]$, proportional to $\mathrm{e}^{\mathcal{H}_{\boldsymbol{\beta}}\left(\,\omega_0^n\,\right)}$: the probability of a block is proportional to the exponential of its potential ('energy'). This approach is therefore quite natural from the statistical mechanics point of view.

[3] Most approaches assume moreover that the pairwise coefficients are symmetric, $\beta_{kl} = \beta_{lk}$, which divides the number of parameters by 2.

J. Stat. Mech. (2013) P03006

The main problem, however, is 'what is the proportionality coefficient?' As shown in [35], the normalization of conditional probabilities does not reduce to the mere division by a constant partition function. This normalization factor is itself dependent on the past activity.

To overcome this dependency, Marre *et al* assumed that the activity respected a detailed balance. In this particular case, it can be shown that the normalization factor becomes, again, a constant. But this is an important reduction that could have implications for the interpretation of the data: for example, with this simplification, it is not possible to give an account of asymmetric cross correlograms.

2.3. General spatio-temporal models

We now present the general formalism which allows us to solve the variational problem 'maximizing entropy under spatio-temporal constraints'. This approach is rigorous and the normalization problem is resolved without requiring additional assumptions such as detailed balance. At the end of this section, we briefly discuss other approaches considering spatio-temporal statistics and their relations to potentials of the form (6).

2.3.1. Constructing the Markov chain. In this section we show how one can generate a Markov chain where transition probabilities are proportional to $e^{\mathcal{H}_{\boldsymbol{\beta}}(\omega_l^{l+D})}$, for a potential $\mathcal{H}_{\boldsymbol{\beta}}$ corresponding to spatio-temporal events. We also solve the normalization problem. This construction is well known and is based on the so-called transfer matrix (see e.g. [21] for a presentation in the context of statistical physics; [44] for a presentation in the context of ergodic theory and [77] for a presentation in the context of spike train analysis).

This matrix is constructed as follows. Consider two spike blocks w_1, w_2 of range $D \geq 1$. The transition $w_1 \to w_2$ is *legal* if w_1 has the form $\omega(0)\omega_1^{D-1}$ and w_2 has the form $\omega_1^{D-1}\omega(D)$. The vectors $\omega(0), \omega(D)$ are arbitrary but the block ω_1^{D-1} is common. Here is an example of a legal transition:

$$w_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}; \qquad w_2 = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

Here is an example of a forbidden transition

$$w_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}; \qquad w_2 = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Any block ω_0^D of range $R = D + 1$ can be viewed as a legal transition from the block $w_1 = \omega_0^{D-1}$ to the block $w_2 = \omega_1^D$ and in this case we write $\omega_0^D \sim w_1 w_2$.

The *transfer matrix* $\mathcal{L}$ is defined as:

$$\mathcal{L}_{w_1,w_2} = \begin{cases} e^{\mathcal{H}_{\boldsymbol{\beta}}(\omega_0^D)} & \text{if } w_1, w_2 \text{ is legal with } \omega_0^D \sim w_1 w_2 \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

From the matrix $\mathcal{L}$ the transition matrix of a Markov chain can be constructed, as we now show. Since observables are assumed to be bounded from below, $\mathcal{H}_{\boldsymbol{\beta}}(\omega_0^D) > -\infty$, thus $e^{\mathcal{H}_{\boldsymbol{\beta}}(\omega_0^D)} > 0$ for each legal transition. As a consequence of the Perron–Frobenius theorem [20, 56], $\mathcal{L}$ has a unique real positive eigenvalue $s_{\boldsymbol{\beta}}$, strictly larger than the

modulus of the other eigenvalues (with a positive gap), and with right, R, and left, L, eigenvectors: $\mathcal{L}R = s_{\boldsymbol{\beta}}R$, $L\mathcal{L} = s_{\boldsymbol{\beta}}L$, or, equivalently[4]:

$$\sum_{\omega(D)\in\{0,1\}^N} \mathrm{e}^{\mathcal{H}_{\boldsymbol{\beta}}(\omega_0^D)} R\left(\omega_1^D\right) = s_{\boldsymbol{\beta}} R\left(\omega_0^{D-1}\right);$$
$$\sum_{\omega(0)\in\{0,1\}^N} L\left(\omega_0^{D-1}\right) \mathrm{e}^{\mathcal{H}_{\boldsymbol{\beta}}(\omega_0^D)} = s_{\boldsymbol{\beta}} L\left(\omega_1^D\right).$$

These eigenvectors have strictly positive entries $R(\cdot) > 0$, $L(\cdot) > 0$, functions of blocks of range D. They can be chosen so that the scalar product $\langle L, R\rangle = 1$. We define:

$$\mathcal{P}(\mathcal{H}_{\boldsymbol{\beta}}) = \log s_{\boldsymbol{\beta}}. \tag{13}$$

called 'topological pressure'. We discuss the origin of this term and its properties in section 2.3.2.

To define a Markov chain from the transfer matrix $\mathcal{L}$ (equation (12)) we introduce the *normalized potential*:

$$\phi(\omega_0^D) = \mathcal{H}_{\boldsymbol{\beta}}(\omega_0^D) - \mathcal{G}_{\boldsymbol{\beta}}(\omega_0^D) \tag{14}$$

with:

$$\mathcal{G}_{\boldsymbol{\beta}}(\omega_0^D) = \log R\left(\omega_0^{D-1}\right) - \log R\left(\omega_1^D\right) + \log s_{\boldsymbol{\beta}}, \tag{15}$$

and a family of transition probabilities:

$$P\left[\omega(D) \,\middle|\, \omega_0^{D-1}\right] \stackrel{\text{def}}{=} \mathrm{e}^{\phi(\omega_0^D)} > 0. \tag{16}$$

These transition probabilities define a Markov chain which admits a unique invariant probability:

$$\mu(\omega_0^{D-1}) = R\left(\omega_0^{D-1}\right) L\left(\omega_0^{D-1}\right). \tag{17}$$

From the general form of block probabilities (2) the probability of blocks of depth $n \geq D$ is, in this case:

$$\mu\left[\omega_0^n\right] = \mathrm{e}^{\sum_{l=0}^{n-D}\phi(\omega_l^{D+l})} \mu\left[\omega_0^{D-1}\right]. \tag{18}$$

thus, from (17),(14),(15):

$$\mu\left[\omega_0^n\right] = \frac{\mathrm{e}^{\mathcal{H}_{\boldsymbol{\beta}}(\omega_0^n)}}{s_{\boldsymbol{\beta}}^{n-D+1}} R\left(\omega_{n-D+1}^n\right) L\left(\omega_0^{D-1}\right), \tag{19}$$

where $\mathcal{H}_{\boldsymbol{\beta}}(\omega_0^n)$ is given by (11).

[4] The right eigenvector R has 2^{ND} entries R_w corresponding to blocks of range D. It obeys $\sum_{w_2}\mathcal{L}_{w_1 w_2}R_{w_2} = s_{\boldsymbol{\beta}}R_{w_1}$, where $w_1 = \omega_0^{D-1}$ and where the sum runs over blocks $w_2 = \omega_1^D$. Since $\mathcal{L}_{w_1 w_2}$ is nonzero only if the entries w_1, w_2 have the block ω_1^{D-1} in common, and since the right-hand side $(s_{\boldsymbol{\beta}}R_{w_1})$ fixes the value of w_1, this sum holds in fact on all possible values of $\omega(D)$. The notation R_w, although natural, does not make explicit the block involved. This is problematic when one wants to handle equations such as (19). As a consequence, we prefer to use the notation $R(\mathit{block})$ to make explicit this dependence. The same remark holds mutatis mutandis for the left eigenvector.

J. Stat. Mech. (2013) P03006

2.3.2. Remarks.

(1) We have been able to compute the probability of any blocks ω_0^n. It is proportional to $e^{\mathcal{H}_{\boldsymbol{\beta}}(\omega_0^n)}$ and the proportionality factor has been computed. In the general case of spatio-temporal events, it depends on ω_0^{D-1} and ω_{n-D+1}^n.
The same arises in statistical mechanics when dealing with boundary conditions. The forms (18), (19), are reminiscent of Gibbs distributions on spin lattices, with lattice translation invariant probability distributions given specific boundary conditions. Given a spin potential of spatial range n, the probability of a spin block depends upon the state of the spin block, as well as the spin states in a neighbourhood of that block. The conditional probability of this block given a fixed neighbourhood is the exponential of the energy characterizing physical interactions, within the block, as well as interactions with the boundaries. In (18), spins are replaced by spiking patterns; space is replaced by time. Spatial boundary conditions are here replaced by the dependence upon ω_0^{D-1} and ω_{n-D+1}^n.
As a consequence, as soon as one is dealing with spatio-temporal events the normalization of conditional probabilities does not reduce to the mere division by:

$$Z_n = \sum\nolimits_{\omega_0^n} e^{\mathcal{H}_{\boldsymbol{\beta}}(\omega_0^n)}, \tag{20}$$

as easily checked in (19).

(2) The topological pressure obeys nevertheless:

$$\mathcal{P}(\mathcal{H}_{\boldsymbol{\beta}}) = \lim_{n\to+\infty} \frac{1}{n} \log Z_n,$$

and is analogous to a thermodynamic potential density (free energy, free enthalpy, pressure). This analogy is also clear in the variational principle (23) below. To our best knowledge the term 'topological pressure' has its roots in the thermodynamic formalism of hyperbolic (chaotic) maps [54, 44, 4]. In this context, this function can be computed as the grand potential of the grand canonical ensemble, as a cycle expansion over unstable periodic orbits. It is therefore equivalent to a pressure[5] depending on topological properties (periodic orbits).

(3) In the case $D = 0$ the Gibbs distribution reduces to (7). One can indeed easily show that:

$$\exp \mathcal{G}_{\boldsymbol{\beta}} = s_{\boldsymbol{\beta}} = \sum\nolimits_{\omega(0)} e^{\mathcal{H}_{\boldsymbol{\beta}}(\omega(0))} = Z_{\boldsymbol{\beta}},$$

where $Z_{\boldsymbol{\beta}}$ is the partition function (8). Additionally, since spike patterns occurring at distinct time are independent in the $D = 0$ case, Z_n in (20) can be written as $Z_n = Z_{\boldsymbol{\beta}}^n$ so that $\mathcal{P}(\mathcal{H}_{\boldsymbol{\beta}}) = \log Z_{\boldsymbol{\beta}}$.

(4) In the general case of spatio-temporal constraints, the normalization requires the consideration of a normalizing function $\mathcal{G}_{\boldsymbol{\beta}}$ *depending as well on the blocks* ω_0^D. Thus, in addition to function $\mathcal{H}_{\boldsymbol{\beta}}$ normalization introduces a second *function* of spike blocks. This consequently increases the complexity of Gibbs potentials and Gibbs distributions compared to the spatial ($D = 0$) case where $\mathcal{G}_{\boldsymbol{\beta}}$ reduces to a constant.

[5] The grand potential Φ obeys $\Phi = -PV$, where P is the physical pressure and V the volume. Therefore, the grand potential density is (minus) the pressure.

J. Stat. Mech. (2013) P03006

2.3.3. The maximum entropy principle. We now show that the probability distribution defined this way solves the variational problem 'maximizing entropy under constraints'.

We define the *entropy rate* (or Kolmogorov–Sinai entropy):

$$h\left[\mu\right] = -\limsup_{n\to\infty}\frac{1}{n+1}\sum\nolimits_{\omega_0^n}\mu\left[\omega_0^n\right]\log\mu\left[\omega_0^n\right], \tag{21}$$

where the sum holds over all possible blocks ω_0^n. Note that in the case of a Markov chain $h\left[\mu\right]$ also reads [16]:

$$h\left[\mu\right] = -\sum\nolimits_{\omega_0^D}\mu\left[\omega_0^D\right]\,P\left[\omega(D)\,\middle|\,\omega_0^{D-1}\right]\log P\left[\omega(D)\,\middle|\,\omega_0^{D-1}\right], \tag{22}$$

whereas, when $D=0$, $h\left[\mu\right]$ reduces to the (5).

As a general result from ergodic theory [54, 29, 12] and mathematical statistical physics [21], there is a unique[6] probability distribution μ such that [54, 29, 12]:

$$\mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right] = \sup_{\nu\in\mathcal{M}_{\rm inv}}\left(h\left[\nu\right] + \nu\left[\mathcal{H}_{\boldsymbol{\beta}}\right]\right) = h\left[\mu\right] + \mu\left[\mathcal{H}_{\boldsymbol{\beta}}\right], \tag{23}$$

where $\mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]$ is given by (13). $\mathcal{M}_{\rm inv}$ is the set of all possible time-translation invariant probabilities on the set of rasters with N neurons and $\nu\left[\mathcal{H}_{\boldsymbol{\beta}}\right] = \sum_{\omega_0^D}\mathcal{H}_{\boldsymbol{\beta}}(\omega_0^D)\,\nu(\omega_0^D)$ is the average value of $\mathcal{H}_{\boldsymbol{\beta}}$ with respect to the probability ν.

Looking at the second equality, the variational principle (23) selects, among all possible probabilities ν, a unique one realizing the supremum. This is exactly the invariant distribution of the Markov chain and is the sought Gibbs distribution. It is clear from (23) that the topological pressure is the formal analogue to a thermodynamic potential density, where $\mathcal{H}_{\boldsymbol{\beta}}$ somewhat fixes the 'ensemble': $\nu\left[\mathcal{H}_{\boldsymbol{\beta}}\right] = \sum_{k=1}^{K}\beta_k\nu\left[\mathcal{O}_k\right]$ plays the role of βE (canonical ensemble), $\beta E - \mu N$ (grand canonical ensemble), ... in thermodynamics [4].

2.3.4. Inferring the parameters β_k. The inverse problem of finding the β_k values from the observables average measured on the data is a hard problem with no exact analytical solution. However, in the context of spatial models with pairwise interactions, the wisdom coming from statistical physics, and especially the Ising model and spin glasses, as well as from the Boltzmann machine learning community, can be used. As a consequence, in this context, several strategies have been proposed. Ackley *et al* [2] proposed a technique to estimate the parameters of a Boltzmann machine. This technique is effective for small networks but it is time consuming. In practice, the time necessary to learn the parameters increases exponentially with the number of units. To speed up the parameter estimation, analytical approximations of the inverse problem have been proposed, which express the parameters β_k as a nonlinear function of the correlations of the activity (see for example [68, 51, 57, 45, 53, 2, 25, 28]).

These methods do not give an exact result, but are computationally fast. We do not pretend to review all of them here, but we quote a few prominent examples. In [57], Sessak and Monasson proposed a systematic small-correlation expansion to solve the inverse Ising problem. They were able to compute couplings up to the third order in the correlations for generic magnetizations, and to the seventh order in the case of zero magnetizations. Their resulting expansion outperforms existing algorithms on the Sherrington–Kirkpatrick spin-glass model [60].

[6] The result is straightforward here since we consider bounded potentials with finite range.

J. Stat. Mech. (2013) P03006

Based on a high-field expansion of the Ising thermodynamic potential, Cocco *et al* [14] designed an algorithm to calculate the parameters in a time polynomial with N, where the couplings are expressed as a weighted sum over the power of the correlations. They did not obtained a closed analytical expression, but their algorithm could run in a time that was polynomial in the number of neurons.

Other methods, based on Thouless–Anderson–Palmer equations [71] and linear response [28], or information geometry [68], initially proposed in the field of spin glasses, have been adapted and applied to spike train analysis (see e.g. the work done by Roudi *et al* [52]).

The success of these approximations depends on the dataset, and there is no a priori guarantee about their efficiency at finding the right values of the parameters. However, by getting closer to the correct solution, they can potentially speed up the convergence of the learning by starting with a seed much closer to the real solution than if taking a random starting point.

Note also that all the techniques mentioned above have been designed for the case where there is no temporal interaction (except [14, 52] which are discussed in section 2.3.5). Now, we explain how the parameter estimation can be done in the spatio-temporal models.

In the general case the parameters β_k can be determined thanks to the following properties.

- $\mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]$ is a log generating function of cumulants. First:

$$\frac{\partial \mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]}{\partial \beta_k} = \mu\left[\mathcal{O}_k\right]. \tag{24}$$

  This is an extension of (9) to the time-dependent case.

- Second:

$$\frac{\partial^2 \mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]}{\partial \beta_k \partial \beta_l} = \frac{\partial \mu\left[\mathcal{O}_k\right]}{\partial \beta_l} = \sum_{n=0}^{+\infty} C_{\mathcal{O}_k \mathcal{O}_l}(n), \tag{25}$$

  where $C_{\mathcal{O}_k\,\mathcal{O}_l}(n)$ is the correlation function between the two observables $\mathcal{O}_k$ and $\mathcal{O}_l$ at time n. Note that correlation functions decay exponentially fast whenever $\mathcal{H}_{\boldsymbol{\beta}}$ has finite range. So that $\sum_{n=0}^{+\infty} C_{\mathcal{O}_k\,\mathcal{O}_l}(n) < +\infty$.

  Equation (25) characterizes the variation in the average value of $\mathcal{O}_k$ when varying β_l (linear response). The corresponding matrix is a susceptibility matrix. It controls the Gaussian fluctuations of observables around their mean (central limit theorem) [54, 44, 12]. This is the generalization of (10) to the time-dependent case. As a particular case, the fluctuations of the empirical average $\pi_{\omega}^{(T)}\left[\mathcal{O}_k\right]$ of $\mathcal{O}_k$ around its mean $\mu\left[\mathcal{O}_k\right]$ are Gaussian with a mean-square deviation $\sqrt{\mu\left[\mathcal{O}_k\right]\left(1-\mu\left[\mathcal{O}_k\right]\right)}/\sqrt{T}$.

  It is clear that the structure of the linear response in the case of spatio-temporal constraints is quite a bit more complex than the case $D = 0$ (see equation (10)). Actually, for $D = 0$, all correlations $C_{\mathcal{O}_k\,\mathcal{O}_l}(n)$ vanish for $n > 0$ (distinct times are independent).

J. Stat. Mech. (2013) P03006

- $\mathcal{P}(\mathcal{H}_{\boldsymbol{\beta}})$ is a convex function of $\boldsymbol{\beta}$. As a consequence, if there is a set of $\boldsymbol{\beta}$ values, $\boldsymbol{\beta}^*$, such that

$$\frac{\partial \mathcal{P}\left[\mathcal{H}_{\boldsymbol{\beta}}\right]}{\partial \beta_k^*} = \mu\left[\mathcal{O}_k\right] = C_k, \tag{26}$$

  then this set is unique. Thus, the solution of the variational problem (23) is unique.

Basically, equations (24)–(26), tell us that techniques based on free energy expansion in spatial models can be extended as well to spatio-temporal cases, where the free energy is replaced by the topological pressure. Obviously, estimating (not to speak of computing) the topological pressure can be a formidable task. Although the transfer matrix technique allows the computation of the topological pressure, the use of this method for large N is hopeless (see section 3.1). However, techniques based on periodic orbit expansion and zeta functions could be useful [44]. Additionally, cumulant expansions of the pressure, equations (24) and (25) corresponding to the two first orders, suggest that the extension of methods based on free energy expansion could be used. In addition to the works quoted above, we can also think of constraint satisfaction problems by Mézard and Mora [37] and approaches based on Bethe free energy [78]. Finally, as we checked, the properties of spatio-temporal Gibbs distributions allows one to extend the parameter estimation methods developed for the spatial case in [17, 7] to spatio-temporal distributions (to be published).

2.3.5. Other spatio-temporal models. Here we shortly review alternative spatio-temporal models. We essentially refer to approaches attempting to construct a Markov chain and related invariant probability by proposing a specific form for the transition probabilities.

A prominent example is provided by the so-called linear–nonlinear (LN) models and generalized linear models (GLM) [5, 36, 43, 74, 48, 46, 3, 47]. Shortly, the idea is to model spike statistics by a point process where the instantaneous firing rate of a neuron is a nonlinear function of the past network activity, including feedbacks and interaction between neurons [63]. This model has been applied in a wide variety of experimental settings [6, 13, 70, 8, 42, 74, 46]. Typically, referring e.g. to [3], the rate r_i has the form (adapting to our notations):

$$r_i = f\left[b_i + K_i.x + \sum\nolimits_j H_{ij} r_j\right] \tag{27}$$

where the kernel K_i represents the ith cell's linear receptive field and x is an input. H_{ij} characterizes the effect of spikes emitted in the past by pre-synaptic neuron j on post-synaptic neuron i. In this approach, neurons are assumed to be conditionally independent given the past. The probability to have a given spike response to a stimulus, given the past activity of the network, reads as the product of firing rates (see e.g. equation (eq2.4) in [3]).

In [3] the authors use several Monte Carlo approaches to learn the parameters of the model for a Bayesian decoding of the rasters. Comparing to the method presented in the previous sections, the main advantages of the GLM are: (i) the transition probability is known (postulated) from the beginning and does not require the heavy normalization imposed by potentials of the form (6); (ii) the model parameters have a neurophysiological interpretation, and their number grows at most as a power law in the number of neurons,

J. Stat. Mech. (2013) P03006

as opposed to (6), where the parameters are delicate to interpret and whose number can become quite large, depending on the set of constraints.

Note, however, that a model of the form (27) can be written as well in the form (6): this is a straightforward consequence of the Hammersley–Clifford theorem [22]. The parameters β_k in (6) are then nonlinear functions of the parameters in (27) (see [10] for an example).

The main drawback of this approach is the assumption of conditional independence between neurons: neurons are assumed independent at time t when the past, which appears in the function H_{ij} in (27), is given, and the probability of a spiking pattern at time t is the product of neuron firing rates. In contrast, the maximal entropy principle does not require this assumption.

It is interesting to note that the conditional independence assumption can be rigorously justified in conductance-based integrate-and-fire models [10, 11] and the form of the function f can be explicitly found (this a sigmoid function instead of an exponential as usually postulated in GLM). This result holds true if only chemical synapses are involved (this is also implicit in the kernel form H_{ij} in (27) [3]), but conditional independence breaks down, for example, as soon as electric synapses (gap junctions) are involved: this can be mathematically shown in conductance-based integrate-and-fire models [15]. Note that, in this case, a large fraction of the correlations are due to dynamical interactions between neurons: as a consequence they persist even if there is no shared input.

Recently, Macke *et al* [34] extended the GLM model to fix the lack of instantaneous correlations between neurons in the GLM. They added a common input function that has a linear temporal dynamics. However, one of the disadvantages of this technique is that its likelihood is not unimodal, and thus computationally expensive expectation–maximization algorithms have to be used to fit parameters.

The GLM model is usually used to model both the stimulus–response dependence as well as the interaction between neurons, while the MaxEnt models usually focus on the latter (but see [72]).

To finish this subsection, we would like to quote two important works dealing with spatio-temporal events too. First, in [14] Cocco and co-workers consider the spiking activity of retinal ganglion cells with a dual approach: on one hand they consider an Ising model (and higher-order spatial terms) where they propose an inverse method based on a cluster expansion to find efficiently the coupling in the Ising model from data; on the other hand, they consider the problem of finding the parameters (synaptic couplings) in a integrate-and-fire model with noise from its spike trains. In the weak noise limit the conditional probability of a spiking pattern, given the past, is given by a least action principle. This probability is a Gibbs distribution whose normalized potential is characterized by the action computed over an optimal path. This second approach allows the characterization of spatio-temporal events. Especially it gives a very good fit of the cross correlograms.

Second, in [52], the authors consider a one step memory Markov chain where the conditional probability has a time-dependent potential of the Ising type. Adapting a Thouless–Anderson–Palmer [71] approach used formerly in the Sherrington–Kirkpatrick mean-field model of spin glasses [60] they propose an inversion algorithm to find the model parameters. As in the GLM, their model assumes conditional independence given the past (see equation (1) in [52]).

J. Stat. Mech. (2013) P03006

2.4. Comparing models

Solving equation (26) provides an optimal choice for the Gibbs distribution μ, *given the observables* $\mathcal{O}_k$. However, changing the set of observables provides distinct Gibbs distributions, which do not approximate the hidden probability with the same accuracy. We need here a way to quantify the 'distance' between the 'model' (the Gibbs distribution fixed by the set of observables) and the exact, hidden, probability $\mu^{(*)}$. Here are several criteria of comparison.

2.4.1. Kullback–Leibler divergence. The Kullback–Leibler divergence between $\mu, \mu^{(*)}$ is given by:

$$d(\mu^{(*)}, \mu) = \limsup_{n\to\infty} \frac{1}{n+1} \sum_{\omega_0^n} \mu^{(*)}[\omega_0^n] \log\left[\frac{\mu^{(*)}[\omega_0^n]}{\mu[\omega_0^n]}\right], \tag{28}$$

which provides some notion of asymmetric 'distance' between μ and $\mu^{(*)}$. The KL divergence accounts for the discrepancy between the predicted probability $\mu[\omega_0^n]$ and the exact probability $\mu^{(*)}[\omega_0^n]$ for all blocks of range n.

This quantity is not numerically computable from (28). However, for μ a Gibbs distribution and $\mu^{(*)}$ a time-translation invariant probability, the following holds:

$$d_{\mathrm{KL}}(\mu^{(*)}, \mu) = \mathcal{P}[\mathcal{H}_{\boldsymbol{\beta}}] - \mu^{(*)}[\mathcal{H}_{\boldsymbol{\beta}}] - h(\mu^{(*)}).$$

The topological pressure $\mathcal{P}[\mathcal{H}_{\boldsymbol{\beta}}]$ is given by (13) while $\mu^{(*)}[\mathcal{H}_{\boldsymbol{\beta}}]$ is estimated by $\pi_\omega^{(T)}[\mathcal{H}_{\boldsymbol{\beta}}] = \sum_{k=1}^K \beta_k \pi_\omega^{(T)}[\mathcal{O}_k] = \sum_{k=1}^K \beta_k C_k$.

Since $\mu^{(*)}$ is unknown, $h(\mu^{(*)})$ is unknown, and can only be estimated from data, i.e. one estimates the entropy of the empirical probability, $h(\pi_\omega^{(T)})$. There exist efficient methods for that. Note that the entropy of a Markov chain is readily given by equation (22), so the entropy $h(\pi_\omega^{(T)})$ is obtained by replacing the exact probability P in equation (22), by the empirical probability $h(\pi_\omega^{(T)})$. As $T \to +\infty$, $h(\pi_\omega^{(T)}) \to h(\mu^{(*)})$, at exponential rate[7]. For finite T, finite size corrections exist, see e.g. Strong *et al* [67]. In figure 1 is plotted an example. For a potential $\mathcal{H}_{\boldsymbol{\beta}}$ with $N = 5$ neurons and range $R = 2$, containing all possible observables, we have plotted the difference between the exact probability (known from (22) and the explicit form (16), (17) of transition probabilities and invariant probability) and the approached entropy $h(\pi_\omega^{(T)})$ obtained by replacing the exact probability P by the empirical probability $h(\pi_\omega^{(T)})$, as a function of raster size T. We have also plotted the finite corrections method proposed by Strong *et al* in [67].

Now, if one wants to compare how two Gibbs distributions μ_1, μ_2 approximate data, one compares the divergence $d_{\mathrm{KL}}(\mu^{(*)}, \mu_1)$, $d_{\mathrm{KL}}(\mu^{(*)}, \mu_2)$, where $h(\mu^{(*)})$ is independent of the model choice. Therefore, the comparison of two models can be done without computing $h(\mu^{(*)})$.

2.4.2. Comparison of observables average. Another criterion, easier to compute, is to compare the expected value of the observables average, $\mu^{(*)}[\mathcal{O}_k]$, known from (24) to the empirical average $\pi_\omega^{(T)}[\mathcal{O}_k]$. Error bars are expected to follow the central limit theorem

[7] The rate is given by the spectral gap of the transfer matrix: the difference between the largest eigenvalue (it is real and positive) and the modulus of the second largest eigenvalue (in modulus).

J. Stat. Mech. (2013) P03006

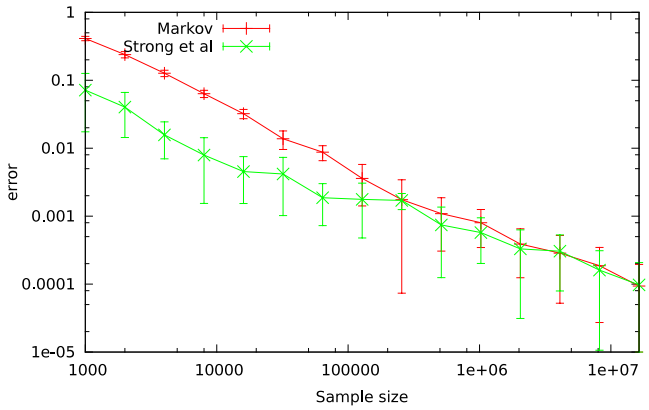

Figure 1. Difference between the exact probability and the approached entropy $h(\pi_\omega^{(T)})$, as a function of raster size T. The potential of test includes all the possible observables where weights are set as random values.

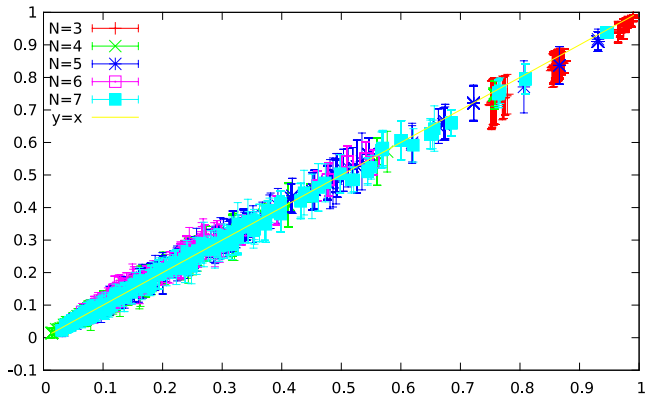

Figure 2. Comparison between the estimated and real values of observable averages.

where fluctuations are given by equation (25). Examples are given in figure 2. Note that the comparison of observables average is less discriminant than minimizing the Kullback–Leibler divergence, since there are infinitely many possible models matching the observables average.

3. Monte Carlo method for spatio-temporal Gibbs distribution

3.1. The advantages and limits of the transfer matrix method

The advantage of the transfer matrix method is that it is mathematically exact: given a potential $\mathcal{H}_{\boldsymbol{\beta}}$, it gives the Gibbs distribution and topological pressure without computing a partition function; given the parametric form (6), where the parameters β_k have to be determined ('learned'), it provides the unique solution. On numerical grounds, this method provides an optimal estimation, in the limits of the error made when observing

J. Stat. Mech. (2013) P03006

the observables empirically, this error being characterized by the central limit theorem. Its main drawback is that the transfer matrix $\mathcal{L}$ has 2^{NR} entries! Although, most of those entries are zero (2^N nonzero entries per row, thanks to the compatibility conditions) it is simply too huge to handle cases where $NR > 24$.

Focusing thus on the huge number of states in the set of blocks, it is clear that any method requiring the extensive description of the phase space fails as NR grows. Additionally, while the accessible phase space is huge, the *observed* phase space (e.g. in an experimental raster) is rather small. Several strategies exist to avoid the extensive description of the phase space. Here, we propose an approach based on Monte Carlo sampling.

The idea is the following. Given a potential $\mathcal{H}_{\boldsymbol{\beta}}$ we find a strategy to approximately compute the average $\mu\left[\mathcal{O}_k\right]$ of observables $\mathcal{O}_k$ under the Gibbs distribution μ, using a statistical Monte Carlo sampling of the phase space. For that purpose, the algorithm generates a raster following the statistics defined by the potential $\mathcal{H}_{\boldsymbol{\beta}}$, and computes the observables on this artificial raster. Thanks to the estimation of the observables, the parameters of the model (β_k) can be found by modifying their values to minimize iteratively the distance between the values of the observables estimated on the real raster, and the values estimated with the Monte Carlo sampling. Powerful algorithms exist for this, taking into account the uncertainty on the empirical averages ruled by the central limit theorem [17, 7].

3.2. The Monte Carlo–Hastings algorithm

The Monte Carlo–Hastings method consists in sampling a target probability distribution μ by constructing a Markov chain whose invariant probability is μ [24]. The transition probability of this Markov chain, between two states $\omega^{(1)}$ and $\omega^{(2)}$, is:

$$P[\omega^{(1)}|\omega^{(2)}] = \max\left(\frac{Q(\omega^{(1)}|\omega^{(2)})}{Q(\omega^{(2)}|\omega^{(1)})}\frac{\mu\left[\,\omega^{(2)}\,\right]}{\mu\left[\,\omega^{(1)}\,\right]}, 1\right). \tag{29}$$

The function $Q(\ )$ can have different forms, allowing in particular the acceleration of the convergence rate of the algorithm. Such specific forms are highly dependent on the form of $\mathcal{H}_{\boldsymbol{\beta}}$, and there is no general recipe to determine Q, given $\mathcal{H}_{\boldsymbol{\beta}}$. The contribution of Q cancels in (29) whenever Q is symmetric ($Q(\omega|\omega') = Q(\omega'|\omega)$). We make this assumption in the following. Practically, we take Q as the uniform distribution corresponding to flipping one spike at each iteration of the method.

In classical Monte Carlo approaches in statistical physics, the normalization factor of the Gibbs distribution, the partition function, cancels when computing the ratio of two block probabilities $\mu\left[\,\omega^{(2)}\,\right]/\mu\left[\,\omega^{(1)}\,\right]$. The situation is different in the presence of spatio-temporal constraints, as shown in equation (19): 'boundary terms' $L\left(\omega_0^{D-1}\right)$, $R\left(\omega_{n-D+1}^{n}\right)$ remain. Actually, the same would hold in a statistical physics problem with spatial interactions if one were to compare the probability of bulk spin chains with distinct boundary conditions.

This problem can however be circumvented thanks to the following remarks:

J. Stat. Mech. (2013) P03006

(1) If one compares the probability of two blocks $\omega^{(1)}, \omega^{(2)}$ of range $n \geq 2D+1$, with $\omega_0^{D-1,(1)} = \omega_0^{D-1,(2)}$ and $\omega_{n-D+1}^{n,(1)} = \omega_{n-D+1}^{n,(2)}$ then (19) reads:

$$\frac{\mu[\omega_0^{n,(2)}]}{\mu[\omega_0^{n,(1)}]} = \mathrm{e}^{\Delta\mathcal{H}_{\boldsymbol{\beta}}(\omega^{(1)},\omega^{(2)},0,n)}$$

with

$$\Delta\mathcal{H}_{\boldsymbol{\beta}}(\omega^{(1)},\omega^{(2)},0,n) = \mathcal{H}_{\boldsymbol{\beta}}(\omega_0^{n,(1)}) - \mathcal{H}_{\boldsymbol{\beta}}(\omega_0^{n,(2)}).$$

Thus, the Monte Carlo transition probability (29) is only expressed as a difference of the potential of the two blocks.

(2) $\Delta\mathcal{H}_{\boldsymbol{\beta}}(\omega^{(1)},\omega^{(2)},0,n) = \sum_{k=1}^{K}\beta_k\Delta\mathcal{O}_k(\omega^{(1)},\omega^{(2)},0,n)$, with:

$$\Delta\mathcal{O}_k(\omega^{(1)},\omega^{(2)},0,n) = \sum_{l=0}^{n-D}[\mathcal{O}_k(\omega_l^{D+l,(2)}) - \mathcal{O}_k(\omega_l^{D+l,(1)})].$$

Since the $\mathcal{O}_k$ are monomials, many terms $\mathcal{O}_k(\omega_l'^{l+D}) - \mathcal{O}_k(\omega_l^{l+D})$ cancel. Assuming that we flip a spike at position (k,t), $k \in \{1,\ldots,N\}$, $t \in \{D, n-D\}$, we have indeed:

$$\Delta\mathcal{O}_k(\omega^{(1)},\omega^{(2)},0,n) = \sum_{l=t-D}^{t}[\mathcal{O}_k(\omega_l^{D+l,(2)}) - \mathcal{O}_k(\omega_l^{D+l,(1)})].$$

Since the difference $\mathcal{O}_k(\omega_l^{D+l,(2)}) - \mathcal{O}_k(\omega_l^{D+l,(1)}) \in \{-1,0,1\}$, the computational cost of $\Delta\mathcal{O}_k(\omega^{(1)},\omega^{(2)},0,n)$ is minimal if one makes a list of monomials affected by the flip of spike (k,r), $r = 0,\ldots,D$.

3.3. Convergence rate

The goal of the Monte Carlo–Hastings algorithm is to generate a sample of a target probability obtained by iteration of the Markov chain defined by equation (29). In our case, this sample is a raster ω_0^{T-1}, distributed according to a Gibbs distribution μ. Call N_{flip} the number of iterations ('flips' in our case) of the Monte Carlo algorithm. As $N_{\text{flip}} \to +\infty$ the probability that the algorithm generates a raster ω_0^{T-1} tends to $\mu\left[\omega_0^{T-1}\right]$. Equivalently, if one generates N_{seed} rasters and denotes $\#\left(\omega_0^{T-1}\right)$ the number of occurrences of a specific bloc ω_0^{T-1}, then:

$$\lim_{N_{\text{seed}}\to+\infty}\lim_{N_{\text{flip}}\to+\infty}\frac{\#\left(\omega_0^{T-1}\right)}{N_{\text{seed}}} = \mu\left[\omega_0^{T-1}\right].$$

The convergence is typically exponential with a rate depending on $\mathcal{H}_{\boldsymbol{\beta}}$.

Now, the goal here is to use a Monte Carlo raster to estimate $\mu[\mathcal{O}_k]$ by performing the empirical average $\pi_\omega^{(T)}[\mathcal{O}_k]$ on that raster. However, as explained in section 2.1.4, even if the raster is distributed according to μ (corresponding thus to taking the limit $N_{\text{flip}} \to +\infty$) the empirical average $\pi_\omega^{(T)}[\mathcal{O}_k]$ is not equal to $\mu[\mathcal{O}_k]$, it converges to $\mu[\mathcal{O}_k]$ as $T \to +\infty$, with an exponential rate (see footnote 7). More precisely, the probability that the difference $\left|\pi_\omega^{(T)}[\mathcal{O}_k] - \mu[\mathcal{O}_k]\right|$ exceeds some $\epsilon > 0$ behaves like $\exp(-T \times I(\epsilon))$, where $I(\epsilon)$, called the large-deviations rate, is the Legendre transform of the topological pressure [12].

J. Stat. Mech. (2013) P03006

When T is large we have:

$$\mu[\,|\,\pi_\omega^{(T)}\,[\,\mathcal{O}_k\,]-\mu\,[\,\mathcal{O}_k\,]\,|>\epsilon]\simeq\exp\left(\frac{-T\times\epsilon^2}{\sigma(\mathcal{O}_k)}\right) \tag{30}$$

where $\sigma(\mathcal{O}_k)=\sqrt{\mu\,[\,\mathcal{O}_k\,](1-\mu\,[\,\mathcal{O}_k\,])}$ is the mean-square deviations of $\mathcal{O}_k$.

As a consequence, to obtain the exact average $\mu\,[\,\mathcal{O}_k\,]$ from our Monte Carlo algorithm we would need to take the limits:

$$\lim_{T\to+\infty}\lim_{N_{\rm seed}\to+\infty}\lim_{N_{\rm flip}\to+\infty}\frac{\#\left(\omega_0^{T-1}\right)}{N_{\rm seed}}, \tag{31}$$

in *that order*: they do not commute. A prominent illustration of this point is illustrated in figure 4.

For consistency of notation we note from now on $T-1\equiv N_{\rm times}$ for the raster length. When dealing with numerical simulations with a finite number of samples, the goal is to minimize the probability that the error is bigger than a real number ϵ, by a suitable choice of:

- The raster length: $T-1=N_{\rm times}$.
- The number of flips: $N_{\rm flip}$.
- The number of seeds: $N_{\rm seed}$.

Let us now establish a few relations between those parameters. First, it is somewhat evident that $N_{\rm flip}$ must be at least proportional to $N\times N_{\rm times}$ in order to give a chance to all spikes in the raster to be flipped at least once. This criterion respects the order of limits in (31).

Since μ is ergodic one can in principle estimate the average of observables by taking $N_{\rm seed}=1$ and taking $N_{\rm times}$ large. However, the larger $N_{\rm times}$ then the larger $N_{\rm flip}$, and too large $N_{\rm times}$ leads to too long simulations. In contrast, one could generate a large number $N_{\rm seed}$ of rasters with a small $N_{\rm times}$. This would have the advantage of reducing $N_{\rm flip}$ as well. However, the error (30) would then be too large. So, one needs to find a compromise: $N_{\rm times}$ large enough to have small Gaussian fluctuations (30) and small enough to limit $N_{\rm flip}$. Then, by increasing $N_{\rm seed}$, one approaches the optimal bound on fluctuations given by (30). Additionally, this provides error bars.

4. Numerical tests

In this section, performance in terms of convergence rate and CPU time for increasing values of N (number of neurons) are discussed. First, we consider potentials (6) where the $\mathcal{O}_k$s are observables of the form (1) ('polynomial potentials'), and we compare the Monte Carlo results to those obtained using the transfer matrix and the Perron–Frobenius theorem. As discussed in section 3.1, the transfer matrix method becomes rapidly numerically intractable, so that the comparison between Monte Carlo averages and exact averages cannot be done for large N. To circumvent this problem, we introduce, in section 4.2, a specific class of potentials for which the analytical computation of the topological pressure as well as observable averages can be analytically done, whatever N, R. This provides another series of tests.

J. Stat. Mech. (2013) P03006

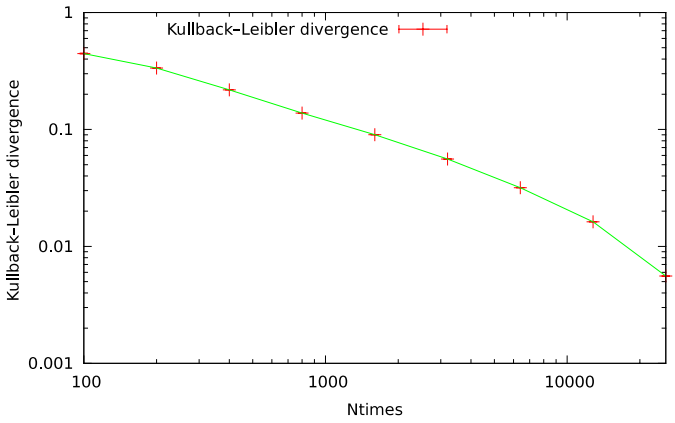

Figure 3. Evolution of Kullback–Leibler divergence (28) as a function N_{times}.

4.1. Polynomial potentials

In this section, we present Monte Carlo simulations with a potential of the form (6), where the $\mathcal{O}_k$ are 100 observables randomly chosen among the 2^{NR} possibilities. More precisely, we select randomly a fraction $1/R$ of 'rates', a fraction $1/R$ of pairwise terms and so on.

4.1.1. Checking observables average. In figure 2 we show the comparison between the exact values of the observable averages (ordinate) and the estimated Monte Carlo values (abscissa). The error bars have been computed with N_{seed} samples. The tests were performed with $N_{\text{seed}} = 20$, $N_{\text{times}} = 10\,000$ and $N_{\text{flip}} = 100\,000$. In this example, N goes from 3 to 8 and $R = 3$. For larger values of NR the numerical computation with the transfer matrix method is no longer possible ($NR = 24$ corresponds to matrices of size $16\,777\,216 \times 16\,777\,216$).

4.1.2. Convergence rate. In this section, we show how the Kullback–Leibler divergence varies as a function of N_{times}. In figure 3 we show the evolution of the Kullback–Leibler divergence between the real distribution and its estimation with the Monte Carlo method.

We also consider the error

$$\text{error} = \max_{k=1\cdots K} \left| 1 - \frac{\pi^{(T)}[\mathcal{O}_k]}{\mu[\mathcal{O}_k]} \right|. \tag{32}$$

As developed in section 3.3, this quantity is expected to converge to 0 if $N_{\text{times}} \to +\infty$ when N_{flip} grows proportional to $N \times N_{\text{times}}$. For finite N_{times}, $N_{\text{flip}} \to +\infty$ the error is controlled by the central limit theorem. The probability that the error on the average of observable $\mathcal{O}_k$ is bigger than ϵ (equation (30)), behaves like $\exp(-N_{\text{times}} \times \epsilon^2/\sigma(\mathcal{O}_k))$.

In contrast, if N_{flip} stays constant while N_{times} grows, the error is expected to first decrease to a minimum, after which it increases. This is because the number of flips is insufficient to reach the equilibrium distribution of the Monte Carlo–Hastings Markov chain. This effect is presented in figure 4. It shows the error (32) as a function of N_{times} for $N = 3$ to 7 neurons with 3 different N_{flip} values (1000, 10 000, 100 000).

J. Stat. Mech. (2013) P03006

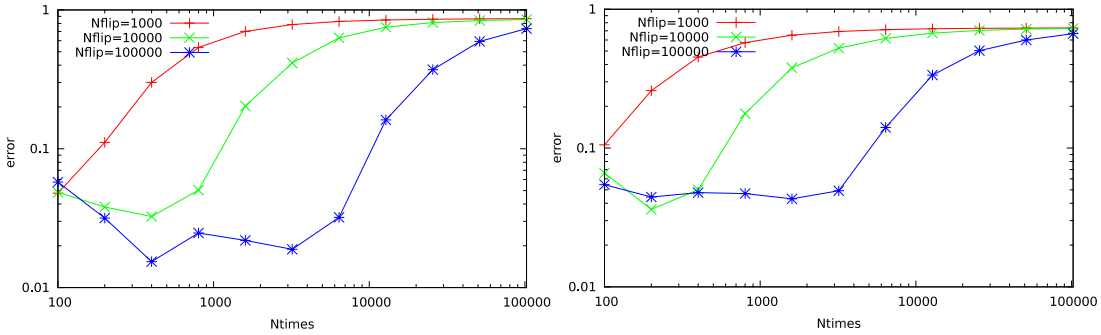

Figure 4. Error as a function of N_{times}, for several values of N_{flip} (1000, 10 000, 100 000). (Left) $N = 3$; (right) $N = 7$.

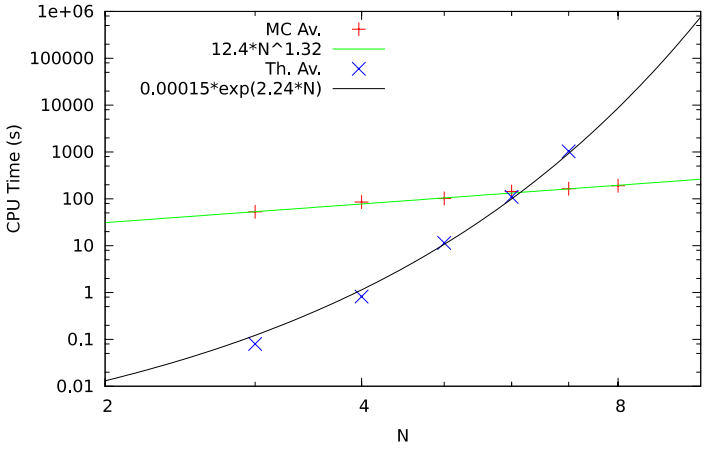

Figure 5. The CPU time necessary to obtain the observable averages presented in figure 2, for the Monte Carlo average (MC Av.) and for the exact average (Th. Av.), as a function of N. The full lines correspond to fit.

Clearly, the number of flips N_{flip} should be at least more than $N \times N_{\text{times}}$ in order to give a chance to all spikes in the raster to be flipped at least once. A value $N_{\text{flip}} = 10 \times N \times N_{\text{times}}$ seems to be enough and computationally reasonable to perform the estimations. With an $N_{\text{seed}} = 20$, we have results with a reasonable error around the mean values.

4.1.3. CPU time. Here we compare the CPU time for a Monte Carlo simulation and the time for a computation with the transfer matrix. We illustrate this in figure 5. We have plotted the CPU time necessary to obtain the observables average presented in figure 2, for the Monte Carlo average and for the exact average, as a function of N. The CPU time for the Monte Carlo method increases slightly more than linearly while the CPU time for the transfer matrix method increases exponentially fast: note that $R = 3$ here, so that $R \log 2 = 2.08$, close to the exponential rate found by fit: 2.24.

We also plot in figure 6 the CPU times as a function of N_{times} with three N_{flip} values (1000, 10 000, 100 000), for 3 and 7 neurons. The CPU time increases in a linear fashion

J. Stat. Mech. (2013) P03006

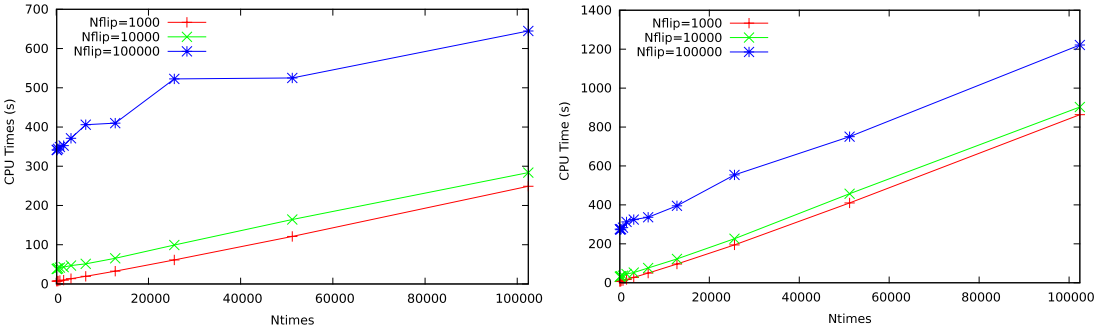

Figure 6. The CPU time (T_{cpu}) as a function of N_{times}. T_{cpu} increases in a linear fashion with N_{time} as $aN + b$, where a is a function of N and b is a function of N_{flip}.

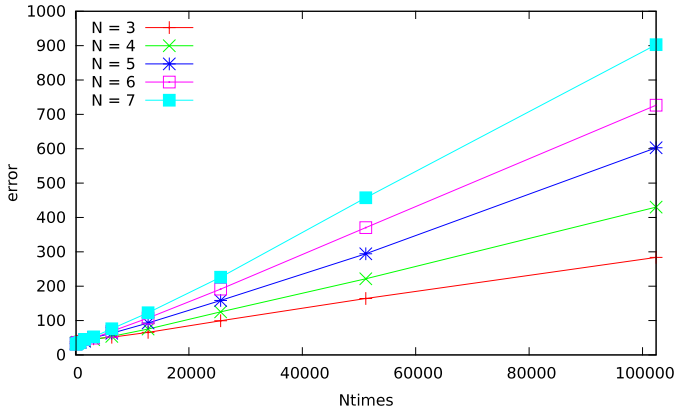

Figure 7. The CPU time (T_{cpu}) as a function of N_{times} for several N values ($N_{\mathrm{flip}} = 10\,000$).

with the N_{times} value. The CPU time also increases linearly with the N_{flip} value (for the same N value). The simulations have been done on a computer with the following characteristics: 7 Intel(R) Xeon(R) 3.20 GHz processors with a 31.5 Gb RAM.

In figure 7 we compare the CPU time increase with N_{times} for several N values. It shows that the CPU time increases linearly with N_{times} (as in figure 6). However, the slope increases with the number of neurons N.

4.2. An analytically solvable example as a benchmark

In this section we consider a specific example of potentials for which the topological pressure is analytically computable, whatever N, R. As a consequence the average of observables and fluctuations can also be computed. This example is obviously rather specific, but its main interests are to provide a didactic illustration of the application of thermodynamic formalism as well as a benchmark for numerical methods.

4.2.1. Analytical setting. We fix the number of neurons N and the range R and we choose L distinct pairs (i_l, t_l), $l = 1 \cdots L$, $i_l \in \{1, \ldots, N\}$, $t_l \in \{0, \ldots, D-1\}$. To this

J. Stat. Mech. (2013) P03006

set is associated a set of $K = 2^L$ events $\mathcal{E}_k = (\omega_{i_1}(t_1), \ldots, \omega_{i_l}(t_l))$, $k = 0, \ldots, K-1$. For example, if $L = 2$, there are 4 possible events $\mathcal{E}_0 = (0,0)$: neuron i_1 is not firing at time t_1 and neuron i_2 is not firing at time t_2; $\mathcal{E}_1 = (0,1)$: neuron i_1 is not firing at time t_1 and neuron i_2 is firing at time t_2; and so on. It is convenient to have a label k corresponding to the binary code of the event.

We define K observables $\mathcal{O}_k$ of range D taking binary values $0, 1$. For a block ω_0^D, $\mathcal{O}_k\left[\omega_0^D\right] = 0$ if the event $\mathcal{E}_k$ is not realized in the bloc ω_0^D and is 1 otherwise. In the example above, $\mathcal{O}_0\left[\omega_0^D\right] = 1$ if neuron i_1 is not firing at time t_1 and neuron i_2 is not firing at time t_2 in the block ω_0^D. Thus, for $N = 3$, $R = 4$, $(i_1, t_1) = (1,0); (i_2; t_2) = (1,1)$,

$$\mathcal{O}_0\left[\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}\right] = \mathcal{O}_0\left[\begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}\right] = 1,$$

while

$$\mathcal{O}_0\left[\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}\right] = \mathcal{O}_0\left[\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}\right] = 0.$$

We finally define a potential $\mathcal{H}$ as in (6), $\mathcal{H}_{\boldsymbol{\beta}} = \sum_{k=1}^K \beta_k \mathcal{O}_k$.

For this type of potential, whatever ω_0^{D-1}, $\sum_{\omega_0^{D-1}} \mathrm{e}^{\mathcal{H}_{\boldsymbol{\beta}}\left(\omega_0^D\right)}$ is independent of $\omega(D)$. As a consequence of the Perron–Frobenius theorem $s_{\boldsymbol{\beta}} = \sum_{\omega_0^{D-1}} \mathrm{e}^{\mathcal{H}_{\boldsymbol{\beta}}\left(\omega_0^D\right)}$ and therefore:

$$\mathcal{P}(\mathcal{H}_{\boldsymbol{\beta}}) = (N-L)\log(2) + \log\left[\sum_{k=1}^K \mathrm{e}^{\beta_k}\right].$$

As a consequence, from (24), giving the average of observable $\mathcal{O}_k$ as the derivative of $\mathcal{P}$ with respect to β_k:

$$\mu\left[\mathcal{O}_k\right] = \frac{\mathrm{e}^{\beta_k}}{\sum_{n=1}^K \mathrm{e}^{\beta_n}}.$$

The fluctuations of observables can also be estimated as well. They are Gaussian with a covariance matrix given by the Hessian of $\mathcal{P}$ (see equation (25)) and the central limit theorem.

Remark. An important assumption here is that observables do not depend on $\omega(D)$. This important simplification as well as the specific form of observables makes the computation of $\mathcal{P}$ tractable. Note however that, although $\mathcal{H}$ does not depend on $\omega(D)$ as well, the *normalized* potential and therefore the conditional probability $P\left[\omega(D) \,\middle|\, \omega_0^{D-1}\right]$ depend on $\omega(D)$ thanks to the normalization factor $\mathcal{G}$ and its dependence in the right eigenvector R of the Perron–Frobenius matrix.

4.2.2. Numerical results for large-scale networks. Let us now use this example as a benchmark for our Monte Carlo method. We have considered a case with range $R = 4$ and $L = 6$ pairs, corresponding to $2^6 = 64$ terms in the potential. We have analysed the convergence rate as a function of N, the number of neurons, and N_{times}, the time length

J. Stat. Mech. (2013) P03006

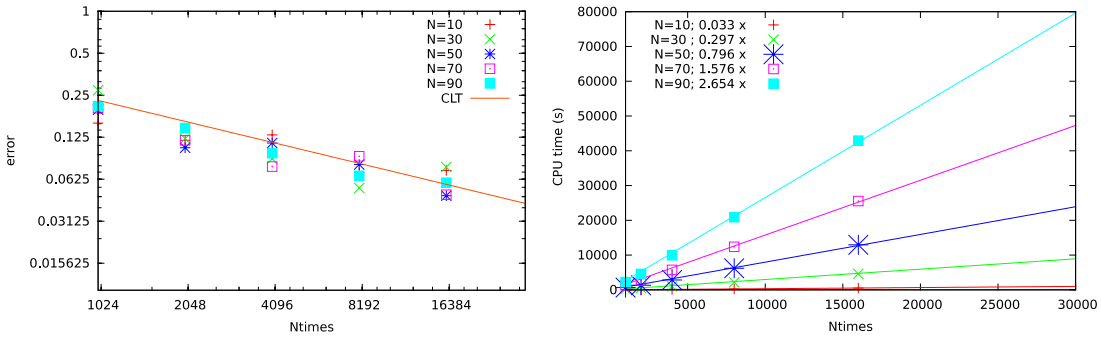

Figure 8. (Left) Relative error as a function of N_{times} for several values of N. CLT indicates the decay expected from the central limit theorem. As expected from the central limit theorem , the error decreases as a power law $N_{\text{times}}^{-1/2}$. (Right) CPU time as a function of N_{times} for several values of N. It increases linearly as $\text{CPU} = aN_{\text{times}}$. In the legend, next to the value of N, we indicate the value of a.

of the Monte Carlo raster. The number of flips, N_{flip} is fixed to $10 \times N \times N_{\text{times}}$, so that, on average, each spike of the Monte Carlo raster is flipped ten times in one trial.

In figure 8 (left) we have shown the relative error (equation (32)) as a function of N_{times} for several values of N. The empirical average $\pi^{(T)}[\mathcal{O}_k]$ is computed on 10 Monte Carlo rasters. That's why we don't write the index ω in the empirical probability. We stopped the simulation when the error is lower than 5%. As expected from the central limit theorem (CLT), the error decreases as a power law $CN_{\text{times}}^{-1/2}$, where the constant C has been obtained by fit.

In figure 8 (right) we have drawn the CPU time as a function of N_{times} for several values of N. It increases linearly with N_{times}, with a coefficient depending on N. The simulation is relatively fast: it takes 150 mins for $N = 60$, $N_{\text{times}} = 8000$ (with 10 Monte Carlo trials) on a 7-processor machine (each of these processors has the following specifications: Intel(R) Xeon(R) CPU 2.27 GHz, 1.55 MB of cache memory size) with a 17.72 GB RAM.

5. Discussion and perspectives

In this paper, we have shown how maximum entropy models can be extended to analyse the spatio-temporal dynamics of the neural activity. This raises specific issues, mainly related to the fact that the normalization of the Gibbs potential depends on the past activity. We have shown that transfer matrix results allow one to handle this problem properly, providing additional crucial information on statistics (especially the average of observables and fluctuations of empirical averages). The challenge is then to be able to fit these models to the recordings. A major step in the fitting process is to compute the observables generated by the model for a given set of parameters. We have proposed a first method, based on the transfer matrix. It gives exact results, but can only be applied to small subsets of neurons. We have then designed a Monte Carlo approach that overcomes this issue, confirmed by several tests.

In fact, matching Gibbs averages of observables is only a first, although crucial, step towards spike train analysis of neuronal activity. The next step consists of fitting

J. Stat. Mech. (2013) P03006

the parameters of a model from an experimental raster. Basically, this corresponds to minimizing the Kullback–Leibler divergence (28) between the model and the empirical measure. We have reviewed some possible techniques in section 2.3.4. The application of our method to fit Gibbs distributions on large-scale retina recordings will be considered in a forthcoming paper.

As a final issue we would now like to discuss shortcomings of maximum entropy models. Although initially proposed as powerful methods for neuroscience applications, many future reports have cast doubt on how useful (spatial pairwise) maximum entropy models are. These criticism include the role of common input [34], the role of higher-order correlations [66, 40], scaling properties [51] and nonstationarity [66]. As a matter of fact, the models presented in section 2.3.5 are clear and efficient alternatives to spatial pairwise maximum entropy models. Let us now comment upon these shortcomings.

First, as we have developed, the maximum entropy approach is definitely not limited to spatial pairwise interactions. In particular, the role of higher-order interactions beyond pairwise equal time ones, provides a clear motivation for including a longer temporal history in statistical models of neural data.

This raises, however, the question of how one chooses the potential. As revealed in section 2.2.1, the possible number of constraints is simply overwhelming and one has to make choices to reduce their number. These choices can be based on ad hoc assumptions (e.g. rates or instantaneous pairwise correlations are essential in neuronal coding) or on empirical constraints (type of cells, spatial localization). However, this combinatorial complexity is clearly a source of troubles and questions about the real efficiency of the maximum entropy problem. This problem is particularly salient when dealing with temporal interactions of increasing memory: even the number of possible pairwise interactions might be too large to fit all of them on finite size recordings. Additionally, using a too complex potential increases the number of parameters necessary to fit the data beyond what the number of available samples allows.

One solution is to try to infer the form of the potential from the data set. Important theorems in ergodic theory [49] as well as in variable length Markov chain estimations can be used [9]. This will be developed in a separate paper.

It is however quite restrictive to stick to potentials expressed as linear combinations of observables, like (6). This form has its roots in thermodynamics and statistical physics, but is far from being the most general form. Nonlinear potentials such as (27) (GLM) can be considered as well. Although such potentials can be expressed in the form (6) from the Hammersley–Clifford theorem [22], this representation induces a huge redundancy in the coefficients β_k. Examples are known of nonlinear potentials with relatively small numbers of parameters λ_l, which, expressed in the form (6), give rise to 2^{NR} parameters β_k, all of them being functions of the λ_l: see [10, 11]. Such potentials constitute relevant alternatives to (6) where the formalism described here fully applies.

More generally, alternatives to maximum entropy models consider different models trying to mimic the origin of the observed correlations. This is the case of the model proposed by [34], where common inputs are added to account for the instantaneous correlations, and the GLM model, where the numbers of parameters is only N^2. However, note that in all these cases the models constrain the correlations to be in a specific form, and might not be a good description of the activity either. Testing these models on data is the only way to distinguish the most relevant ones. Note that the discrepancy between

J. Stat. Mech. (2013) P03006

model and data will probably be more and more obvious with a larger set of neurons. The validity of a model will also depend on size of the recorded population.

Another, even deeper, question is the translation invariance assumption intrinsic to the maximum entropy principle. When dealing, for example, with transient responses to temporary stimuli this assumption is clearly highly controversial. Note however that although the maximum entropy principle does not extend to nontranslationally invariant statistics, the concept of Gibbs distribution extends to that case [21]. Here, Gibbs distributions are constructed via transition probabilities, possibly with an infinite memory. Examples of applications to neuronal networks can be found in [10, 11]. However, the application of this concept to analysing real data, especially the problem of parameter estimation, remains to our knowledge an open challenge.

Acknowledgments

We are grateful to G Tkacik, T Mora, S Kraria, T Viéville and F Hebert for advice and help. This work was supported by the INRIA, ERC-NERVI number 227747, KEOPS ANR-CONICYT and European Union Project # FP7-269921 (BrainScales) projects to BC and HN and ANR OPTIMA to OM. Finally, we would like to thank the reviewers for helpful comments and remarks.

References

[1] Abeles M, 1982 *Local Cortical Circuits: An Electrophysiological study* (Berlin: Springer)
[2] Ackley H, Hinton E and Sejnowski J, *A learning algorithm for Boltzmann machines*, 1985 *Cogn. Sci.* **9** 147–69
[3] Ahmadian Y, Pillow J W and Paninski L, *Efficient Markov chain Monte Carlo methods for decoding neural spike trains*, 2011 *Neural Computat.* **23** 46–96
[4] Beck C and Schloegl F, 1995 *Thermodynamics of Chaotic Systems: An Introduction* (Cambridge: Cambridge University Press)
[5] Brillinger D R, *Maximum likelihood analysis of spike trains of interacting nerve cells*, 1988 *Biol. Cybern.* **59** 189–200
[6] Brillinger D R, *Nerve cell spike train data analysis—a progression of technique*, 1992 *J. Am. Statist. Assoc.* **87** 260–71
[7] Broderick T, Dudik M, Tkacik G, Schapire R E and Bialek W, *Faster solutions of the inverse pairwise Ising problem*, 2007 submitted (see arXiv:0712.2437)
[8] Brown E N, Kass R E and Mitra P P, *Multiple neural spike train data analysis: state-of-the-art and future challenges*, 2004 *Nature Neurosci.* **7** 456–61
[9] Buehlmann P and Wyner A J, *Variable length Markov chains*, 1999 *Ann. Stat.* **27** 480–513
[10] Cessac B, *A discrete time neural network model with spiking neurons II. Dynamics with noise*, 2011 *J. Math. Biol.* **62** 863–900
[11] Cessac B, *Statistics of spike trains in conductance-based neural networks: rigorous results*, 2011 *J. Math. Neurosci.* **1** (8) 1–42
[12] Chazottes J R and Keller G, *Pressure and equilibrium states in ergodic theory*, 2008 *Isr. J. Math.* **131** 1
[13] Chichilnisky E J, *A simple white noise analysis of neuronal light responses*, 2001 *Netw. Comput. Neural Syst.* **12** 199–213
[14] Cocco S, Leibler S and Monasson R, *Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods*, 2009 *Proc. Nat. Acad. Sci.* **106** 14058–62
[15] Cofré R and Cessac B, *Dynamics and spike trains statistics in conductance-based integrate-and-fire neural networks with chemical and electric synapses*, 2012 *Chaos Solitons Fractals* submitted
[16] Cornfeld I P, Fomin S V and Sinai Ya G, 1982 *Ergodic Theory* (Berlin: Springer)
[17] Dudík M, Phillips S J and Schapire R E, *Performance guarantees for regularized maximum entropy density estimation*, 2004 *Proc. 17th Annu. Conf. on Computational Learning Theory*
[18] Ganmor E, Segev R and Schneidman E, *The architecture of functional interaction networks in the retina*, 2011 *J. Neurosci.* **31** 3044–54

J. Stat. Mech. (2013) P03006

[19] Ganmor E, Segev R and Schneidman E, *Sparse low-order interaction network underlies a highly correlated and learnable neural population code*, 2011 *Proc. Nat. Acad. Sci.* **108** 9679–84
[20] Gantmacher F R, 1998 *The Theory of Matrices* (Providence, RI: AMS Chelsea Publishing)
[21] Georgii H-O, 1988 *Gibbs Measures and Phase Transitions* (*De Gruyter Studies in Mathematics* vol 9) (Berlin: de Gruyter)
[22] Hammersley J M and Clifford P, *Markov fields on finite graphs and lattices*, 1971 unpublished
[23] Harris K D, Henze D A, Hirase H, Leinekugel X, Dragoi G, Czurko A and Buzsaki G, *Spike train dynamics predicts theta-related phase precession in hippocampal pyramidal cells*, 2002 *Nature* **417** 738–41
[24] Hastings W K, *Monte Carlo sampling methods using Markov chains and their applications*, 1970 *Biometrika* **57** 97–109
[25] Higuchi S and Mezard M, *Susceptibility propagation for constraint satisfaction problems*, 2009 arXiv:0903.1621
[26] Ikegaya Y, Aaron G, Cossart R, Aronov D, Lampl I, Ferster D and Yuste R, *Synfire chains and cortical songs: temporal modules of cortical activity*, 2004 *Science* **304** 559–64
[27] Jaynes E T, *Information theory and statistical mechanics*, 1957 *Phys. Rev.* **106** 620
[28] Kappen H J and Rodriguez F B, *Boltzmann machine learning using mean field theory and linear response correction*, 1998 *Advances in Neural Information Processing Systems* (Cambridge, MA: MIT Press) pp 280–6
[29] Keller G, 1998 *Equilibrium States in Ergodic Theory* (Cambridge: Cambridge University Press)
[30] Kenet T, Bibitchkov D, Tsodyks M, Grinvald A and Arieli A, *Spontaneously emerging cortical representations of visual attributes*, 2003 *Nature* **425** 954–6
[31] Lampl I, Reichova I and Ferster D, *Synchronous membrane potential fluctuations in neurons of the cat visual cortex*, 1999 *Neuron* **22** 361–74
[32] Louie K and Wilson M A, *Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep*, 2001 *Neuron* **29** 145–56
[33] Luczak A, Barthó P and Harris K D, *Spontaneous events outline the realm of possible sensory responses in neocortical populations*, 2009 *Neuron* **62** 413–25
[34] Macke J H, Busing L, Cunningham J P, Yu B M, Shenoy K V and Sahani M, *Empirical models of spiking in neural populations*, 2011 *Advances in Neural Information Processing Systems* ed J Shawe-Taylor *et al* pp 1350–8
[35] Marre O, El Boustani S, Frégnac Y and Destexhe A, *Prediction of spatiotemporal patterns of neural activity from pairwise correlations*, 2009 *Phys. Rev. Lett.* **102** 138101
[36] McCullagh P and Nelder J A, 1989 *Generalized Linear Models* 2nd edn (London: Chapman and Hall)
[37] Mézard M and Mora T, *Constraint satisfaction problems and neural networks: a statistical physics perspective*, 2009 *J. Physiol. Paris* **103** 107–13
[38] Mokeichev A, Okun M, Barak O, Katz Y, Ben-Shahar O and Lampl I, *Stochastic emergence of repeating cortical motifs in spontaneous membrane potential fluctuations in vivo*, 2007 *Neuron* **53** 413–25
[39] Nirenberg S H and Victor J D, *Analyzing the activity of large populations of neurons: how tractable is the problem?*, 2007 *Curr. Opin. Neurobiol.* **17** 397–400
[40] Ohiorhenuan I E, Mechler F, Purpura K P, Schmid A M, Hu Q and Victor J D, *Sparse coding and high-order correlations in fine-scale cortical networks*, 2010 *Nature* **466** 617–21
[41] Oram M W, Wiener M C, Lestienne R and Richmond B J, *Stochastic nature of precisely timed spike patterns in visual system neuronal responses*, 1999 *J. Neurophysiol.* **81** 3021–33
[42] Paninski L, Fellows M, Shoham S, Hatsopoulos N and Donoghue J, *Superlinear population encoding of dynamic hand trajectory in primary motor cortex*, 2004 *J. Neurosci.* **24** 8551–61
[43] Paninski L, *Maximum likelihood estimation of cascade point-process neural encoding models*, 2004 *Netw. Comput. Neural Syst.* **15** 243–62
[44] Parry W and Pollicott M, 1990 *Zeta Functions and the Periodic Orbit Structure of Hyperbolic Dynamics* (Paris: Société Mathématique de France) pp 187–8
[45] Peterson C and Anderson J R, *A mean field theory learning algorithm for neural network*, 1987 *Complex Syst.* **1** 995–1019
[46] Pillow J W, Shlens J, Paninski L, Sher A, Litke A M, Chichilnisky E J and Simoncelli E P, *Spatio-temporal correlations and visual signaling in a complete neuronal population*, 2008 *Nature* **454** 995–9
[47] Pillow J W, Ahmadian Y and Paninski L, *Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains*, 2011 *Neural Comput.* **23** 1–45
[48] Pillow J W, Paninski L, Uzzell V J, Simoncelli E P and Chichilnisky E J, *Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model*, 2005 *J. Neurosci.* **25** 11003–13
[49] Pollicott M and Weiss H, *Free energy as a dynamical invariant (or can you hear the shape of a potential?)*, 2003 *Commun. Math. Phys.* **240** 457–82
[50] Puchalla J L, Schneidman E, Harris R A and Berry M J, *Redundancy in the population code of the retina*, 2005 *Neuron* **46** 493–504

J. Stat. Mech. (2013) P03006

[51] Roudi Y, Nirenberg S and Latham P E, *Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't*, 2009 *PLOS Computat. Biol.* **5** e1000380
[52] Roudi Y and Hertz J, *Mean field theory for non-equilibrium network reconstruction*, 2011 *Phys. Rev. Lett.* **106** 048702
[53] Roudi Y, Tyrcha J and Hertz J A, *Ising model for neural data: model quality and approximate methods for extracting functional connectivity*, 2009 *Phys. Rev.* E **97** 051915
[54] Ruelle D, 1978 *Thermodynamic Formalism* (Reading, MA: Addison-Wesley)
[55] Schneidman E, Berry M J, Segev R and Bialek W, *Weak pairwise correlations imply strongly correlated network states in a neural population*, 2006 *Nature* **440** 1007–12
[56] Seneta E, 2006 *Non-negative Matrices and Markov Chains* (Berlin: Springer)
[57] Sessak V and Monasson R, *Small-correlation expansions for the inverse Ising problem*, 2009 *J. Phys. A: Math. Theor.* **42** 055001
[58] Shadlen M and Newsome W, *The variable discharge of cortical neurons: implications for connectivity*, 1998 *J. Neurosci.* **18** 3870–96
[59] Shenoy K V, Kaufman M T, Sahani M and Churchland M M, *A dynamical systems view of motor preparation: implications for neural prosthetic system design*, 2011 *Progr. Brain Res.: Enhancing Perform. Action Perception* **192** 33
[60] Sherrington D and Kirkpatrick S, *Solvable model of a spin-glass*, 1975 *Phys. Rev. Lett.* **35** 1792
[61] Shlens J, Field G D, Gauthier J L, Grivich M I, Petrusca D, Sher A, Litke A M and Chichilnisky E J, *The structure of multi-neuron firing patterns in primate retina*, 2006 *J. Neurosci.* **26** 8254
[62] Shlens J, Field G D, Gauthier J L, Greschner M, Sher A, Litke A M and Chichilnisky E J, *The structure of large-scale synchronized firing in primate retina*, 2009 *J. Neurosci.* **29** 5022–31
[63] Simoncelli E P, Paninski J P, Pillow J and Schwartz O, *Characterization of neural responses with stochastic stimuli*, 2004 *The Cognitive Neurosciences* (Cambridge, MA: MIT Press) p 327
[64] Singer W and Gray C M, *Visual feature integration and the temporal correlation hypothesis*, 1995 *Annu. Rev. Neurosci.* **18** 555–86
[65] Softky W R and Koch C, *The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps*, 1993 *J. Neurosci.* **13** 334–50
[66] Staude B, Grun S and Rotter S, *Higher-order correlations in non-stationary parallel spike trains: statistical modeling and inference*, 2010 *Front. Computat. Neurosci.* **4** 16
[67] Strong S P, Koberle R, de Ruyter van Steveninck R R and Bialek W, *Entropy and information in neural spike trains*, 1998 *Phys. Rev. Lett.* **80** 197–200
[68] Tanaka T, *A theory of mean field approximation*, 1998 *Adv. Neural Inform. Process. Syst.* **48** 351–60
[69] Tang A, Jackson D, Hobbs J, Chen W, Smith J L, Patel H, Prieto A, Petrusca D, Grivich M I, Sher A, Hottowy P, Dabrowski W, Litke A M and Beggs J M, *A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro*, 2008 *J. Neurosci.* **28** 505–18
[70] Theunissen F E, David S V, Singh N C, Hsu A, Vinje W E and Gallant J L, *Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli*, 2001 *Network* **12** 289–316
[71] Thouless D J, Anderson P W and Palmer R G, *Solution of solvable model of a spin glass*, 1977 *Phil. Mag.* **35** 593–601
[72] Tkacik G, Prentice J S, Balasubramanian V and Schneidman E, *Optimal population coding by noisy spiking neurons*, 2010 *Proc. Nat. Acad. Sci.* **107** 14419–24
[73] Tkacik G, Schneidman E, Berry M J II and Bialek W, *Spin glass models for a network of real neurons*, 2009 arXiv:0912.5409v1
[74] Truccolo W, Eden U T, Fellows M R, Donoghue J P and Brown E N, *A point process framework for relating neural spiking activity to spiking history, neural ensemble and extrinsic covariate effects*, 2005 *J. Neurophysiol.* **93** 1074–89
[75] Tsodyks M, Kenet T, Grinvald A and Arieli A, *Linking spontaneous activity of single cortical neurons and the underlying functional architecture*, 1999 *Science* **286** 1943–6
[76] Vaadia E, Haalman I, Abeles M, Bergman H, Prut Y, Slovin H and Aertsen A, *Dynamics of neuronal interactions in monkey cortex in relation to behavioural events*, 1995 *Nature* **373** 515–8
[77] Vasquez J C, Marre O, Palacios A G, Berry M J and Cessac B, *Gibbs distribution analysis of temporal correlation structure on multicell spike trains from retina ganglion cells*, 2012 *J. Physiol. Paris* **106** 120–7
[78] Welling M and Teh Y W, *Approximate inference in Boltzmann machines*, 2003 *Artif. Intell.* **143** 19–50
[79] Yu S, Huang D, Singer W and Nikolic D, *A small world of neuronal synchrony*, 2008 *Cereb. Cortex* **18** 2891

J. Stat. Mech. (2013) P03006