Analysis of Spike-train Statistics with Gibbs Distributions: Theory, Implementation and Applications

Juan Carlos Vasquez Betancur

PhD THESIS

prepared at
INRIA Sophia Antipolis

and presented at the
University of Nice-Sophia Antipolis
Graduate School of Information and Communication Sciences

A dissertation submitted in partial fulfillment
of the requirements for the degree of

DOCTOR OF SCIENCE
Specialized in Mathematics and its interactions

Analysis of Spike-train Statistics with Gibbs Distributions: Theory, Implementation and Applications

Juan Carlos VASQUEZ BETANCUR

Date of defence 7th March 2011.

Adviser	Pr. Bruno CESSAC, INRIA Sophia Antipolis, France
Co-Adviser	Pr. Thierry VIEVILLE, INRIA Sophia Antipolis, France
Reviewers	Pr. Simon Schultz, Imperial College of London, UK.
	Pr. Jean-Pierre Nadal, Ecole Normale Superieur, Paris, France
Examiners	Pr. Alain Destexhe, UNIC CNRS, France
	Pr. Adrian Palacios, CINV, Universidad de Valparaiso, Chile

UNIVERSITÉ NICE-SOPHIA ANTIPOLIS - UFR Sciences

École Doctorale STIC
(Sciences et Technologies de l'Information et de la Communication)

THÈSE

pour obtenir le titre de
DOCTEUR EN SCIENCES
de l'UNIVERSITÉ de Nice-Sophia Antipolis
Discipline: Mathématiques et leurs intéractions

présentée et soutenue par

Juan Carlos VASQUEZ BETANCUR

Analyse des Statisques de trains d'Impulsion: Théorie, Implémentation et Applications

Thèse dirigée par Bruno CESSAC et Thierry VIEVILLE

Date de soutenance, 7 Mars 2011

Composition du jury:

Directeur	Pr. Bruno CESSAC, INRIA Sophia Antipolis, France
Co-Directeur	Pr. Thierry VIEVILLE, INRIA Sophia Antipolis, France
Rapporteurs	Pr. Simon Schultz, Imperial College of London, UK.
	Pr. Jean-Pierre Nadal, Ecole Normale Superieur, Paris, France
Examinateurs	Pr. Alain Destexhe, UNIC CNRS, France
	Pr. Adrian Palacios, CINV, Universidad de Valparaiso, Chile

Contents

Abstract

We propose a generalization of the existing maximum entropy models used for spike trains statistics analysis. We bring a simple method to estimate Gibbs distributions, generalizing existing approaches based on Ising model or one step Markov chains to arbitrary parametric potentials. Our method enables one to take into account memory effects in dynamics. It provides directly the Kullback-Leibler divergence between the empirical statistics and the statistical model. It does not assume a specific Gibbs potential form and does not require the assumption of detailed balance. Furthermore, it enables the comparison of different statistical models and offers a control of finite-size sampling effects, inherent to empirical statistics, by using large deviations results. A numerical validation of the method is proposed. Applications to biological data of multi-electrode recordings from retina ganglion cells in animals are presented. Additionally, our formalism permits to study the evolution of the distribution of spikes caused by the variation of synaptic weights induced by synaptic plasticity. We provide an application to the analysis of synthetic data from a simulated neural network under Spike-time Dependent Plasticity $STDP$.

Résumé

Nous proposons une gÃľnÃľralisation des modÃÍles actuels, utilisÃľs en neuroscience pour l'analyse des statistiques de trains de potentiels d'action, et basÃľ sur le paradigme de maximisation de l'entropie statistique sous contraintes. Notre mÃľthode permet d'estimer des distributions de Gibbs avec un potentiel paramÃľtrique arbitraire, gÃľnÃľralisant les modÃÍles actuels $IsingouchainesdeMarkovdupremierordre$. Notre mÃľthodologie permet de tenir compte des effets de mÃľmoire dans la dynamique. Elle fournit de maniÃÍre directe la divergence de Kullback-Leibler entre le statistique empirique et le modÃÍle statistique. Elle ne prÃľsuppose pas une forme spÃľcifique du potentiel de Gibbs et ne nÃľcessite pas l'hypothÃÍse de bilan dÃľtaillÃľ En outre, elle permet la comparaison de diffÃľrents modÃÍles statistiques et offre un contrÃt'le des effets de taille finie, propres a la statistique empirique, par le biais de rÃľsultats des grandes dÃľviations. Nous avons Ãľgalement dÃľveloppÃľ un logiciel implÃľmentant cette mÃľthode et nous prÃľsentons des rÃľsultats d'application Ãă des donnÃľes biologiques issues d'enregistrements par multi-Ãľlectrode sur des cellules ganglionnaires de rÃľtines animaux. De plus, notre formalisme permet d'Ãľtudier l'Ãľvolution de la distribution des potentiels d'action lors de la variation des poids synaptiques induits par plasticitÃľ. Nous montrons une application a l'analyse de donnÃľes synthÃľtiques issues d'un rÃľseau neuronal simulÃľ soumis a de la plasticitÃľ de type STDP

Chapter 1

Introduction

Contents

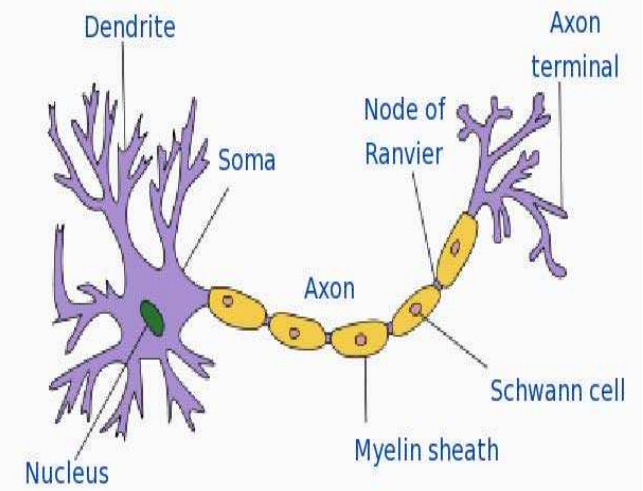

Figure 1.1: Basic scheme of a single neuron, by Quasar Jarosz at en.wikipedia

1.1 From Neurons Structure to spike trains

After recognition of the nervous networks as responsible of transport and process of the information in most forms of animal life, it was Galvani in 1791 who discovered the electric nature of nervous signals. But only until the early 20th century the anatomist Ramon y Cajal showed that those networks were made of an assembly of well defined cells called neurons (see figure 1.1), which communicate via localized junctions called synapses, where special neural interaction process might be induced electrically or chemically (shown in figure 1.1) [48].

Later, in the middle of the Twentieth century, Hodgkin and Huxley [84] explained the mechanism of creation and propagation of neuronal electric signals . The lipid cell membrane of many types of cells, and in particular neurons, contains voltage-gated ion channels which enable the cell to modify the charge imbalances between the interior and the exterior environment. As a result, a local variation of the membrane's electric potential can be generated and might also propagate along the cell as an electrical impulse called "action potential" also called "spikes" within neuroscience community (see figure 1.1). Let's see this process in more detail. In the neurons, there is usually an equilibrium electric membrane potential, called the resting potential, due to the potential difference produced by the normal concentration differences of potassium, sodium, chlor and other species of ions inside and outside the cell. A perturbation of this resting potential

[2]NIA - Alzheimer's Disease: Unraveling the Mystery. National Institutes of Health. http://www.nia.nih.gov/Alzheimers/Publications/Unraveling/

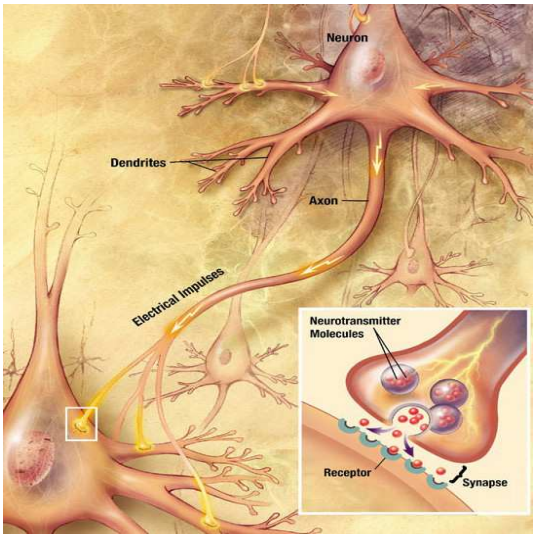

Figure 1.2: Basic scheme of an synaptic chemical transmission process between neurons.Illustration from [2]

starts when a stimuli opens some Sodium-ion channels of the cell membrane which enables an ion flux causing a depolarization. If this initial depolarization makes the electric potential to exceed a certain threshold value, more sodium channels get opened causing a bigger depolarization that reaches a peak. During the depolarizations process the sodium channels get deactivated and potassium channels get opened to redistribute the ion flux repolarizing the cell and returning it to the initial resting potential. After this process that takes about few milliseconds, most of sodium-channels remain deactivated and consequently less sensitive to stimuli for about at least 1 ms resulting in a refractory period where the generation of another spike is much difficult hence unlikely.

Once the action potential is produced the perturbation propagates from the soma through the axon until it reaches the axon terminal where it causes voltage-gated calcium channels to be opened, allowing calcium ions to enter the terminal. Calcium causes synaptic vesicles filled with neurotransmitter molecules to fuse with the membrane, releasing their contents into the synaptic cleft. The neurotransmitters travel across the synaptic cleft and activate receptors on the post-synaptic neuron, i.e., its dendrites (See figure 1.1). The synapses can be excitatory or inhibitory and will increase or decrease activity (i.e. the number of open ion channels) in the target neuron. This process is one way since axons do not posses chemo-receptors and dendrites can not release neurotransmitters. Since there are neurons having up to thousands of dendrite branches and simi-

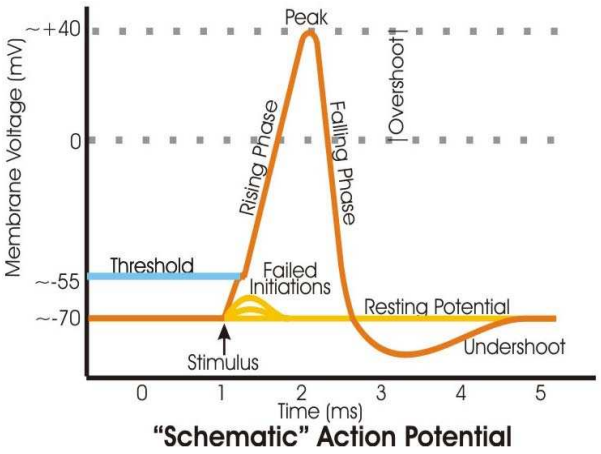

Figure 1.3: Basic scheme of an action potential, i.e, the time course of the membrane electric potential (perturbation) produced in neural cells. Artwork by Synaptidude at en.wikipedia

larly dendrites receiving thousands of synapses, almost any type of network structure is theoretically possible.

The activity of a neural network can be measured using different techniques such as Multi-Electrode Arrays (MEA). Although what is measured consist in the local variation of membrane potentials, studies often focus on the occurrence of spikes[3]. We call a "spike-train" the succession of spikes emitted by a neuron or a group of neurons. In this thesis we focus on describing neural activity by such spike trains.

1.2 Spike Train Statistics

As we have described,in previous section, neurons activity results from complex and nonlinear mechanisms (which includes also a variety of network architectures) [84, 42, 48, 70], leading to a wide variety of dynamical behaviors [42, 79]. This activity is revealed by the emission of action potentials or "spikes". While the shape of an action potential is essentially always the same for a given neuron, the succession of spikes emitted by this neuron can have a wide variety of patterns (isolated spikes, periodic spiking, bursting, tonic spiking, tonic bursting, etc ...) [90, 19, 197], depending on physiological parameters, but also on excitations coming either from other neurons or from external inputs. To characterize this activity and more precisely, to

[3]Which are extracted, in MEA recordings, by means of techniques called "Spike-sorting" that we do not discuss here. We assume that this process of extracting spikes from MEA recordings is "perfect" which is actually not the case.

understand how it is used for neural processing and encoding of information obtained from external stimuli by means of neuronal dynamics is a very active research field [154], although still much of the role of neural assemblies and their internal interactions remains unknown [145].

1.2.1 Spike trains as a “neural code”.

As a step towards the characterization of neural activity, it seems natural to consider spikes as “information quanta” or “bits” and to seek the information exchanged by neurons in the structure of spike trains. Doing this, one switches from the description of neurons in terms of membrane potential dynamics, to a description in terms of spike trains. This point of view is used, in experiments, by the analysis of *raster plots*, i.e. the activity of a neuron is represented by a mere vertical bar each time this neuron emits a spike. Though this change of description raises many questions, it is commonly admitted in the computational neuroscience community that spike trains constitute a "neural code”.

This raises however other questions:

- How is “information” encoded in a spike train: rate coding [2], temporal coding [193], rank coding [141, 50], correlation coding [95] ?
- How to measure the information content of a spike train ?

There is a wide literature dealing with these questions [133, 93, 10, 132, 5, 178, 67, 137], which are inherently related to the notion of *statistical characterizations* of spike trains, see [153, 48, 70] and references therein for a review. As a matter of fact, a prior to handle “information” in a spike train is the definition of a suitable probability distribution that matches the empirical averages obtained from measures. Thus, in some sense that we make precise in this thesis, the choice of a set of quantities to measure (observables) constrains the form of the probability characterizing the statistics of spike trains.

As a consequence, there is currently a wide debate on the canonical form of these probabilities. While Poisson statistics, based on the mere knowledge of frequency rates, are commonly used with some success in biological experiments [69, 68], other investigations evidence the role of spikes coincidence or correlations [74, 75] and some people have proposed non Poisson probabilities, such as Ising like Gibbs distributions, to interpret their data [166, 196]. This controversies have been recently pushed a step further, since [56] have found very small amount of neural correlations (of an order lower than those reported in [166]) between retina ganglion cells sharing a common visual stimuli whereas [152] have build an recurrent neural network

model capable to present low pairwise correlations despite of substantial amounts of shared input. More generally, it appears that there might be a large variety of possible neural coding/decoding strategies in both biological and artificial networks. However, determining the form of the probability characterizing the statistics of spike trains is extremely difficult even in simple neural models, whenever one attempts to characterize a collective behavior. In the present thesis, we show that Gibbs distribution are natural candidates whenever a set of quantities to measure (observables) has been prescribed. As a matter of fact Poisson distributions and Ising' like distributions are specific examples, but, certainly, do not constitute the general case.

1.2.2 Current state-of-the-art

The simultaneously recording of the activity of groups of neurons (up to several hundreds) over a dense configuration, supplies a critical database to unravel the role of specific neural assemblies. Indeed, the increase of database size is likely exponential and the kind of recording techniques quite diverse [188] (see fig 1.2.2). In complement of descriptive statistics (e.g. by means of cross-correlograms or joint peri-stimulus time histograms), somehow difficult to interpret for a large number of units (review in [21, 99]), is the specific analysis of multi-units spike-patterns, as found e.g. in [1]. This approach develops algorithms to detect common patterns in a data block, as well as performing combinatorial analysis to compute the expected probability of different kind of patterns. In particular, one might test the hypothesis that the more frequent patterns of synchronization were not produced by chance [78]. The main difficulty with such type of approaches is that they rely on a largely controversial assumption, Poissonian statistics (see [146, 147, 166]), which moreover, is a minimal statistical model largely depending on the belief that firing rates are essentially the main characteristic of spike trains.

A different approach has been proposed in [166]. They have shown that a model taking into account pairwise synchronizations between neurons in a small assembly (10-40 retinal ganglion cells) describes most (90%) of the correlation structure and of the mutual information of the block activity, and performs much better than a non-homogeneous Poissonian model. Analogous results were presented the same year in [176]. The model used by both teams is based on a probability distribution known as the Gibbs distribution of the Ising model[4] which comes from statistical physics. The parameters of this distribution relating, in neural data analysis, to the firing rate of neurons and to their probability of pairwise synchronization have to be de-

[4]These notions are defined in chapter 2

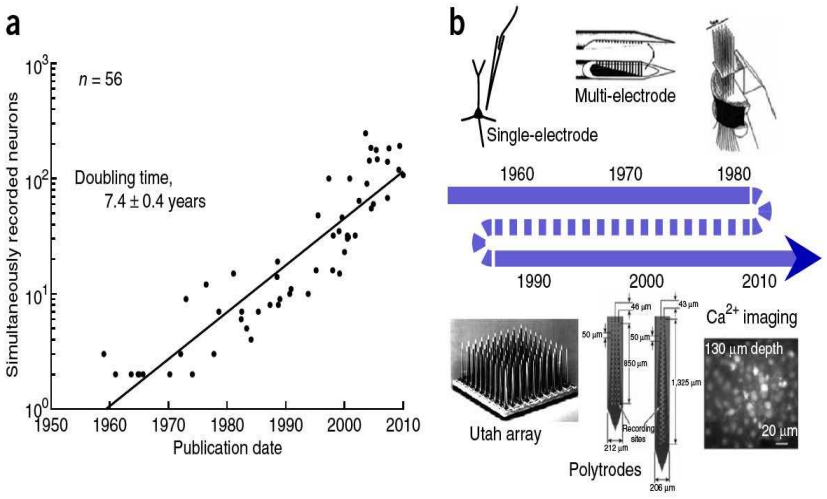

Figure 1.4: Figure and artwork reproduced from [188]. Exponential growth in the number of recorded neurons. **(a)** Examining 56 studies published over the last five decades, we found that the number of simultaneously recorded neurons doubled approximately every 7 years. **(b)** A timeline of recording technologies during this period shows the development from single-electrode recordings to multi-electrode arrays and in vivo imaging techniques.

termined from empirical data. Note that this approach has been previously presented in neuroscience, but in a slightly different and more general fashion, by [121, 108, 122] (it was referred as "log-linear models"). The use of Ising model in neural data analysis (especially as response of visual stimuli) mainly on retina ganglion cells but also on visual cortex cell has been largely exploited by several other authors [38, 138, 175, 191, 211]. In particular, it is believed by some of them [38] that the pairwise coupling terms inferred from simultaneous spikes corresponds, in the model, to effective couplings between cells. In this spirit, computing the parameters of the Ising model would provide an indirect access to neural connections. In addition, an increasing number of different theoretical and numerical developments of this idea have recently appeared. In particular, in [196], the authors propose a modified learning scheme and thanks to concepts taken from physics, such as heat capacity, explore new insights like the distribution of the underlying density of states. More recently, in [195], the authors use this framework to study the optimal population coding and the stimuli representation. Additionally, in [160, 157], the authors study and compare several approximate, but faster, estimation methods for learning the couplings and apply them on experimental and synthetic data drawing several interesting results. Finally, in a recent paper [165], it is shown convincingly that Ising model can be used to build a decoder capable of predicting, on a millisecond timescale,

the stimulus represented by a pattern of neural activity in the mouse visual cortex.

On the other hand, in [159], it has been shown that although Ising model is good for small populations, this is an artifact of the way data is binned and the small size of the system. Moreover, it might be questionable whether more general forms of Gibbs distributions (e.g. involving n-uplets of neurons) could improve the estimation and account for deviations to Ising-model ([175, 196, 134]) and provide a better understanding of the neural code from the point of view of the maximal entropy principle [92]. As a matter of fact, back to 1995, [122] already considered multi-unit synchronizations and proposed several tests to understand the statistical significance of those synchronizations and the real meaning of their corresponding value in the energy expansion. A few years later, [121] generalized this approach to arbitrary spatio-temporal spike patterns and compared this method to other existing estimators of high-order correlations or Bayesian approaches. They also introduced a method comparison based on the Neyman-Pearson hypothesis test paradigm. Though the numerical implementation they have used for their approach presented strong limitations, they have applied this methods successfully to experimental data from multi-units recordings in the pre-frontal cortex, the visual cortex of behaving monkeys, and the somato-sensory cortex of anesthetized rats. Several papers have pointed out the importance of temporal patterns of activity at the network level [114, 206, 169], and recently [191] have shown the insufficiency of Ising model to predict the temporal statistics of the neural activity. As a consequence, a few authors ([120, 3, 158]) have attempted to define time-dependent Gibbs distributions on the base of a Markovian approach (1-step time pairwise correlations). In particular, in [120] it is convincingly showed a clear increase in the accuracy of the spike train statistics characterization. Namely, this model produces a lower Jensen-Shannon Divergence, when analyzing raster data generated by a Glauber spin-glass model, but also *in vivo* multi-neurons data from cat parietal cortex in different sleep states.

To summarize, the main advantages of all these 'Ising-like' approaches are:

- (i) to be based on a widely used principle, the maximal entropy principle [92] to determine statistics from the empirical knowledge of *(ad hoc)* observables;
- (ii) to propose statistical models having a close analogy with Gibbs distributions of magnetic systems, hence disposing of several deep theoretical results and numerical methods (Markov Chain Monte-Carlo

methods, Mean-Field approximations, series expansions), resulting in a fast analysis of experimental data from large number (up to few hundreds) of neurons.

However, as we argue in this thesis, 'Ising-like' approaches present also, in their current state, several limitations.

- (i) The maximal entropy principle leads to a parametric form corresponding to choosing a finite set of *ad hoc* constraints. This only provides an approximation of the real statistics, while the distance, measured e.g. by the Kullback-Leibler divergence, between the model and the hidden distribution can be quite large [43]. Especially, Ising statistics is somewhat minimal since constraints are only imposed to first order moments and pairwise synchronizations. Therefore, it is mandatory to develop methods allowing one to handle more general forms of statistical models and to *compare* them. We propose such a method in chapter 2.

- (ii) The extension of the Ising model to more general form of Gibbs potential, including time dependent correlations, requires a proper renormalization in order to be related to the equilibrium probability of Markov chain. The normalization is not straightforward and does not reduce to the mere division of e^{ψ} by a constant, where ψ is the Gibbs potential. As emphasized in this thesis, this limitation can be easily resolved using spectral methods which, as a by-product, allows the numerical computation of the "free energy" without computing a "partition function".

- (iii) Ising-like related methods do not allow to treat in a straightforward way the time-evolution of the distribution (e.g. induced by mechanisms such as synaptic plasticity) although it can be applied to time-varying couplings, as very recently claimed in [158]. On the opposite, the analysis of the time-evolution of parametric Gibbs distributions induced by a synaptic plasticity mechanism has been addressed by us in [26] using the present formalism which may extends to time-dependent parameters. This is discussed in chapter 5.

1.3 What is this thesis about.

In this thesis, we propose a generalization of the maximal entropy models used for spike trains statistics analysis. The probability distributions obtained from the maximal entropy principle (Jaynes 1957, [92])- are usually

known as Gibbs distributions. We remark that the most usual examples of Gibbs distribution in literature (e.g Ising models) are memory-less distributions, which constitute a particular and less general subclass. Depending of the research field, Gibbs distributions (but mainly the memory-less subclass) is known also as log-linear-exponential or multinomial logarithm statistical models. Jaynes maximum entropy principle has been used on very different fields as discrimination Markov networks ([213]), inference in protein interaction networks ([196]) or identification of gene regulatory interactions ([53]). On the other hand, as pointed in ([175]), computing a maximum entropy distribution which is "memory-less" (see chapter 2 for details) is equivalent to computing the maximum likelihood estimate of the Lagrange multipliers λ_i associated to the statistical memory-less observables[5], given the data. It is also a particular case of parameter estimation on a discrete Markov random field or equivalently a Boltzmann machine.

Within this generalization of the maximal entropy models used for spike trains statistics analysis, we propose a spectral method which provides directly the "free energy density" and the Kullback-Leibler divergence between the empirical statistics and the statistical model. This method does not assume a specific potential form thus it is not limited to firing rates models, pairwise synchronizations as [166, 176, 175] or 1-step time pairwise correlations models as [120, 3], but deals with general form of Gibbs distributions, with parametric potentials corresponding to a spike n-uplets expansion, with multi-units and multi-times terms. The method is exact (in the sense that is does not involve heuristic minimization techniques) and allows us to handle correctly non-normalized potentials. It does not require the assumption of detailed balance (necessary to apply Markov Chain Monte-Carlo (MCMC) methods) and offers a control of finite-size sampling effects, inherent to empirical statistics thanks to large deviation theory results. Moreover, we perform fast and reliable estimate of quantities such as the Kullback-Leibler divergence allowing a comparison between different models, as well as the computation of standard statistical indicators, and a further analysis about convergence rate of the empirical estimation.

Preliminary applications to biological data of multi-electrode recordings from retina ganglion cells in animals are also presented. Additionally, our formalism permits to study the evolution of the distribution of spikes caused by the variation of synaptic weights induced by synaptic plasticity. We provide an application to the analysis of synthetic data from a simulated neural network under Spike-time Dependent Plasticity (STDP). Finally, we also provide perspectives concerning the limitation of the method and possible

[5]The equivalency of both set-up in the general case, namely distributions and observables with memory, have not been proved up to the authors knowledge.

ways of improvement.

1.3.1 Thesis organization and main conclusions

In this manuscript the general Gibbs distribution framework adapted for the analysis of spike trains statistics is presented in Chapter 2. A numerical implementation of it is has been provided in chapter 3 and indeed has resulted in a numerical library (more precisely, it is C++ header), freely available at `http://enas.gforge.inria.fr/`. The proposed implementation of the formalism is tested in chapter 4. Applications to biological data of multi-electrode recordings from retina ganglion cells in animals are presented also in chapter 4. Additionally, since our formalism permits to study the evolution of the distribution of spikes caused by the variation of synaptic weights induced by synaptic plasticity, we provide in chapter 5 an application to the analysis of synthetic data from a simulated neural network under Spike-time Dependent Plasticity (STDP). For the latter, we include a short discussion about bases of synaptic plasticity and a general way to integrate it to neural dynamics. When dealing with artificial neural networks we mostly consider IF neurons (leaky or generalized) which are shortly presented in chapter 5 but more general details can be founded in one of the annexed publications.

In conclusion, the major points of our work are :

1. A numerical method to analyze empirical spike trains with statistical models going beyond Ising.
2. A method to select a statistical model among several ones, by minimizing the Kullback-Leibler divergence between the empirical measure and the statistical model.
3. A framework to handle properly finite-size effects using large deviations results.
4. A way of generating artificial (stationary) spike train with an arbitrary statistics.
5. A numerical library (more precisely, it is C++ header), freely available at
   `http://enas.gforge.inria.fr/`, designed to be a plug-in to existing software such as Neuron or MVA-Spike.
6. A numerical study of different aspects of our method such as convergence rate and binning effects, using basic statistical models. .
7. A framework to study the evolution of statistical models under evolution of system parameters.

8. Applications of the framework to analysis of real and synthetic data.

1.3.2 Chapter Summary

In chapter 2 we present the framework of Gibbs Distribution from the perspective of Markov chains of higher order where the Gibbs measure corresponds to the invariant measure. We start with a short presentation of Markov chains and its pertinacy to spike trains statistics. and then we connect the Markov chains results with the notion of a Gibbs potential. Next, we discuss the parametric expansion of Gibbs distributions, we give some examples and present how to determine the statistical properties of the distribution from its parametric potential. In the following we describe the computation of a Gibbs distribution from the empirical data thanks to the Kullback-Leibler divergence criteria, detailing several aspects of this procedure, in particular the finite sample-size effects, and the methods to evaluate the goodness of fit.

In chapter 3, we propose a numerical method (included in ENAS library) based on the presented framework to parametrically estimate, and possibly compare, models for the statistics of simulated multi-cell-spike trains. Additionally, we comment briefly about several maximum entropy available tools and some neuroscience software tools for analysis of multicellular recordings and how they compare/complete our ENAS library. Finally, we include some discussion about possible improvements for the computational implementation.

In chapter 4, contains two parts: the method numerical validation and its application to artificial data. and the application to biological data. First, we perform a large battery of tests enabling us to experimentally validate the method. In particular we analyze the numerical precision of parameters estimation. Second, we generate synthetic data with a given statistics, and compare the estimation obtained using these data for several models. Moreover, we simulate a neural network and propose the estimation of the underlying Gibbs distribution parameters whose analytic form is known [25]. Third, we study numerically the behavior of estimation with finite size effects for simple statistical models. Additionally we show results on the parameters estimation from synthetic data generated by a statistical model with time-varying coefficients. The biological application starts with a discussion about biological recordings data treatment. Then we show preliminary analysis over several data sets from retina recordings in Frog, human, and salamander. These results are preliminary because they treat in most cases smaller sets of neurons, forthcoming results over larger sets -up to the memory capacity limit inherent to our implementation (discussed in chapter 3)- will be added later. Finally we show simple case study of binning effects

in the Gibbs distribution estimation is presented.

In chapter 5, we discuss the effects of Synaptic plasticity on spike train statistics. With this aim we present a short about neural dynamics as a dynamical system and later we give a brief outline of Gibbs measures in ergodic theory. Then, we develop analytical results about the evolution of the Gibbs Distributions under smooth evolution of internal parameters of the system such as synaptic weights. . We apply these results in the scope of Leaky-Integrate-and-Fire discrete-time recurrent neural networks with STDP. We also perform the estimation for several models using the data obtained from this simulated neural network under stationary dynamics after Spike-Time dependent synaptic plasticity process. In chapter 6, we present a summary of conclusions, provide some comparisons with respect to related approaches used in Neuroscience and finally we discuss the framework limitations and as well as its theoretical and numerical perspectives.

1.3.3 Alternative approaches we shall not discuss.

In this manuscript we will not comment some alternative approaches for the analysis of spike train statistics. The first we would like to mention, is a non equivalent but related approach that has been proposed recently by Globerson et al. ([73]), where the distribution will be fitted in order to minimize the mutual information between the input stimuli and the output response while agreeing with data measurements.

We also need to mention a current popular alternative methodology, the model-based maximum-likelihood approaches also known as Generalized Linear Models (GLM) where the multicellular activity is pre-processed and finally fitted as a very general Cox process. A good description and bibliography summary for the reader interested in this methodology can be found on ([189]). Recently, it have also appeared a different analysis method called Point Process History models where the spiking history is used to perform maximum likelihood of the instantaneous firing rates of each neuron [204].

Finally, from a more detailed point of view, [198] have proposed a Stochastic calculus approach that has been use to characterize neurons activity in noisy environments and find the probability distribution attached to their spike trains.

Chapter 2

Spike trains statistics from a theoretical perspective

Overview

In this chapter we present the framework of Gibbs Distribution from the perspective of Markov chains of higher order where the Gibbs measure corresponds to the invariant measure. We start with a short presentation of Markov chains and its pertinacy to spike trains statistics. and then we connect the Markov chains results with the notion of a Gibbs potential. Next, we discuss the parametric expansion of Gibbs distributions, we give some examples and present how to determine the statistical properties of the distribution from its parametric potential. In the following we describe the computation of a Gibbs distribution from the empirical data thanks to the Kullback-Leibler divergence criteria, detailing several aspects of this procedure, in particular the finite sample-size effects, and the methods to evaluate the goodness of fit. Finally, we discuss the estimation method behavior under non stationary statistics.

Contents

2.1 Spike trains statistics and Markov chains

We consider the evolution of a network of N neurons. We assume the network parameters (synaptic weights, currents, etc..) to be fixed in this context (see [28] for a discussion). This means that we assume observing a period of time where the system parameters are essentially constant. In other words, we focus here on *stationary* dynamics. This restriction is further discussed at the end of the chapter and it is the partially overcome for Gibbs distributions with slow time-varying parameter in Section 4.1.4. Nevertheless a general presentation aiming to describe non-stationary systems is given in chapter 5.

We assume that there is a minimal time scale $\delta > 0$ corresponding to the minimal resolution of the spike time, constrained by biophysics and by measurements methods [28]. As a consequence, the expression "neurons fire at time t" must be understood as "a spike occurs between t and $t+\delta$". Without loss of generality (change of time units) we set $\delta = 1$.

One associates to each neuron i a variable $\omega_i(t) = 1$ if neuron i fires at time t and $\omega_i(t) = 0$ otherwise. A "spiking pattern" is a vector $\omega(t) \stackrel{\text{def}}{=} [\omega_i(t)]_{i=0}^{N-1}$ which tells us which neurons are firing at time t. A *spike block* is a finite ordered list of spiking patterns, written:

$$[\omega]_{t_1}^{t_2} = \{\omega(t)\}_{\{t_1 \leq t \leq t_2\}},$$

where spike times have been prescribed between the times t_1 to t_2. We call a "raster plot" a bi-infinite sequence $\omega \stackrel{\text{def}}{=} \{\omega(t)\}_{t=-\infty}^{+\infty}$, of spiking patterns. Although we consider infinite sequences in the abstract setting we consider later on finite sequences. We denote $\Sigma \stackrel{\text{def}}{=} \{0,1\}^Z$ the set of raster plots.

2.1.1 Transition probabilities

The probability that a neuron emits a spike at some time t depends on the history of the neural network. However, it is impossible to know explicitly its form in the general case since it depends on the past evolution of all variables determining the neural network state. A possible simplification is to consider that this probability depends *only* on the spikes emitted in the past by the network. In this way, we are seeking a family of transition probabilities of the form $Prob\left(\omega(t) \mid [\omega]_{t-R}^{t-1}\right)$, where R is the *memory depth* of the probability i.e. how far in the past does the transition probability depend on the past spike sequence. These transition probabilities, from which all spike trains statistical properties can be deduced, are called *conditional intensity* in [94, 20, 36, 98, 203, 135, 202, 148] and they are essential to characterize the spike trains statistics.

Although it is possible to provide an example of neural network model where R is infinite [25] it is clearly desirable, for practical purposes to work with finite memory R. In this way, $Prob\left(\omega(t) \mid [\omega]_{t-R}^{t-1}\right)$ generates a Markov chain.

2.1.2 Markov chains

The properties of a Markov chain are easily expressed using matrix/vectors representation. For this purpose, we choose a symbolic representation of spike blocks of length R. For a fixed memory depth $R > 0$ there are $M = 2^{NR}$ such possible spike blocks, requiring, to be represented, NR symbols ('0"s and '1"s). Instead, we associate to each block $[\omega]_k^{k+R-1}$ an integer:

$$w_k = \sum_{t=0}^{R-1} \sum_{i=0}^{N-1} 2^{i+Nt} \omega_i(t+k). \tag{2.1}$$

We write $w_k \sim [\omega]_k^{k+R-1}$. We note:

$$\mathcal{A} \stackrel{\text{def}}{=} \{0, \ldots, M-1\}, \tag{2.2}$$

the set of values taken by the w's (space state of the Markov chain).

Now, for integer s, t such that $s \leq t$, $t - s \geq R$, a spike sequence $[\omega]_s^t = \omega(s)\,\omega(s+1)\ldots\omega(s+R-1)\ldots\omega(t)$ can be viewed as a sequence of integers $w_s, w_{s+1} \ldots w_{t-R+1}$. Clearly, this representation introduces a redundancy since successive blocks w_k, w_{k+1} have a strong overlap. But what we gain is a convenient representation of the Markov chains in terms of matrix/vectors.

We note:

$$\mathcal{H} \stackrel{\text{def}}{=} \mathbb{R}^M. \tag{2.3}$$

We focus here on homogeneous Markov chains where transition probabilities do not depend on t (stationarity). In this setting the transition probability $Prob\left(\omega(t) \mid [\omega]_{t-R}^{t-1}\right)$ is represented by a $M \times M$-matrix $\mathcal{M}$, with entries $\mathcal{M}_{w'\,w}$ such that:

$$\mathcal{M}_{w'\,w} = \begin{cases} Prob\left(\omega(t) \mid [\omega]_{t-R}^{t-1}\right), & \text{if} \quad w' \sim [\omega]_{t-R}^{t-1},\, w \sim [\omega]_{t-R+1}^{t}\,; \\ 0, & \text{otherwise.} \end{cases} \tag{2.4}$$

As a comment to this definition note that to define the matrix we have to consider all pairs w', w. But, among these M^2 pairs, only those such that w', w corresponds to consecutive blocks ($w' \sim [\omega]_{t-R}^{t-1}$, $w \sim [\omega]_{t-R+1}^{t}$) are non-zero[1]. Consequently, although $\mathcal{M}$ has M^2 entries, it is a *sparse matrix* since each line has, at most, 2^N non-zero entries.

[1]Although, for some blocks $[\omega]_{t-R}^{t}$, $Prob\left(\omega(t) \mid [\omega]_{t-R}^{t-1}\right)$ may also vanish.

In the same way, the probability of spike blocks of length R is represented by a M-dimensional vector $P \in \mathcal{H}$ with entries:

$$P_{w'} = Prob\left([\omega]_{t-R}^{t-1}\right) \quad \text{with} \quad w' \sim [\omega]_{t-R}^{t-1},$$

such that $\sum_{w' \in \mathcal{A}} P_{w'} = 1$.

Since we are dealing with stationary dynamics $Prob\left(\omega(t) \mid [\omega]_{t-R}^{t-1}\right)$ does not depend on t and we are free to choose it. From now on, we therefore take $t = 0$ so that we consider probability transitions $Prob\left(\omega(0) \mid [\omega]_{-R}^{-1}\right)$.

We now briefly summarize the main properties of Markov chains.

- *Normalization.*

$$\forall w' \in \mathcal{A}, \quad \sum_{w \in \mathcal{A}} M_{w'\,w} = 1, \tag{2.5}$$

- *Invariant probability distribution.*

  A probability distribution μ is invariant for the Markov chain if

$$\mu \;=\; \mu\, M. \tag{2.6}$$

  Note that for each entry $\mu_{w'}$ of the vector μ, it corresponds to the probability $P_{w'}$ of the event w' (which we recall is a block og length R) under this stationary distribution.

  The existence (and uniqueness) is guaranteed by the *Perron-Frobenius theorem* stated in the next section. From now on we assume that μ is unique and we consider statistics of spike blocks with respect to μ.

- *Probability of a spike sequence.*

  Taking the transition probability matrix M, the invariant measure vector μ, such that $\mu_{w'} = P_{w'}, \quad \forall w'$ and using the Chapmann-Kolmogorov equation, the probability for any sequence $[\omega]_1^n$, for any integer $n > R$ is given by

$$Prob\left([\omega]_1^n\right) = \mu_{w_1} \prod_{k=1}^{n-1} M_{w_k\, w_{k+1}}. \tag{2.7}$$

  On the opposite, for blocks of size $0 < n < R$ then we can compute the probability by the corresponding marginal distribution given by

$$Prob\left([\omega]_1^n\right) = \sum_{w \ni [\omega]_1^n} \mu(w),$$

  where the sum holds on each word w containing the block $[\omega]_1^n$.

  Due to the stationarity and the invariance of μ one obtains likewise the probability of blocks $[\omega]_k^{k+n}$ for any integer k.

2.1.3 The Perron-Frobenius theorem

We state now the Perron-Frobenius (PF) theorem [171, 66] which is a key result for our approach. Since this theorem holds in a more general context than Markov chains and since we need this general context for our purposes we state it in its general form. We consider a $M \times M$ matrix $\mathcal{L}$ with $\mathcal{L}_{w'w} \geq 0$, but *we do not assume the normalization property* (2.5). We assume $\mathcal{L}$ to be primitive, equivalently irreducible and aperiodic, i.e. $\exists n > 0$, such that, $\forall w, w'$, $\mathcal{L}^n_{w'w} > 0$ where $\mathcal{L}^n$ is the n-th power of $\mathcal{L}$. In the context of Markov chains (i.e. when $\mathcal{L}$ is normalized) this property means that there is a time $n > 0$ such that, for any pair w', w there is a path $w', w_1, \ldots, w_{n-2}, w$ in the Markov chain, of length n, joining w' and w, with positive probability[2].

Then:

Theorem 1. *$\mathcal{L}$ has a unique maximal and strictly positive eigenvalue s associated with a right eigenvector $b\rangle$ and a left eigenvector $\langle b$, with positive and bounded entries, such that $\mathcal{L}b\rangle = sb\rangle$, $\langle b\mathcal{L} = s\langle b$. Those vectors can be chosen such that $\langle b.b\rangle = 1$ where $.$ is the scalar product in $\mathcal{H}$. The remaining part of the spectrum is located in a disk in the complex plane, of radius strictly lower than s. As a consequence, for any vector v in $\mathcal{H}$ not orthogonal to $\langle b$,*

$$\frac{1}{s^n}\mathcal{L}^n v \to b\rangle\langle b\,.\,v, \tag{2.8}$$

as $n \to +\infty$.

When this matrix is *normalized* (prop. (2.5)) then the following additional properties hold.

- $s = 1$.
- $\forall w \in \mathcal{A}, \langle b_w = \alpha$ where α is a constant (which can be taken equal to 1).
- There is a unique invariant measure μ for the Markov chain (of memory/order R) whose components are given by:

$$\mu_w = b_w\rangle\langle b_w. \tag{2.9}$$

[2] The matrix $\mathcal{L}$ defined in (2.12) below is primitive by construction. But $\mathcal{L}$ is intended to provide a statistical model for a realistic network where primitivity remains a theoretical assumption. What we now is that primitivity holds for Integrate-and-Fire models with noise [25] and is likely to hold for more general neural networks models where noise renders dynamics ergodic and mixing. Note, on the opposite, that if this assumption is not fulfilled the uniqueness of the Gibbs distribution is not guaranteed. In this case, one would have a situation where statistics depend on initial conditions, which would considerably complicate the analysis, although not rendering it impossible. In this case, the statistical model would have to be estimated for different sets of initial conditions. This situation may happen for systems close to phase transitions.

This means that the probability of the event w' under the stationary distribution is equal to the product of its associate index entries w' on the left and right eigenvectors corresponding to the maximal eigenvalue ($s = 1$) of the transitition matrix.

As a consequence, the PF theorem provides the invariant distribution of the Markov chain, given directly by the left and right eigenvector associated with the eigenvalue s.

Now, we note an important property. If $\mathcal{L}$ in theorem 1 is not normalized it is always possible to associate it with the (normalized) probability transition of a Markov chain $\mathcal{M}$ by the relation :

$$\mathcal{M}_{w'w} = \mathcal{L}_{w'w} \frac{1}{s} \frac{b_w\rangle}{b_{w'}\rangle}, \tag{2.10}$$

since $\forall w' \in \mathcal{A}$, $\sum_{w\in\mathcal{A}} \mathcal{M}_{w'w} = \frac{1}{s}\frac{1}{b_{w'}\rangle} \sum_{w\in\mathcal{A}} \mathcal{L}_{w'\,w} b_w\rangle = \frac{s\, b_{w'}\rangle}{s\, b_{w'}\rangle} = 1$.

As a consequence, the probability (2.7) of a spike sequence reads:

$$Prob\left([\omega]_1^n\right) = \mu_{w_1} \frac{1}{s^n} \frac{b_{w_n}\rangle}{b_{w_1}\rangle} \prod_{k=1}^{n-1} \mathcal{L}_{w_k\, w_{k+1}}. \tag{2.11}$$

The latter equation is simply states the Chapman-Kolmogorov relation applied to a Markov chain of arbitrary order that naturally arises and which can be generated by normalization of Perron-Frobenius matrix

2.2 From Markov chain to Gibbs distributions

We now show how conditional intensities and Markov chains formalism are naturally related with Gibbs distributions.

2.2.1 Range-$R+1$ potentials

We call "potential" a function ψ which associate to a raster plot ω a real number. Some regularity conditions, associated with a sufficiently fast decay of the potential at infinity ([102]), are also required. We formally state them later in section 5.1. A potential has range $R+1$ if it is only a function of $R+1$ consecutive spiking patterns in the raster (e.g. $\psi([\omega]_{-\infty}^0) = \psi([\omega]_{-R}^0)$. Coding spikes blocks of length R with (2.1) we may write a range $R+1$-potential as matrix $\psi_{w'\,w}$ where $\psi_{w'\,w}$ is finite if $w' \sim [\omega]_{-R}^{-1}$, $w \sim [\omega]_{-R+1}^0$ and takes the value $-\infty$ otherwise.

To this potential we associate a $M \times M$ matrix $\mathcal{L}(\psi)$ given by:

$$\mathcal{L}_{w'\,w}(\psi) \;=\; e^{\psi_{w'\,w}}, \tag{2.12}$$

It is primitive by construction so it obeys Perron-Frobenius theorem. Indeed, take two symbols w', w correspond to blocks $w' \sim \alpha'_1 \dots \alpha'_R$, $w \sim \alpha_1 \dots \alpha_R$ where $\alpha'_k, \alpha_k \in \mathcal{A}, k = 1 \dots R$. Either these block overlap, say, on $l \leq R$ spiking patterns i.e. $\alpha'_{R-l} = \alpha_1, \dots, \alpha'_R = \alpha_l$. Then in l time steps one goes from w' to w. Or, they do not overlap. Then, the block $\alpha'_1 \dots \alpha'_R \alpha_1 \dots \alpha_R$ (concatenation) corresponds to a path $w_1\, w_2\, \dots w_R\, w_{R+1}\, w_{R+2}$ where $w_1 = w'$, $w_2 = \alpha'_2 \dots, \alpha'_R\, \alpha_1$, …, $w_{R+1} = \alpha'_R\, \alpha_1 \dots \alpha_{R-1}$, $w_{R+2} = w$. So there is a path going from from w to w' in $R+2$ time steps. Since $\psi_{w'\, w}$ is finite for contiguous blocks each matrix element $\mathcal{L}_{w_k\, w_{k+1}}(\psi)$ is positive and therefore the matrix element $\mathcal{L}^{R+2}_{w'\, w}$ is positive.

For s the largest eigenvalue of $\mathcal{L}$, we define:

$$P(\psi) = \log s, \tag{2.13}$$

called the "topological pressure" in the context of ergodic theory and "free energy density" in the context of statistical physics (see below for more links with statistical physics).

We say that ψ is *normalized* if $\mathcal{L}(\psi)$ is normalized. In this case, the log of this potential, with range $R+1$, corresponds to the conditional intensities of a Markov chain with a memory depth R. Moreover, $P(\psi) = 0$.

2.2.2 Gibbs distribution

The probability (2.11) of a spike sequence then reads:

$$Prob\left([\omega]_1^n\right) = \mu_{w_1} \frac{1}{e^{nP(\boldsymbol{\psi})}} \frac{b_{w_n}\rangle}{b_{w_1}\rangle} e^{\sum_{k=1}^{n-1} \psi_{w_k\, w_{k+1}}}. \tag{2.14}$$

This is an example of *a Gibbs distribution associated with the potential* ψ. We present in section 5.1 a more formal and general definition of Gibbs distribution. From now on we note μ_ψ the probability distribution associated to a Gibbs potential ψ and acting on the complete space of events, but a comment for clarity is necessary here.

Since the probabilities of spike blocks are referring to $\mu_{\boldsymbol{\psi}}$ we note $\mu_{\boldsymbol{\psi}}\left([\omega]_1^n\right)$ instead of $Prob\left([\omega]_1^n\right)$. Nevetheless, when refering explicitely to events of length R, i.e $[\omega]_0^{R-1}$, we prefer to maintain use of the symbols w and w' which were previously associated to them since eq.2.12, and also we propose to drop the potential dependence ψ wherever there is no place for confussion regarding the associated potential, thus, $\mu_{\boldsymbol{\psi}}\left([\omega]_0^{R-1}\right) \equiv \mu_w$. Indeed this simplifications are already introduced in eq.(2.14)

To make a connection with the classical setting for Gibbs distributions let us introduce another potential (formal Hamiltonian)

$$H_{w'\, w} = \psi_{w'\, w} + \log(b_w\rangle),$$

and a "conditional" partition function

$$Z(w') = e^{P(\psi)} b_{w'}\rangle.$$

Note that

$$Z(w') = \sum_{w \in \mathcal{A}} e^{H_{w'\,w}} \tag{2.15}$$

since $\sum_{w\in\mathcal{A}} e^{H_{w'\,w}} = \sum_{w\in\mathcal{A}} L_{w'w} b_w\rangle = s b_{w'}\rangle$ from the PF theorem.

Then, for two successive blocs:

$$Prob(w', w) \;=\; \mu_{w'} \frac{1}{s} \frac{b_w\rangle}{b_{w'}\rangle} e^{\psi_{w'\,w}} \;=\; \mu_{w'}\, \frac{1}{Z(w')} e^{H_{w'\,w}},$$

so that:

$$\mathcal{M}_{w'\,w} \;=\; Prob(w\,|w') \;=\; \frac{1}{Z(w')} e^{H_{w'\,w}}, \tag{2.16}$$

which has the classical form "$\frac{1}{Z} e^{-\beta H}$" but where Z depends[3] on w' explaining the terminology "conditional partition function" (compare with eq. (1) in ref [120]). Let us insist that $Z(w')$ depends on w'. Moreover, its computation requires the knowledge of the eigenvalue s *and* eigenvector $b\rangle$ even when using the sum form (2.15) since H depends on $b\rangle$.

Remarks. We note that a normalization factor described by $\sum_{w'\in\mathcal{A},\, w\in\mathcal{A}} e^{\psi_{w'\,w}}$ is not adequate (although there is a notable exception, discussed in section 2.3.5: the Ising model, and more generally memory-less processes). Indeed, such a normalization factor would correspond to a *joint* probability[4] while $e^{\psi_{w'\,w}}$ is related to a *conditional probability*. A formal normalization requires the knowledge of the largest eigenvalue *and* of the corresponding right eigenvector. with the partition function given by (2.15), although its computation is useless in our setting since s provides directly the "free energy density" $P(\psi) = \log s$ from which all required information is obtained, as exposed below, hence becoming one of the main advantages of the proposed framework.

[3] A similar situation arises in statistical physics for systems with boundaries where the partition function depends on the boundary.

[4] The joint probability in our case is

$$Prob\,(w',\, w) = \mu_{w'} \mathcal{M}_{w',\,w} = \mu_{w'} \mathcal{L}_{w'w}\, \frac{1}{s}\, \frac{b_w\rangle}{b_{w'}\rangle} = b_{w'}\rangle \langle b_{w'} e^{\psi_{w'w}}\, \frac{1}{s}\, \frac{b_w\rangle}{b_{w'}\rangle} = \frac{1}{s} b_w\rangle\, \langle b_{w'} e^{\psi_{w'w}},$$

which is correctly normalized since

$$\sum_{w'\in\mathcal{A},\, w\in\mathcal{A}} \frac{1}{s} b_w\rangle\, \langle b_{w'} e^{\psi_{w'w}} = \frac{1}{s} \sum_{w\in\mathcal{A}} b_w\rangle \sum_{w'\in\mathcal{A}} \langle b_{w'} e^{\psi_{w'w}} = \frac{1}{s} \sum_{w\in\mathcal{A}} b_w\rangle s \langle b_w = 1.$$

2.3 From parametric Gibbs potential to Markov chains

In the previous section we have seen how starting from conditional intensities $Prob\left(\omega(t) \mid [\omega]_{t-R}^{t-1}\right)$ one can construct a Markov chain whose invariant probability is a Gibbs distribution μ_ψ. However, in real situations neither $Prob\left(\omega(t) \mid [\omega]_{t-R}^{t-1}\right)$ nor μ_ψ are known. As a consequence, one is lead to extrapolate them from empirical measurement. We now show how this can be done by starting from a generic guess form for the Gibbs potential. Here, we start from a Gibbs potential and infer the form of the Markov chain and its related statistical properties, that are then compared to the statistics of empirical data. For this we need to consider a generic form of range-$R+1$ potentials.

2.3.1 Parametric forms of range-$R+1$ potentials

A natural approach consists of seeking a generic and parametric form of potentials decomposing as a linear combination of characteristic events. Dealing with spike trains natural characteristic events[5] have the form "neuron i_1 fires at time t_1, neuron i_2 at time t_2, ... neuron i_n fires at time t_n" (spike-uplets). To such an event one can associate a function $\omega \to \omega_{i_1}(t_1)\dots\omega_{i_n}(t_n)$ which takes values in $\{0,1\}$ and is 1 if and only if this event is realized. We call an *order-n monomial* a product $\omega_{i_1}(t_1)\dots\omega_{i_n}(t_n)$, where $0 \le i_1 \le i_2 \le \dots \le i_n \le N-1$, $-\infty \le t_1 \le t_2 \le \dots \le t_n \le +\infty$ and such that there is no repeated pair (i_k, t_k), $k = 1 \dots n$.

Monomials constitute a natural basis for Gibbs potential, in the sense that any range-$R+1$ potential can be decomposed as:

$$\psi_{w_k, w_{k+1}} = \sum_{n=1}^{R} \sum_{(i_1,t_1),\dots,(i_n,t_n)\in\mathcal{P}(N,R),} \lambda^{(n)}_{i_1,t_{i_1},\dots,i_n,t_{i_n}} \omega_{i_1}(k+t_{i_1})\dots\omega_{i_n}(k+t_{i_n}), \tag{2.17}$$

where $\mathcal{P}(N,R)$ is the set of non repeated pairs of integers with $i \in \{0,\dots,N-1\}$ and $t_{i_l} \in \{-R,\dots,0\}$. Since we are dealing with stationary dynamics ψ is translation invariant (the $\lambda^{(n)}_{i_1,t_{i_1},\dots,i_n,t_{i_n}}$'s do not depend on k) and we may define it for $k=0$.

This form can be rigorously justified in the LIF model with noise (see chapter 5 for an example, [25]) and is nothing but a Taylor expansion of $\log(P\left[\omega(0) \mid [\omega]_{-R}^{-1}\right])$, where one collects all terms of the form $\omega_{i_1}^{k_1}(t_{i_1})\dots\omega_{i_n}^{k_n}(t_{i_n})$, for integer $k_1,\dots k_n$'s, using that $\omega_i^k(t) = \omega_i(t)$, for any

[5] Although our formalism affords the consideration of events of a different kind, such as the appearance of a specific block.

$k > 0$ and any i, t. In this case the coefficients $\lambda^{(n)}_{i_1,t_{i_1},\ldots,i_n,t_{i_n}}$ are explicit functions of the network parameters (e. g. synaptic weights). They are also determined as Lagrange multipliers in the Maximal Entropy approach (see section 2.5.3).

2.3.2 Further approximations.

The potential (2.17) remains quite cumbersome since the number of terms in (2.18) explodes combinatorially as N, R growth. As a consequence, one is typically lead to consider parametric forms where monomials have been removed (or, sometimes, added) in the expansion. This constitutes a coarser approximation to the exact potential, but more tractable from the numerical or empirical point of view. To alleviate notations we write, in the rest of paper, the parametric potential in the form:

$$\psi_{w'\,w} = \sum_{l=1}^{L} \lambda_l \phi_l(w',\, w), \tag{2.18}$$

where ϕ_l's are monomials of range $\leq R+1$. If ϕ_l is the monomial $\omega_{i_1}(t_1)\ldots\omega_{i_n}(t_n)$, $-R \leq t_k \leq 0$, $k = 1\ldots n$, then $\phi_l(w',\, w) = 1$ if $w' \sim [\omega]^{-1}_{-R}$, $w \sim [\omega]^{0}_{-(R+1)}$ with $\omega_{i_k}(t_k) = 1$, $k = 1\ldots n$ and is zero otherwise. In other words, $\phi_l(w',\, w) = 1$ if and only if the block $w'w$ contains the event "neuron i_1 fires at time t_1, $\ldots$, neuron i_n fires at time t_n". The choice of the parametric form (2.18) defines what we call a "statistical model", namely a Gibbs distribution.

2.3.3 Generic parametric forms : a new computation paradigm

The idea is now to start from a parametric form of a potential and to infer from it the statistical properties of the corresponding Markov chain that we will be compared to empirical statistics. However, when switching from the potential (2.17), which is the polynomial expansion of the log of the conditional intensity, to a generic parametric form (2.18), one introduces several biases. First, one may add terms which are not in the original potential. Second, since (2.17) is the $\log$ of a probability, it is normalized which certainly imposes specific relations between the coefficients $\lambda^{(l)}_{i_1,n_{i_1},\ldots,i_l,n_{i_l}}$. On the opposite, in (2.18), the coefficients λ_l are arbitrary and do not satisfy the normalization constraint.

Nevertheless, the Perron-Frobenius give us exact relations to go from an arbitrary parametric potential to the normalized potential of Markov chain. This is the key aspect that conduces to a the change of paradigm for the parametric estimation from Markov chain sampling to matrix computations

(see section (3.1.3) for the numerical development). The price to pay is to compute the largest eigenvalue (and related eigenvectors) of a $2^{NR} \times 2^{NR}$ matrix. While the dimension increases exponentially fast with the range of the potential and the number of neurons, note that we are dealing with a *sparse matrix* with at most 2^N non-zero entries per line. Then, the largest eigenvalue and corresponding eigenvectors are easily obtained by a power iteration algorithm (though, Krylov subspace iterations methods are also available [52], see discussion in chapter 6.).

2.3.4 Cohomology

. Before giving some examples, we need to state explicitly that performing a time shift by adding an identical constant value to the time arguments of monomials that compose any potential, will result in an equivalent potential. By equivalent we mean here "cohomologous". Two potentials are cohomologous if there exists a positive function $b(w)$ such that $\boldsymbol{\psi}'(w) = \boldsymbol{\psi}(w) - \log(b(\sigma w)) + \log(b(w)) - P[\boldsymbol{\psi}]$ and it can be shown that cohomologous potentials have the same Gibbs measure [102]. Although it might be cumbersome to show how this happens by merely comparing the new structure of the PF matrix, it is straightforward to interpret (see section 2.5.3, since this means that the Gibbs measure is constrained by the average value of the observable (i.e, the monomials), thus the obtained stationary measure should not depend on a time shift in the monomial itself as this operation remains the monomial average unchanged. Consequently, we are allowed to write potentials with time delays centered at zero (eg. eq.2.23) but it do not be confused/related to a given constraint in the future time of the sequence.

2.3.5 Examples of potentials

Range-1 potentials. The easiest examples are range-1 potentials which correspond to a Markov chain without memory, where therefore the spiking pattern $w_t = \boldsymbol{\omega}(t)$ is independent of $w_{t-1} = \boldsymbol{\omega}(t-1)$. In this case, $\psi_{w'\,w} \equiv \psi_{w'}$ and $\mathcal{L}_{w'\,w} = e^{\psi_{w'}}$ does not depend on w. As a consequence, all lines of $\mathcal{L}$are linearly dependent which implies that there are $N-1$ 0-eigenvalues while the largest eigenvalue is $s = \sum_{w' \in \mathcal{A}} e^{\psi_{w'}} \stackrel{\text{def}}{=} Z$. The corresponding left eigenvector is $\langle b = (1, \ldots, 1)$ and the right eigenvector is $b_{w'}\rangle = \frac{e^{\psi_{w'}}}{Z}$, so that $\langle b.b\rangle = 1$. Thus, the Gibbs distribution is, according to (2.9), $\mu_{\boldsymbol{\psi}_{w'}} = \frac{e^{\psi_{w'}}}{Z}$. Here, we have the classical form of a Gibbs distribution with a constant partition function Z and the normalization only consists of removing the log of Z to ψ.

Bernoulli potentials. This is the simplest case of range-1 potential where:

$$\psi_{w'} = \sum_{i=0}^{N-1} \lambda_i \omega_i(0). \tag{2.19}$$

Then, $e^{\psi_{w'}} = \prod_{i=0}^{N-1} e^{\lambda_i \omega_i(0)}$ and $Z = \prod_{i=0}^{N-1}(1 + e^{\lambda_i})$. Therefore, the corresponding Gibbs distribution provides a statistical model where neurons are independent, and where $Prob\left(\omega_i(0) = 1\right) = \frac{e^{\lambda_i}}{1+e^{\lambda_i}}$. Hence, the parameter λ_i is directly related to the so-called firing rate, $r_i = Prob\left(\omega_i(0) = 1\right)$.

"Ising" like potentials. This type of range-1 potential has been used by Schneidman and collaborators in [166]. It reads, in our notations,

$$\psi_{w'} = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \lambda_{ij} \omega_i(0) \omega_j(0). \tag{2.20}$$

Note that this authors use instead the spin state variables $\sigma_i \in \{-1, +1\}; i = 1, ..., N$, as they are the natural variables for spin systems. This representation differs simply by a change of variables, namely $\omega_i(0) = \frac{\sigma_i+1}{2}$, not affecting our conclusions. The variable change reflects in a constant factor added to the transformed Gibbs potential while our $\lambda_{i,ij}$ parameters can be expressed as linear combinations of the new set of parameters which are usually denoted by h_i and J_{ij} respectively. The corresponding Gibbs distribution provides a statistical model where synchronous pairwise correlations between neurons are taken into account, but neither higher order spatial correlations nor other time correlations are considered. As a consequence, the corresponding "Markov chain" is memory-less. Therefore the obtained probability distribution acts on single time events, and in this case eq. 2.14 simplifies largely

$$Prob\left([\omega]_1^1\right) = \frac{1}{e^{P(\boldsymbol{\psi})}} e^{\sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \lambda_{ij} \omega_i(0) \omega_j(0).}. \tag{2.21}$$

Since the Ising model is well known in statistical physics the analysis of spike statistics with this type of potential benefits from a diversity of methods leading to really efficient algorithms ([157, 160, 38]).

Range-$k+1$ potentials. There are several compelling ideas that suggest the use of parametric forms of potentials with memory larger than 1. First, there is the analytical result discussed in section 5.2.1 of the Gibbs potential (5.23) for a discrete-time LIF model (5.22), whose explicit form contains memory dependent terms. Additionally, there is the fact that range-1 potentials, as for instance "Ising" models do not account for the natural synaptic

propagation delays at least at the natural sampling rate scale (the one determined by the neuron's refractory period). Note that artificial synchronization effects might appear due to binning (see section 4.2.1).

Pairwise range-2 potentials. A natural extension of the previous cases is to consider Markov chains with memory 1. The potential has the form:

$$\psi(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \sum_{\tau=-1}^{0} \lambda_{ij\tau} \omega_i(0)\omega_j(\tau). \tag{2.22}$$

This case has been investigated in [120] using the computation of the conditional partition function (2.15) with a detailed balance approximation.

Pairwise Time-Dependent-k potentials with rates (RPTD-k). An easy generalization of the previous examples is:

$$\psi(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \sum_{\tau=-k}^{k} \lambda_{ij\tau} \omega_i(0)\omega_j(\tau), \tag{2.23}$$

called *Pairwise Time-Dependent k (RPTD-k) with Rates* potentials in the sequel.

Pairwise Time-Dependent k (PTD-k) potentials.

A variation of (2.23) is to avoid the explicit constraints associated to firing rates :

$$\psi(\omega) = \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \sum_{\tau=-k}^{k} \lambda_{ij\tau} \omega_i(0)\omega_j(\tau), \tag{2.24}$$

called *Pairwise Time-Dependent k (PTD-k)* potentials in the sequel.

2.4 Determining the statistical properties of a Gibbs distribution with parametric potentials.

In this section we deal with the following problem: Given a parametric potential how can one infer the main characteristics of the corresponding Gibbs distribution?

2.4.1 Computing averages of monomials

Denote $\mu_{\boldsymbol{\psi}}[\phi_l]$ the average of ϕ_l with respect to $\mu_{\boldsymbol{\psi}}$. Since

$$\mu_{\boldsymbol{\psi}}[\phi_l] = \sum_{w',\, w \in \mathcal{A}} \mu_{\boldsymbol{\psi}}(w', w)\, \phi_l(w',\, w)$$

one obtains:

$$\mu_{\boldsymbol{\psi}}[\phi_l] = \sum_{w', w \in \mathcal{A}} \mu_{\boldsymbol{\psi}_{w'}} M_{w'\, w}\, \phi_l(w',\, w), \tag{2.25}$$

where $\mu_{\boldsymbol{\psi}_{w'}}$ is given by (2.9) and $M_{w'\,w}$ by (2.10). This provides a fast way to compute $\mu_{\boldsymbol{\psi}}[\phi_l]$.

2.4.2 The topological pressure.

The PF theorem gives a direct access to the topological pressure $P(\boldsymbol{\psi})$ which is the logarithm of the leading eigenvalue s, easily obtained by a power method (see eq. (5.11)). In the case of range-$R+1$ potentials (2.18) where the topological pressure $P(\boldsymbol{\psi})$ becomes a function of the parameters $\boldsymbol{\lambda} = (\lambda_l)_{l=1}^{L}$, we write $P(\boldsymbol{\lambda})$. One can show that the topological pressure is the generating function for the cumulants of the monomials ϕ_l:

$$\frac{\partial P(\boldsymbol{\lambda})}{\partial \lambda_l} = \mu_{\boldsymbol{\psi}}[\phi_l]. \tag{2.26}$$

Higher order cumulants are obtained likewise by successive derivations. Especially, second order moments related to the central limit theorem obeyed by Gibbs distributions [17, 102] are obtained by second order derivatives. As a consequence of this last property, the topological pressure's Hessian is positive and the topological pressure is *convex* with respect to $\boldsymbol{\lambda}$.

2.4.3 The Maximal Entropy (MaxEnt) principle.

Gibbs distributions obeys the following variational principle (see section 5.1):

$$P(\psi) \stackrel{\text{def}}{=} h(\mu_\psi) + \mu_\psi(\psi) = \sup_{\mu \in m^{(inv)}(\Sigma)} h(\mu) + \mu(\psi), \tag{2.27}$$

where $m^{(inv)}(\Sigma)$ is the set of invariant probability measures on Σ, the set of raster plots, and where $h(\mu_\psi)$ is the statistical entropy[6]. When $\mu(\psi)$ is imposed (e.g. by experimental measurement), this corresponds to finding a probability that maximizes the statistical entropy under constraints [92]. The value of the maximum is the topological pressure.

[6] This is:

$$h\left[\mu\right] = \lim_{n \to +\infty} \frac{h^{(n)}\left[\mu\right]}{n}, \tag{2.28}$$

where

$$h^{(n)}\left[\mu\right] = -\sum_{\omega \in \Sigma^{(n)}} \mu\left([\omega]_0^{n-1}\right) \log \mu\left([\omega]_0^{n-1}\right), \tag{2.29}$$

$\Sigma^{(n)}$ being the set of admissible sequences of length n. This quantity provides the exponential rate of growth of admissible blocks having a positive probability under μ, as n growths. It is positive for chaotic system and it is zero for periodic systems.

2.4.4 Computing the entropy

From the previous expression one obtains, for a parametric potential:

$$h\left[\mu_{\boldsymbol{\psi}}\right] = P(\boldsymbol{\lambda}) - \sum_{l} \lambda_l \mu_{\boldsymbol{\psi}}\left[\phi_l\right]. \tag{2.30}$$

2.4.5 Comparing several Gibbs statistical models.

The choice of a potential (2.18), i.e. the choice of a set of observables, fixes a statistical model for the statistics of spike trains. Clearly, there are many choices of potentials and one needs to propose a criterion to compare them. The Kullback-Leibler divergence,

$$d(\nu, \mu_{\boldsymbol{\psi}}) = \limsup_{n\to\infty} \frac{1}{n} \sum_{[\omega]_0^{n-1}} \nu\left([\omega]_0^{n-1}\right) \log\left[\frac{\nu\left([\omega]_0^{n-1}\right)}{\mu_{\boldsymbol{\psi}}\left([\omega]_0^{n-1}\right)}\right], \tag{2.31}$$

where ν and $\mu_{\boldsymbol{\psi}}$ are two absolutely continuous invariant probability measures, provides some notion of asymmetric "distance" between $\mu_{\boldsymbol{\psi}}$ and ν.

The computation of $d(\nu, \mu)$ is delicate but, in the present context, the following holds. For ν an invariant measure and $\mu_{\boldsymbol{\psi}}$ a Gibbs measure with a potential ψ, both defined on the same set of sequences Σ, one has [17, 161, 102, 31]:

$$d\left(\nu, \mu_{\boldsymbol{\psi}}\right) = P(\boldsymbol{\psi}) - \nu(\psi) - h(\nu). \tag{2.32}$$

This is the key of the algorithm that we have developed.

2.5 Computing the Gibbs distribution from empirical data.

2.5.1 Empirical Averaging

Assume now that we observe the spike trains generated by the neural network. We want to extract from these observations information about the set of monomials ϕ_l constituting the potential and the corresponding coefficients λ_l.

Typically, one observes, from $\mathcal{N}$ repetitions of the same experiment, i.e. submitting the system to the same conditions, $\mathcal{N}$ raster plots $\omega^{(m)}, m = 1 \ldots \mathcal{N}$ on a finite time horizon of length T. These are the basic data from which we want to extrapolate the Gibbs distribution. The key object for this is the *empirical* measure. For a fixed $\mathcal{N}$ (number of observations) and a fixed

T (time length of the observed spike train), the *empirical average* of a function $f : \Sigma \to R$ is:

$$\bar{f}^{(\mathcal{N},T)} = \frac{1}{\mathcal{N}T} \sum_{m=1}^{\mathcal{N}} \sum_{t=1}^{T} f(\sigma^t \omega^{(m)}), \tag{2.33}$$

where the left shift σ^t denotes the time evolution of the raster plot, namely, it shifts the raster left-wise (one time step forward). This notation is compact and well adapted to the next developments than the classical formula, reading, e.g., for firing rates $\frac{1}{\mathcal{N}T} \sum_{m=1}^{\mathcal{N}} \sum_{t=1}^{T} f(\omega^{(m)}(t))$.

Typical examples are $f(\omega) = \omega_i(0)$ in which case the empirical average of f is the firing rate[7] of neuron i; $f(\omega) = \omega_i(0)\omega_j(0)$ then the empirical average of f measures the estimated probability of spike coincidence for neuron j and i; $f(\omega) = \omega_i(\tau)\omega_j(0)$ then the empirical average of f measures the estimated probability of the event "neuron j fires and neuron i fires τ time step later" (or sooner according to the sign of τ).

The empirical measure is the probability distribution $\pi^{(T)}$ such that, for any function[8] $f : \Sigma \to R$,

$$\pi^{(T)}(f) = \bar{f}^{(\mathcal{N},T)}. \tag{2.34}$$

Equivalently, the empirical probability of a spike block $[\omega]_{t_1}^{t_2}$ is given by:

$$\pi^{(T)}\left([\omega]_{t_1}^{t_2}\right) = \frac{1}{\mathcal{N}T} \sum_{m=1}^{\mathcal{N}} \sum_{t=1}^{T} \chi_{[\omega]_{t_1}^{t_2}}(\sigma^t \omega^{(m)}), \tag{2.35}$$

where $\chi_{[\omega]_{t_1}^{t_2}}$ is the indicatrix function of the block $[\omega]_{t_1}^{t_2}$ so that $\sum_{t=1}^{T} \chi_{[\omega]_{t_1}^{t_2}}(\sigma^t \omega^{(m)})$ simply counts the number of occurrences of the block $[\omega]_{t_1}^{t_2}$ in the empirical raster $\omega^{(m)}$.

2.5.2 Estimating the potential from empirical average

The empirical measure is what we get from experiments while it is assumed that spike statistics is governed by a(n) (hidden) Gibbs distribution μ_{ψ^*} with Gibbs potential ψ^* that we want to determine or approximate. Clearly there are infinitely many *a priori* choices for this distribution, corresponding to infinitely many a priori choices of a "guess" Gibbs potential ψ. However, the ergodic theorem (the law of large number) states that $\pi^{(T)} \to \mu_{\psi^*}$ as $T \to \infty$ μ_{ψ^*} almost-surely. Equivalently, the Kullback-Leibler divergence $d\left(\pi^{(T)}, \mu_{\psi^*}\right)$ between the empirical measure and the sought Gibbs distribution *tends to* 0 as $T \to \infty$.

[7] Recall that we assume dynamics is stationary so rates do not depend on time.

[8] In fact, it is sufficient here to consider monomials.

Since we are dealing with finite samples the best that we can expect is to find a Gibbs distribution μ_ψ which *minimizes* this divergence. This is the core of our approach. Indeed, using[9] eq. (2.32) :

$$d(\pi^{(T)}, \mu_{\boldsymbol{\psi}}) = P(\boldsymbol{\psi}) - \pi^{(T)}(\psi) - h(\pi^{(T)}), \tag{2.36}$$

for any Gibbs potential ψ. Now, the hidden Gibbs potential ψ^* is such that this distance is minimal among all possible choices of Gibbs potentials. The advantage is that this quantity can be numerically estimated, since for a given choice of ψ the topological pressure is known from the Perron-Frobenius theorem, while $\pi^{(T)}(\psi)$ is directly computable. Since $\pi^{(T)}$ is fixed by the experimental raster plot, $h(\pi^{(T)})$ is independent of the Gibbs potential, so we can equivalently minimize:

$$\tilde{h}\,[\psi] = P\,[\psi] - \pi^{(T)}(\psi), \tag{2.37}$$

without computing the entropy $h(\pi^{(T)})$.

This relation holds for any potential. In the case of a parametric potential of the form (2.18) we have to minimize

$$\tilde{h}\,[\boldsymbol{\lambda}] = P\,[\boldsymbol{\lambda}] - \sum_{l=1}^{L} \lambda_l \pi^{(T)}(\phi_l). \tag{2.38}$$

Thus, from (2.26) and (5.2),(2.33), given the parametric form, the set of λ_l's minimizing the KL divergence are given by:

$$\mu_{\boldsymbol{\psi}}\,[\phi_l] = \pi^{(T)}(\phi_l), \qquad l = 1 \ldots L. \tag{2.39}$$

Before showing why this necessary condition is also sufficient, we want to comment this result in connection with standard approaches ("Jaynes argument").

2.5.3 Inferring statistics from empirical averages of observables ("Jaynes argument") and performing model comparison.

The conditions (2.39) impose constraints on the sought Gibbs distribution. In view of the variational principle (2.27) the minimization of KL divergence *for a prescribed parametric form of the Gibbs potential* is equivalent to *maximizing the statistical entropy under the constraints (2.39)*, where the λ_l's appear as adjustable Lagrange multipliers. This is the Jaynes argument [92] commonly used to introduce Gibbs distributions in statistical physics textbooks, and also used in the funding paper of Schneidman et al. [166]. There

[9] This is an approximation because $\pi^{(T)}$ is not invariant [102]. It becomes exact as $T \to +\infty$.

is however an important subtleties that we want to outline. The Jaynes argument provides the Gibbs distribution which minimizes the KL divergence with respect to the empirical distribution *in a specific class of Gibbs potentials*. Given a parametric form for the potential it gives the set of λ_l's which minimizes the KL divergence for the set of Gibbs measures having *this form of potential* [43]. Nevertheless, the divergence can still be quite large and the corresponding parametric form can provide a poor approximation of the sought measure. So, in principle one has to minimize the KL divergence with respect to several parametric forms. This is a way to compare the statistical models. The best one is the one which minimizes (2.38), i.e. knowing if the " model"ψ_2 is significantly "better" than ψ_1, reduces to verifying:

$$\tilde{h}\left[\psi_2\right] \ll \tilde{h}\left[\psi_1\right], \tag{2.40}$$

easily computable at the implementation level, as developed below. Note that $\tilde{h}$ has the dimension of entropy. Since we compare entropies, which units are bits of information, defined in base 2, the previous comparison units is well-defined.

2.5.4 Alternatives to perform model comparison.

A previously proposed manner to compare the quality of different statistical models when describing a data sequence was used in [166, 157]. It is based on the Multi-information criteria which can be seen as a generalization of mutual information. In these works, the multi-information was applied only to memoryless models (range-1). Let us to define k the correlation order of a candidate memoryless statistical model (i.e. a range-1) as the maximal number of factors composing the largest monomial of the parametric form of the potential. Then, the multi-information $I(r,s)$ between to candidate models of correlation order r,s respectively , corresponds to their entropy difference $I(r,s) = h(\mu(\boldsymbol{\psi}_r)) - h\left(\mu(\boldsymbol{\psi}_s)\right)$. To perform the model comparison the first stage is estimating parameters of candidate models by the maximum entropy criteria under the constraints of the empirical data. Finally, noting that for a finite set of N neurons, the largest possible correlation order of an memoryless statistical model is N, we proceed to compute the multi-information ratio $\eta = \frac{I(1,s)}{I(1,N)})$ which enables one to compare which amount of the most complex model (of correlation order N) is "contained" by the model of correlation order s. This method is related with $\tilde{h}$ and eq (2.40) since from equations (2.36, 2.37, 2.38) we can write $I(r,s) = \left[\tilde{h}(\boldsymbol{\psi}_r) - d(\pi^{(T)}, \mu(\boldsymbol{\psi}_r)\right] - \left[\tilde{h}\left(\boldsymbol{\psi}_s\right) - d(\pi^{(T)}, \mu(\boldsymbol{\psi}_s))\right]$.

In a parallel work, [97], authors have use in analogous way the difference between KL divergences of different memoryless models of successive in-

creasing order as a complexity measure with the aim to differentiate various dynamical complexity levels and to detect long range interactions or emerging high order correlations. They applied it successfully on several type of finite-alphabet sequences. Finally, we remark that so far, Multi-information methods have not been extended to compare multi-unit and multi-time models (i.e, models with memory). Let us to explain why: Increasing interactions always reduce the entropy of the statistical model, hence the increasing correlation order in memoryless models defines directly a monotonic decreasing sequence S_i towards the true entropy [166]. However, establishing such a hierarchy within statistical models with memory is not straightforward.

Multi-information techniques as introduced by [166] appear as a reliable way to compare memoryless models. A more formal treatment of multi-information from the point of view of memoryless Gibbs distribution is presented in [58]. In particular in that paper, authors study how multi-information of an Ising-model on a finite lattice behaves in the thermodynamic limit (infinite lattice) and its relation to phase coexistence. However no temporal dependencies are considered.

In fact, both multi-information and Kullback-Leibler based comparison techniques need to be extended to take account of temporal dependencies in order to be able to compare models with different memory, but we remark that this is far from straightforward. In fact, up to our knowledge, there are only few theoretical steps in this direction. One of them have appeared in [7] where authors try to generalize KL divergence for capturing temporal interdependencies within the setting of Markov chains of "equal memory depth". In that work basically, instead of applying the log-likelihood ratio to marginal probabilities, the log-likelihood ratio is computed on the conditional probabilities (i.e. the transition matrix).

2.5.5 Convexity and estimation well-definability.

The topological pressure is convex with respect to $\boldsymbol{\lambda}$. As being the positive sum of two (non strictly) convex criteria $P[\psi]$ and $-\pi^{(T)}(\psi)$ in (2.38), the minimized criterion is convex. This means that the previous minimization method intrinsically converges towards a global minimum, although convergence speed depends on the "flatness" of $P[\psi]$ (see section 2.5.7).

Let us now consider the estimation of an hidden potential $\psi^* = \sum_{l=1}^{L} \lambda_l^* \phi_l$ by a test potential $\psi^{(test)} = \sum_{l=1}^{L^{(test)}} \lambda_l^{(test)} \phi_l^{(test)}$. As a consequence, we estimate ψ^* with a set of parameters $\lambda_l^{(test)}$, and the criterion (2.38) is minimized with respect to *those parameters* $\lambda_l^{(test)}$, $l = 1 \dots L^{(test)}$.

Several situations are possible. First, ψ^* and $\psi^{(test)}$ have the same set of monomials, only the λ_l's must be determined. Then, the unique minimum is reached for $\lambda_l^{(test)} = \lambda_l^*, l = 1 \dots L$. Second, $\psi^{(test)}$ contains all the monomials

of ψ^* plus additional ones (*overestimation*). Then, the $\lambda_l^{(test)}$'s corresponding to monomials in ψ converge to λ_l^* while the coefficients corresponding to additional monomials converge to 0. The third case corresponds to *underestimation*. $\psi^{(test)}$ contains less monomials than ψ^* or distinct monomials. In this case, there is still a minimum for the criterion (2.38), but it provides a statistical model (a Gibbs distribution) at *positive KL distance* from the correct potential [43]. In this case adding monomials to $\psi^{(test)}$ will eventually improve the estimation (provided their relevancy). More precisely, if for a first test potential the coefficients obtained after minimization of $\tilde{h}$ are $\lambda_l^{(test)}, l = 1 \dots L^{(test)}$ and for a second test potential they are $\lambda_l^{'(test)}, l = 1 \dots L^{'(test)}, L^{'(test)} > L^{(test)}$ then $\tilde{h}(\lambda_1^{(test)}, \dots, \lambda_{L^{(test)}}^{(test)}) \geq \tilde{h}(\lambda_1^{'(test)}, \dots, \lambda_{L^{'(test)}}^{(test)})$. Note that for the same l the coefficients $\lambda_l^{(test)}$ and $\lambda_l^{'(test)}$ can be quite different.

We remark that these different situations are not inherent to our procedure, but to the principle of finding a hidden probability by maximizing the statistical entropy under constraints, when the full set of constraints is not known. Examples of these cases are provided in section 4.1. As a matter of fact, we have therefore two strategies to estimate an hidden potential. Either starting from a minimal form of test potential (e.g. Bernoulli) and adding successive monomials (e.g. based on heuristic arguments such as "pairwise correlations do matter") to reduce the value of $\tilde{h}$. The advantage is to start from potentials with a few number coefficients, but where the knowledge of the coefficients at a given step cannot be used at the next step, and where one has no idea on "how far" we are from the right measure. The other strategy, we call "model decimation" consists of starting from the largest possible potential with range $R+1$ (see the next paragraph for more details about memory estimation). . In this case it is guarantee that the test potential is at the minimal distance from the sought one, in the set of range-$R+1$ potentials, while the minimization will remove irrelevant monomials (their coefficient vanishes in the estimation). The drawback is that one has to start from a large number of "effective monomials" L_{eff} (more precisely [10], $L_{eff} < 2^{N(R+1)}$) which reduces the number of situations one can numerically handle. These two approaches are used in section 4.1.

2.5.6 A priori memory-range estimation

In order to alleviate the course to) dimensionality problem and the different guess potential choices, we would like to be able to extract from data

[10] In the perspective of Jaynes method only a set of non-redundant monomials is needed. In other words, some monomials corresponds to the same average constraint. For example, the terms $\omega_i(0)$ and $\omega_i(1)$ identify both the same constraint, namely the firing rate of neuron i.

such information as the memory depth of the process. Indeed, this problem of estimating the memory order of the underlying Markov chain to a given sequence, which means, in our framework, to find the the potential range, has been a well known difficult question in coding and information theory [128]. One of the current available tests is the Morvai-Weiss estimator (which includes the Peres-Shields estimator [212]) can be interpreted as sharp transition. The main idea is to observe the behavior of

$$\Delta_k = \sup_{1\leq i} \sup_{\left([\omega]^1_{-k-i+1}:\left([\omega]^0_{-k-i+1},\omega(1)\right)\right)\in\mathcal{S}_{k+i}} \left| P(\omega(1)|\,[\omega]^0_{-k+1}) - P(\omega(1)|\,[\omega]^0_{-k-i+1}) \right|.$$

where $[\omega]^1_{-k-i+1}$ is the concatenation of patterns $[\omega]^0_{-k-i+1}$ and $\omega(1)$, and $\mathcal{S}_{k+i}$ is an adequate subset of words (some conditions are required to guarantee the statistical relevancy of words in the analyzed sequence, but for a precise definition we refer the reader to the original references). In fact, this difference between blocks conditional probabilities should present a maximum for a word selected within the subset of words (block) with a length $k+i$ corresponding to the correct memory depth of the underlying process.

Besides , [30] has indicated that if the underlying sequence dynamics was given by a finite range Gibbs potential of range-$R+1$ the entropy rate difference $H_{k+1} - H_k$, where $H_k = \sum_{[\omega]_0^{k-1}} P([\omega]_0^{k-1}) \log(P([\omega]_0^{k-1}))$ should become constant after $k \geq R+1$. However, the Morvai-Weiss method is more general since it does not make explicit assumptions about the source character and might be able to find the optimal memory-range despite the fact that the source is *a priori* not be Gibbsian. In particular, [45] have used this method to estimate the "prediction suffix tree" of a given data set, which means to find the maximal length required to define a Variable Length Markov chain-VLMC describing that data.

The above methods methods might offer additional algorithmic tools that could be explored in the future. We also briefly note that variable length schemes (VLMC) might offer an alternative perspective with respect to fixed memory Markovian descriptions of symbolic sequences. In [61], presents VLMC framework to estimate general stationary sequences and explore memory properties of the sequence. The main idea of their procedure is to prune the branches of a maximal Context Tree (which is a tree structure which contains the information of every pattern present the sequence), and the prunning is done by using a Kullback-Leibler divergence criteria. However up to our knowledge, these schemes can not take benefit from the Gibbs Distribution computational/analytical results.

2.5.7 Finite sample effects and Large Deviations.

A very important issue is to determine how reliable are the Gibbs Distribution estimation procedure, and more precisely, how we this relates with the total available information, namely the amount of empirical data. In fact, in practice, recordings are performed during a finite amount of time, providing finite sample sizes. The estimation over finite samples results in fluctuations over the estimated quantities. This is a central problem, not inherent to our approach but to all statistical methods where one tries to extract statistical properties from finite empirical sample. Since T can be small in practical experiments, this problem can be circumvented by using an average over several samples (see eq. (2.33) and related comments). In the following we develop an estimation of finite sampling effects, which can be addressed by the large deviations properties of Gibbs distributions. Despite the fact that these results have not yet been implemented in our numerical library ENAS (see Chapter 3), some numerical explorations confirming the exponential decay of deviations with the sample size are presented in section 4.1.2.

Rate of Convergence for the estimation

For each observable ϕ_l, $l = 1 \dots L$, the following holds, as $T \to +\infty$ [51]:

$$\mu_{\boldsymbol{\psi}} \left\{ |\pi^{(T)}(\phi_l) - \mu_{\boldsymbol{\psi}}(\phi_l)| \geq \epsilon \right\} \sim \exp(-T I_l(\epsilon)), \tag{2.41}$$

where $I_l(x) = \sup_{\lambda_l \in R} (\lambda_l x - P[\boldsymbol{\lambda}])$, is the Legendre transform of the pressure $P[\boldsymbol{\lambda}]$.

$I_l(x)$ is the rate function of large deviation theory and it can be proven that it is concave and that $I_l(0) = 0$. It can be shown that a series expansion of $I_l(\epsilon)$ yields to

$$I_l(\epsilon) \approx -\frac{\epsilon^2}{2} \left(\left. \frac{\partial^2 P}{\partial \lambda_l^2} \right|_{\lambda_l = 0} \right)^{-1} + o(\epsilon^3) \tag{2.42}$$

Hence, for small ϵ,

$$\mu_{\boldsymbol{\psi}} \left\{ |\pi^{(T)}(\phi_l) - \mu_{\boldsymbol{\psi}}(\phi_l)| \geq \epsilon \right\} \sim C \exp\left(-\frac{\epsilon^2}{2} T \left(\left. \frac{\partial^2 P}{\partial \lambda_l^2} \right|_{\lambda_l = 0} \right)^{-1} \right), \tag{2.43}$$

which is a Gaussian distribution with variance $\sigma = \sqrt{\frac{\left. \frac{\partial^2 P}{\partial \lambda_l^2} \right|_{\lambda_l = 0}}{T}}$. This result provides an approximation of the convergence rate with respect to T but it also shows that the speed of convergence for the estimation of the l-th monomial, ϕ_l, depends on the second partial derivative (i.e, the flatness) of $P[\psi]$ with respect to λ_l. This means that the convergence rate of the estimation

algorithm might be different for each parameter but scales as the inverse square root of the sample size. Furthermore, when the Topological pressure is zero (inflexion point), we have a what is called critical point. In that case, the series expansion of eq.2.42 would require one to consider terms $o(\epsilon^3)$ and more generally, the change of convexity conduces to non-uniqueness of the Gibbs measure, as it appears in classical models spin glasses[199].

The precedent result is very important, since, once the Gibbs distribution is known, one can infer the length T of the time windows over which averages must be performed in order to obtain reliable statistics. This is of particular importance when applying statistical tests such as Neymann-Pearson for which large deviations results are available in the case of Markov chains and Gibbs distributions with finite range potentials [130]. A basic numerical study of finite size effects is presented in section 4.1.2, although this more analytical expressions are not yet numerically implemented. Also the integration with the mentioned Neymann-Pearson results are still to be done.

The Bernoulli approximation It can be shown that (cf. section 2.4.2) that the variance of monomial ϕ_l is given by

$$Var_{\boldsymbol{\psi}}(\phi_l) \doteq \mu_{\boldsymbol{\psi}}({\phi_l}^2) - \mu_{\boldsymbol{\psi}}(\phi_l)^2 = \mu_{\boldsymbol{\psi}}(\phi_l)(1-\mu_{\boldsymbol{\psi}}(\phi_l)) = \left.\frac{\partial^2 P}{\partial \lambda_l^2}\right|_{\lambda_l=0}. \tag{2.44}$$

If one assumes that the time of re-apparition of each block $\bar{w} \doteq [\omega]_0^n$ in sequence of length $T \gg n$ is larger than the correlation scale, then the probability of the event $X = \bar{w}$ at any time is a Bernoulli distributed and using the Central limit theorem for identically distributed variables it can be shown that $P_{\bar{w}}$ tends to a Gaussian distribution with standard deviation (i.e, size of fluctuations) given by

$$\sigma_{\bar{w}} = \sqrt{\frac{\mu_{\boldsymbol{\psi}}(\bar{w})(1-\mu_{\boldsymbol{\psi}}(\bar{w}))}{T}}. \tag{2.45}$$

It is clear, that the large deviation result 2.43, gives formal grounds and establish the limitations for the Bernoulli assumption which has been used by some authors [120], although a more detailed calculation is required because each block $\bar{w}$ might simultaneously (contribute to)/(be contained in) different monomials ϕ_l (recall that this degeneracy results from spike-time events not constrained by the monomial which is defined only by means of spiking events, i.e. a product of ω_i variables).

Probability of incorrect estimation

Another important large deviations property also results from the present formalism [102, 29, 51]. Assume that the sought Gibbs distribution

has potential ψ^*, and assume that we propose, as a statistical model, a Gibbs distribution with potential $\psi^{(test)} \neq \psi^*$. Now, the probability $\mu_{\psi^*}\left\{\|\pi^{(T)} - \mu_{\psi^{(test)}}\| < \epsilon\right\}$ that $\pi^{(T)}$ is ϵ-close to the "wrong" probability $\mu_{\psi^{(test)}}$ decays exponentially fast as:

$$\mu_{\psi^*}\left\{\|\pi^{(T)} - \mu_{\psi^{(test)}}\| < \epsilon\right\} \sim \exp(-T \inf_{\mu,\, \|\mu-\mu_{\psi^{(test)}}\|<\epsilon} d(\mu, \mu_{\psi^*}))). \qquad (2.46)$$

Thus, this probability decreases exponentially fast with T, with a rate given (for small ϵ) by $T\, d(\mu_{\psi^{(test)}}, \mu_{\psi^*})$. Therefore, a difference of η in the Kullback-Leibler divergences $d(\pi^{(T)}, \mu_{\psi^*})$ and $d(\pi^{(T)}, \mu_{\psi^{(test)}})$ leads to a ratio $\frac{\mu_{\psi^*}\left\{\|\pi^{(T)}-\mu_{\psi^*}\|<\epsilon\right\}}{\mu_{\psi^*}\left\{\|\pi^{(T)}-\mu_{\psi^{(test)}}\|<\epsilon\right\}}$ of order $\exp -T\eta$. Consequently, for $T \sim 10^8$ a divergence of order $\eta = 10^{-7}$ leads to a ratio of order $\exp(-10)$. Illustrations of this are given in section 4.1.

2.5.8 Other criteria for parametric estimation of distributions.

Markov chain related approaches. In the goal of approximating binary time-series sequences by Markov chains of memory r, several authors [192, 142] have considered a different estimation criteria for low-dimensional systems: the minimal Euclidean norm of the difference between the probability vector of symbolic blocks of length r (i.e., a $L_0 = 2^r$ size vector) on the target system and the corresponding probability vector on the estimated system. This means matching the blocks frequencies. This minimum is searched over the full parameter space and initial conditions space. Likewise, they have searched the minimal Frobenius norm (i.e, the matrix-natural extension of the Euclidean norm $\|M\| = \sum_{i,j}^{N} M_{ij}^2$) between the conditional probability matrix of the Markov chains (size ${L_0}^2$) in the target and the corresponding matrix in the estimated system. This means matching the conditional probabilities of blocks. Piccardi [142] has claimed that even in the case of short time sequences, this procedure still makes sense if it is seen as a sought of estimation of a "convenient coding of the available data" to "simply store its available statistics" instead of considering it as an estimation looking for the invariant distribution that describes the data.

The above approaches assume a prior knowledge of the parametric analytical form of the underlying dynamical systems, in order to generate a synthetic sequence from the estimated system to compute the blocks probabilities. However, only heuristic methods are available for both choices of criteria (i.e. matching the probability vector or transition probability matrix). Simulated Annealing techniques were used in [192] while genetic algorithms were used [142]. Although good results were obtained, heuristic

search in larger dimensional spaces becomes very soon a quite heavy computational task.

Specific Gibbs Distributions methods and tools. The K-L divergence minimization can be completed with other standard criteria for which some analytical results are available in the realm of Gibbs distributions. Fluctuations of monomial averages about their mean are Gaussian, since Gibbs distribution obey a central limit theorem with a variance controlled by the second derivative of $P(\boldsymbol{\lambda})$ (see section 2.5.7). Then, using a χ^2 test seems natural, since Cochran's theorem [39] shows that the sample variance of those fluctuations follows a scaled chi-square distribution.

Bootstrapping. This statistical method, is currently widely used for the data analysis of multicellular recordings [191, 120]. Bootstrapping is a method for estimating properties of a given estimator (e.g its variance) by measuring those properties when sampling from an approximating distribution, in particular the empirical distribution of the observed data. Assume that the set observations is issued from an independent and identically distributed population, this method can be implemented by constructing a number of "resamples" of the observed data-set having the same size by random sampling with replacement from the original data-set.

A great advantage of bootstrap is its simplicity [144]. It is straightforward to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution (e.g correlation coefficients). Additionally, it results in an appropriate way to control and check the stability of the results although it can not increase the amount of measured information. Nevertheless, this method does not provide information about fluctuations due to finite-sample. In particular it is always possible for the sampling data-set to be already biased with respect to the true distribution. Finally, we remark that the method simplicity contrast with the concerns about the important assumptions (e.g. independence of samples) made when performing bootstrap analysis since they might not apply for all cases [208, 144]

2.5.9 Other criteria for distribution comparison or 'test Statistics'.

General purpose goodness-of-fit methods. In order to compare the similarity between arbitrary probability distributions, which is known as goodness-of-fit (GOF), there exist several techniques. At the descriptive level, we propose the use of box plots that are standardly use in descriptive statistics. The box-plot method [65] is intended to graphically depict groups of numerical data through their 'five-number summaries' namely: the smallest observation (sample minimum), lower quartile (Q1), median (Q2), upper

quartile (Q3), and largest observation (sample maximum). We remark that the probabilities of spike blocks usually may differ by several orders of magnitude, depending on the block length (probabilities of shorter blocks are usually higher than those of larger blocks). Hence, in order to allow the largest blocks to contribute equally than shorter blocks, we compare the distributions of the block probability fluctuations weighted by their standard deviation, instead of using the distribution of blocks probability directly (see eq. 4.2).

On the other hand, quantitative methods to establish GOF are numerous and can be classified in families of 'test Statistics', the most important being the following:

- Power-Divergence methods (eg. Pearson-χ^2 test) [41]. This family evaluates the fit of the observed frequencies $\{O_i; i = 1..., k\}$ on a sample of length T, to expected frequencies $\{E_i; i = 1..., k\}$ for a uni-variate probability distribution whose domain has been equally split into k-non overlapping bins. The power-divergence statistics $\{I^\alpha; \alpha \in \mathcal{R}\}$ is defined as

  $$2TI^\alpha = \frac{2}{\alpha(\alpha+1)} \sum_{i=1}^{k} O_i \left(\left(\frac{O_i}{E_i} \right) - 1 \right); \quad \alpha \in \mathcal{R}$$

  It is easily seen that the Pearson's χ^2 corresponds to $\alpha = 1$.However, for all tests the value of the "test statistic" depends on how the data is binned (usually a standard of at least 5 events per bin is suggested) hence a sample size T large enough is required in order for the test to be valid.

- The Generalized Kolmogorov-Smirnov (KS) tests (eg. the KS and the Watson-Darling test). This family of tests consider the greatest vertical distance between the empirical Cumulative Distribution Functions (ECDF) ($F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \leq x)$). However, in standard form, they only applies to continuous univariate distributions although an extension to discontinuous distributions exist. Another drawback is that they tend to be more sensitive near the center of the distribution than at the tails. Finally, and perhaps the most serious limitation is that the critical region of the KS test is no longer valid if the fitting distribution compared to data is not fully specified previously (i.e. when location, scale, and shape parameters are estimated from the data set). Despite these drawbacks, Tang et al. [191] have used this test to compare the empirical distribution from biological data and a resampled [11] sequence, with the aim to use it for maximum entropy estimation of Ising models.

[11] see below a short paragraph about resampling

- Phi-Divergence statistics family (eg. Crammer-von Mises test, Lehmann (Bickel)-Rossenblat test)[35, 110]. This family basically corresponds to an integrated kernel which depends on the deviations of the ECDF (their difference) and the ECDF itself. As a result of their flexibility they appear as the most straightforward option within quantitative methods.

We would like to note that the hierarchical order within the GOF-tests (all families together) established by their increasing statistical power, given a fixed amount of sampling data, has been shown to depend on the form of the fitted distribution [187]. This fact is important to see that no test is a priori the better, since in the case of Gibbs distributions, the spiking blocks of fixed length which can be mapped into the interval $[0, 1)$ might present all variety of behaviors. In particular, the support of the measure usually concentrates on a multifractal set where the Gibbs potential is the local fractal dimension [59].

Specific Gibbs Distributions methods. Additionally, in order to discriminate between 2 Gibbs measures there is available a Neyman-Pearson criteria for the discrimination between 2 Markov chains, thanks to large deviations theory that allows to obtain a formula for Neyman-Pearson risk. [130]. Given to hypothesis $H0, H1$, an type I Error consist in failing to accept $H0$ when it is indeed true. For an hypothesis test, its probability of having an Type-I error denoted by $alpha$ is called the power of the test. There is also a Type-II error which consist in accept $H1$ when it is in fact false and its probability is denoted by β. The Neyman Pearson criteria is the optimal hypothesis test, where optimal means that given a fixed power $0 < \alpha < 1$ it is the test with the minimum β error. Negaev [130] construct explicitly a Neyman-Pearson test for discriminate between 2 different Markov distributions $P_0 and P_1$, namely, $H_i = \pi^{(T)} \sim P_i$ though its definition is cumbersome and will be here omitted. Given a fixed power α for the test, the author shows that β error follows a Large deviation principle and its asymptotic behavior in terms of the sample size T is obtained. In particular, it can be shown[12], that in order to have $\beta = 1 - \alpha$ one requires

$$T = \frac{\log(A) - \log(1 - \alpha)}{d(P_0, P1)},$$

where A is a value to be determined from the Markov chains (see ref. [130] for details) such that $\log(A) > 0$, $d(\dot{;})$ is the KL divergence between the Markov chains and clearly $\log(1 - \alpha) < 0$.

In the present manuscript we have limited our analysis to the most standard tests (diagonal representations of blocks (words) probabilities, box plots, χ^2). Examples are given in section 4.1.

[12]Details will be presented eventually elsewhere, since this is in current development

2.5.10 A short discussion about Non-Stationary data

Standard MCMC methods are able to estimate (reconstruct) the invariant measure of **reversible** Markov process satisfying that it satisfies the detailed balance condition, which in our notations writes

$$\mu_{w_1} M_{w_1\, w_2} = \mu_{w_2} M_{w_2\, w_1}; \quad \forall w_1, w_2 \ni [\omega]_0^{R-1}\,.$$

On the other hand, the Gibbs distribution framework does not make this assumption but a weaker condition of primitivity is required instead (eq.2.12). This property might enable the Gibbs distribution parametric estimation to be reliable for non-stationary data generated by certain well-behaved non-stationary processes, for instance a Gibbs distribution with preserved potential form but with slow time-varying coefficients, namely the λ parameters. Although a more detailed analytical study of this idea and the limits of this hypothesis is still to be done. Some successful numerical results with memoryless models have been already obtained and are reported in section 4.1.4. However to test more general models with memory requires further work concerning the additional question of how to adequately generate samples from higher-order non-stationary Markov chains.

In statistical physics, Gibbs distributions are considered on lattices where no spatial homogeneity is considered. This notion transposes to time-dependent processes on $\mathcal{Z}$ and leads to a notion of Gibbs distributions for non stationary processes. The most classical case considered in non-equilibrium statistical physics is the case of slowly varying parameters (adiabatic approach) but the notion of Gibbs distributions extends beyond this case. Furthermore, as will be exposed in chapter 5, the Gibbs distribution framework allows us to consider under certain assumptions, the evolution of the full statistical model (concerning not only its parametric coefficients). An application of this aspect for analyzing neural networks with synaptic plasticity is developed there.

Chapter 3

ENAS Numerical implementation for Gibbs Distributions

Overview

In this chapter , we propose a numerical method based on the presented framework aimed to parametrically estimate, and possibly compare, models for the statistics of simulated multi-cell-spike trains. This implementation is included in ENAS library, which more precisely consist in a C++ header, freely available at `http://enas.gforge.inria.fr/`, and designed to be a plug-in to existing software such as Neuron or MVA-Spike. Additionally, we comment briefly about several maximum entropy available tools and some neuroscience software tools for analysis of multicellular recordings that might compare to/complete our ENAS library. Finally, we include some discussion about possible improvements for the computational implementation.

Contents

3.1 Application: parametric statistic estimation.

Let us now discuss how the theory introduced in the previous chapter enables us to estimate, at a very general level, parametric statistics of spike trains.

We observe N neurons during a stationary period of observation T, assuming that statistics is characterized by an unknown Gibbs potential of range $R+1$. The algorithmic[1] procedure proposed here decomposes in three steps:

1. *Choosing a statistical model*, i.e. fixing the potential (2.18) (equivalently, the relevant monomials).
2. *Computing the empirical average of observables*, i.e. determine them from the raster, using eq. (2.33).
3. *Performing the parametric estimation*, i.e. use a variational approach to determine the Gibbs potential.

Let us describe and discuss these three steps, and then discuss the design choices.

3.1.1 The choices of statistical model: rate, coincidence, spiking pattern and more.

The meaning of monomials.

In order to understand the power of representation of the proposed formalism, let us start reviewing a few elements discussed at a more theoretical level in the previous section.

We start with a potential limited to a unique monomial.

- If $\psi = \omega_i(0)$, its related average value measures the firing probability or *firing rate* of neuron i;
- If $\psi(\omega) = \omega_i(0)\,\omega_j(0)$, we now measure the probability of spikes coincidence for neuron j and i, as pointed out at the biological level by, e.g, [74] and developed by [166];
- If $\psi(\omega) = \omega_i(\tau)\,\omega_j(0)$, we measure the probability of the event "neuron j fires and neuron i fires τ time step later" (or sooner according to the sign of τ); in this case the average value provides[2] the *cross-correlation* for a delay τ and the auto-correlation for $i = j$;

[1] The code is available at `http://enas.gforge.inria.fr/classGibbsPotential.html`

[2] Subtracting the firing rates of i and j.

- A step further, if, say, $\psi(\omega) = \omega_i(0)\,\omega_j(0)\,\omega_j(1)$, we now take into account triplets of spikes in a specific pattern (i.e. one spike from neuron i coinciding with two successive spikes from neuron j);

These examples illustrate the notion of "design choice": the first step of the method being to choose the "question to ask", i.e. what is to be observed over the data. In this framework, this translates in: "choosing the form of the potential". Let us enumerate a few important examples.

Taking only rate or synchronization into account: Bernoulli and Ising potentials.

Rate potential are range-1 potentials, as defined in eq. (2.19). Such models are not very interesting as such, but have two applications: they are used to calibrate and study some numerical properties of the present methods, and they are also used to compare the obtained conditional entropy with more sophisticated models.

Besides, there are the Ising potentials widely studied since independent works by Schneidman and collaborators ([166]) and Shlens and collaborators ([176]) but previously introduced in neuroscience by other authors(see [121, 3] for historical references). These potentials take in account rate and synchronization of neurons pairs, as studied in, e.g. [74]. This form is justified by the authors using the Jaynes argument.

Let us now consider potentials not yet studied (or only partially studied), up to our best knowledge, in the present literature.

Taking rate and correlations into account: RPTD-k potentials.

These potentials defined previously by eq. (2.23) constitute a key example for the present study. On one hand, the present algorithmic was developed to take not only Bernoulli or Ising-like potential into account, but a large class of statistical model, including a *general second order model* (redundant monomial being eliminated), i.e. taking rate, *auto-correlation* (parametrized by $\lambda_{i\tau}$) and *cross-correlation* (parametrized by $\lambda_{ij\tau}$) into account. Only the case $k = 1$, has been developed in the literature ([120, 3, 158]).

Being able to consider such type of model is an important challenge, because it provides a tool to analyze not only synchronization between neurons, but more general temporal relations (see e.g. [54, 74, 18] for important applications).

Let us now turn to a specific example related to the neuronal network dynamics analysis.

Taking plasticity into account: "STDP" potentials

In chapter 5 we consider Leaky-Integrate-and-Fire time-discrete neural networks with Spike-Time Dependent Plasticity of type:

$$W'_{ij} = \epsilon \left[r_d W_{ij} + \frac{1}{T} \sum_{t=T_s}^{T+T_s} \omega_j(t) \sum_{u=-T_s}^{T_s} f(u)\, \omega_i(t+u) \right], \tag{3.1}$$

where W_{ij} is the synaptic weight from neuron j to neuron i, $-1 < r_d < 0$ a term corresponding to passive LTD, T a large time, corresponding to averaging spike activity for the synaptic weights update, and,

$$f(x) = \begin{cases} A_- e^{\frac{x}{\tau_-}}, & x < 0, \quad A_- < 0; \\ A_+ e^{-\frac{x}{\tau_+}}, & x > 0, \quad A_+ > 0; \\ 0, & x = 0; \end{cases}$$

with $A_- < 0$ and $A_+ > 0$, is the STDP function as derived by Bi and Poo [12]. $T_s \stackrel{\text{def}}{=} 2\max(\tau_+, \tau_-)$ is a characteristic time scale. We argue that this synaptic weights adaptation rule produces, when it has converged, is described by a Gibbs distribution with potential approached by:

$$\psi(\omega) = \sum_{i=0}^{N} \lambda_i^{(1)} \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \lambda_{ij}^{(2)} \sum_{u=-T_s}^{T_s} f(u) \omega_i(0) \omega_j(u). \tag{3.2}$$

The general case: Typical number of observed neurons and statistics range.

The previous piece of theory allows us to take any statistics of memory R, among any set of N neurons into account. At the numerical level, the situation is not that simple, since it appears, as detailed in the two next sections, that both the memory storage and computation load are in $O(2^{NR})$. Hopefully, we are going to see that estimation algorithms are rather efficient and lead to a complexity smaller than $O(2^{NR})$.

It is clear that the present limitation is *intrinsic* to the problem, since we have *at least*, for a statistics of memory R, to count the number of occurrences of blocks of N neurons of size R, and there are (at most) 2^{NR} of them. Fastest implementations must be based on the *partial* observation of only a subset of, e.g., the most preeminent occurrences.

Quantitatively, we consider "small" values of N and R, typically a number of neurons equal to $N \in \{1, \simeq 8\}$, and Markov chain of range $R = \{1, \simeq 16\}$, in order to manipulate quantities of dimension $N \leq 8$, and $R \leq 16$, and such that $N(R+1) \leq 18$. The handling of larger neural ensembles and/or ranges will require an extension of the current implementation, using par-

allel computing algorithms, sparse matrix storage techniques and/or distributed memory

3.1.2 Computing the empirical measure: prefix-tree construction.

For one sample ($\mathcal{N} = 1$) the empirical probability (5.2) of the block $[\omega]_{-D}^{t}$, $-D < t \leq 0$ is given by

$$\pi^{(T)}([\omega]_{-D}^{t}) = \frac{\#\,[\omega]_{-D}^{t}}{T}.$$

thus obtained counting the number of occurrences $\#\,[\omega]_{-D}^{t}, -D < t \leq 0$ of the block $[\omega]_{-D}^{t}$ in the sequence $[\omega]_{-T}^{0}$. Since we assume that dynamics is stationary we have, $\pi^{(T)}([\omega]_{-D}^{t}) = \pi^{(T)}([\omega]_{0}^{t+D})$.

We observe that the data structure size has to be of order $O(2^{NR})$ (lower if the distribution is sparse), but does not depends on T. Since many distributions are sparse, it is important to use a sparse data structure, without storing explicitly blocks of occurrence zero.

Furthermore, we have to study the distribution at several ranges R and it is important to be able to factorize these operations. This means counting in one pass, and in a unique data structure, block occurrences of different ranges.

The chosen data structure is a tree of depth $R + 1$ and degree 2^N. The nodes at depth D count the number of occurrences of each block $[\omega]_{-D+t}^{t}$, of length up to $D \leq R + 1$[3]. It is known (see, e.g., [76] for a formal introduction) that this is a suitable data structure (faster to construct and to scan than hash-tables, for instance) in this context. It allows to maintain a computation time of order $O(TR)$, which does not depends on the structure size.

The prefix-tree algorithm.

Since we use such structure in a rather non-standard way compared to other authors, e.g. [76, 67], we detail the method here.

See [76] for a formal introduction and Fig. 3.1 for a graphical definition. We consider a spike train ω_{-T}^{0}, where time is negative. The prefix-tree data structure for the present estimation procedure is constructed iteratively.

1. Each spiking pattern at time t, $\omega(t)$, is encoded by an integer $w(t)$.

2. This given, before any symbol has been received, we start with the empty tree consisting only of the root.

[3]The code is available at `http://enas.gforge.inria.fr/classSuffixTree.html`.

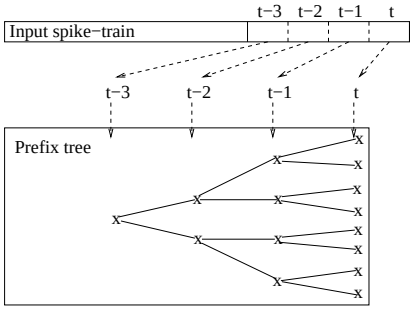

Figure 3.1: Schematic representation of a prefix-tree of depth $D = 4$ and degree $2 = 2^N, N = 1$, thus a binary tree for a single neuron spike train. Each block of depth D is counted in the prefix tree from the leaf corresponding to $\omega(t)$ back to the root corresponding to $\omega(t - D + 1)$.

3. Then suppose for $-D < t \leq 0$ that the tree $\mathcal{T}([\omega]_{-T}^{t-1})$ represents $[\omega]_{-T}^{t-1}$. One obtains the tree $\mathcal{T}([\omega]_{-T}^{t})$ as follows:

   (a) One starts from the root and takes branches corresponding to the observed symbols $\omega(t - D + 1), \cdots, \omega(t)$.

   (b) If ones reaches a leaf before termination, one replaces this leaf by an internal node and extends on the tree.

   (c) Each node or leaf has a counter incremented at each access, thus counting the number of occurrence $\#[\omega]_{-D}^{t}, -D < t \leq 0$ of the block $[\omega]_{-D}^{t}$ in the sequence $[\omega]_{-T}^{0}$.

The present data structure not only enable us to perform the empirical measure estimation over a period of time T, but can also obviously be used to aggregate several experimental periods of observation. It is sufficient to add all observations to the same data structure.

Generalization to a sliding window.

Though we restrict ourselves to stationary statistics in the present work, it is clear that the present mechanism can be easily generalized to the analysis of non-stationary data set, using a sliding window considering the empirical measure in $[t, t + T[$, then $[t + 1, t + 1 + T[$, etc. This is implemented in the present data structure by simply counting the block occurrences observed at time t and adding the block occurrences observed at time T, yielding a minimal computation load. The available implementation has already this functionality (see section 4.1.4 for an example).

3.1.3 Performing the parametric estimation

In a nutshell, the parametric estimation reduces to minimizing (2.36), hence (2.37), by calculating the topological pressure $P(\boldsymbol{\psi}) \equiv P(\boldsymbol{\lambda})$ using (5.11) and the related theorem. We remark that as consequence, our framework induces a change of estimation paradigm from Markov Chain sampling to matrix computations, namely eigenvalue and eigenvector computations. This opens by itself interesting perspectives from a computational point of view which are empowered additionally by the sparse character of the Perron-Frobenius matrix and the fact that we only require the maximal eigenvalue and its eigenvectors (instead of a complete eigendecomposition). The process decomposes into the following steps.

Potential eigen-elements calculation.

It has been shown in the theoretical section that the Perron-Frobenius matrix eigen-elements permits one to derive all characteristics of the probability distribution. Let us now describe at the algorithmic level how to perform these derivations.

1. The first step is to calculate the right-eigenvector $b\rangle$ of the $\mathcal{L}(\psi)$ matrix, associated to the highest eigenvalue, using a standard power-method series [4]:

$$\begin{aligned} s^{(n)} &= \|\mathcal{L}(\psi)\, v^{(n-1)}\| \\ v^{(n)} &= \tfrac{1}{s^{(n)}}\, \mathcal{L}(\psi)\, v^{(n-1)} \end{aligned}$$

   where $v^{(n)}$ is the n-th iterate of an initial vector $v^{(0)}$ and $s^{(n)}$ is the n-th iterate of an initial real value $s^{(0)}$. With this method the pair $(s^{(n)}, v^{(n)})$ converges to $(s(\psi), b\rangle)$ as given by (5.11). In our case, after some numerical tests, it appeared a good choice to either set $v^{(0)}$ to an uniform value, or to use the previous estimated value of $b\rangle$, if available. This last choice is going to speed up the subsequent steps of the estimation algorithm.

   The key point, in this iterative calculation, is that $\mathcal{L}(\psi)$ is a sparse $2^{NR} \times 2^{NR}$ matrix, as outlined in the section 2.1.2. As a consequence calculating $\mathcal{L}(\psi)\, v$ is a $O(2^{N+NR}) \ll O(2^{2\,NR})$ operation.

   The required precision on $(s(\psi), b\rangle)$ must be very high, for the subsequent steps to be valid, even if the eigenvector dimension is huge (it is equal to 2^{NR}), therefore the iteration must be run down to the smallest reasonable precision level (10^{-24} in the present implementation).

[4] This choice is not unique and several alternative numerical methods exists (e.g Krylov subspace methods) [52].

We have experimented that between 10 to 200 iterations are required for an initial uniform step in order to attain the required precision (for $NR \in 2..20$), while less than 10 iterations are sufficient when starting with a previously estimated value.

From this 1st step we immediately calculate:

(a) The topological pressure $P(\psi) = \log(s(\psi))$.

(b) The normalized potential is also stored in a look-up table. This gives us the transition matrix $\mathcal{M}$, which can be used to generate spike trains distributed according the Gibbs distribution $\mu_{\boldsymbol{\psi}}$ and used as benchmarks in the section 4.1.

2. The second step is to calculate the left eigenvector $\langle b$, this calculation having exactly the same characteristics as for $b\rangle$.

   From this 2nd step one immediately calculates:

   (a) The Gibbs probability of a block w given by (2.9), from which probabilities of any block can be computed (eq.2.7, 2.11).

   (b) The theoretical value of the observables average $\mu_{\boldsymbol{\psi}}(\phi_l)$, as given in (2.25).

   (c) The theoretical value of the distribution entropy $h[\mu_{\boldsymbol{\psi}}]$, as given in (2.30).

   After both steps, we obtain all useful quantities regarding the related Gibbs distribution: probability measure, observable value prediction, entropy. These algorithmic loops are direct applications of the previous piece of theory and show the profound interest of the proposed framework: given a Gibbs potential, all other elements can be derived directly.

Estimating the potential parameters.

The final step of the estimation procedure is to find the parameters $\boldsymbol{\lambda}$ such that the Gibbs measure fits at best with the empirical measure. We have discussed why minimizing (2.36) is the best choice in this context. Since $h(\pi^{(T)})$ is a constant with respect to λ, it is equivalent to minimize $\tilde{h}[\psi_\lambda]$ eq. (2.38), where $\mu_{\boldsymbol{\psi}}(\phi_l)$ is given by (2.25. Equivalently, we are looking for a Gibbs distribution $\mu_{\boldsymbol{\psi}}$ such that $\frac{\partial P[\psi_\lambda]}{\partial \lambda_l} = \pi^{(T)}(\phi_l)$ which expresses that $\pi^{(T)}$ is tangent to P at ψ_λ [102].

Matching theoretical and empirical observable values.

As pointed out in the theoretical part, the goal of the estimation is indeed to find the parameters $\boldsymbol{\lambda}$ for which theoretical and empirical observable values match. The important point is that this is exactly what is performed by the proposed method: minimizing the criterion until a minimum is reached, i.e. until the gradient vanishes corresponding to a point where $\mu_{\boldsymbol{\psi}}(\phi_l) = \pi^{(T)}(\phi_l)$, thus where theoretical and empirical observable values are equal. Furthermore, this variational approach provides an effective method to numerically obtain the expected result.

At the implementation level, the quantities $\pi^{(T)}(\phi_l)$ are the empirical averages of the observables, i.e. the observable averages computed on the prefix tree. They are computed once from the prefix tree. For a given $\boldsymbol{\lambda}$, $P(\lambda)$ is given by step 1.a of the previous calculation, while $\mu_{\boldsymbol{\psi}}(\phi_l)$ is given by the step 2.b. It is thus now straightforward[5] to delegate the minimization of this criterion to any standard powerful non-linear minimization routine.

We have implemented such a mechanism using the GSL[6] implementation of non-linear minimization methods. We have also made available the GSL implementation of the simplex algorithm of Nelder and Mead which does not require the explicit computation of a gradient like in eq. (2.38). This alternative is usually less efficient than the previous methods, except in situations, discussed in the next section, where we are at the limit of the numerical stability. In such a case the simplex method is still working, whereas other methods fail.

Measuring the precision of the estimation.

Once the quantity $\tilde{h}[\psi] = P[\psi] - \pi^{(T)}(\psi)$ (eq. (2.38)) has been minimized the Kullback-Leibler divergence $d(\pi^{(T)}, \mu_{\boldsymbol{\psi}}) = \tilde{h}[\psi] - h(\pi^{(T)})$ determines a

[5] Considering a simple gradient scheme, there is always a $\epsilon^k > 0$, small enough for the series $\boldsymbol{\lambda}_l^k$ and $\tilde{h}^k$, defined by:

$$\boldsymbol{\lambda}_l^{k+1} = \boldsymbol{\lambda}_l^k + \epsilon^k \frac{\partial \tilde{h}}{\partial \boldsymbol{\lambda}_l}(\boldsymbol{\lambda}_l^k)$$

$$0 \leq \tilde{h}^{k+1} < \tilde{h}^k,$$

to converge, as a bounded decreasing series, since:

$$\tilde{h}^{k+1} = \tilde{h}^k - \epsilon^k \left| \frac{\partial \tilde{h}}{\partial \boldsymbol{\lambda}_l} \right|^2 + O\left((\epsilon^k)^2\right).$$

[6] The GSL `http://www.gnu.org/software/gsl` multi-dimensional minimization algorithms taking the criteria derivatives into account used here is the Fletcher-Reeves conjugate gradient algorithm, while other methods such as the Polak-Ribiere conjugate gradient algorithm, and the Broyden-Fletcher-Goldfarb-Shannon quasi-Newton method appeared to be less efficient (in precision and computation times) on the benchmarks proposed in the result section. Anyway, the available code `http://enas.gforge.inria.fr/classIterativeSolver.html` allows us to consider these three alternatives, thus allowing to tune the algorithm to different data sets.

notion of "distance" between the empirical measure $\pi^{(T)}$ and the statistical model μ_{ψ}. Though it is not necessary to compute $d(\pi^{(T)}, \mu_{\psi})$ for the comparison of two statistical models $\mu_{\psi}, \mu_{\psi'}$, the knowledge of $d(\pi^{(T)}, \mu_{\psi})$, even approximate, is a precious indication of the method precision. This however requires the computation of $h(\pi^{(T)})$.

Though the numerical estimation of $h(\pi^{(T)})$ is a far from obvious subject, we have implemented the entropy estimation using definitions (2.28) and (2.29). In order to interpolate the limit (2.29), we have adapted an interpolation method from [76] and used the following interpolation formula. Denote by $h(\pi^{(T)})^{(n)}$ the entropy estimated from a raster plot of length T, considering cylinders of size n. We use the interpolation formula $h(\pi^{(T)})^{(n)} \simeq h^{\infty} + \frac{k}{n^c}$, where $h^{\infty}, k, c > 0$ are free parameters, with $h(\pi^{(T)})^{(n)} \to h^{\infty}$, as $n \to +\infty$. The interpolation formula has been estimated in the least square sense, calculating $h(\pi^{(T)})^{(n)}$ on the prefix-tree. The formula is linear with respect to h^{∞} and k, thus has a closed-form solution with respect to these two variables. Since the formula is non-linear with respect to c, an iterative estimation mechanism is implemented.

3.1.4 Design choices: genesis of the algorithm.

Let us now discuss in details the design choices behind the proposed algorithm.

The fact that we have an implementation able to efficiently deal with higher-order dynamics is the result of computational choices and validations, important to report here, in order for subsequent contributor to benefit from this part of the work.

Main properties of the algorithm.

Convexity. As indicated in the section 2.5.5 there is a unique minimum of the criterion. However, if the guess potential $\psi^{(test)}$ does not contain some monomials in ψ^*, the sought potential, the procedure converges but there is an indeterminacy in the λ_l's corresponding to those monomials. The solution is not unique, there is a subspace of equivalent solutions. The rank of the topological pressure Hessian is an indicator of such a degenerate case. Note that these different situations are not inherent to our procedure, but to the principle of finding an hidden probability by maximizing the statistical entropy under constraints, when the full set of constraints is not known [43].

Finite sample effects. As indicated in the section 2.5.7 the estimations crucially depend on T. This is a central problem, not inherent to our ap-

proach but to all statistical methods where one tries to extract statistical properties from finite empirical sample. Since T can be small in practical experiments, this problem can be circumvented by using an average over several samples. In the present formalism it is possible to have an estimation of the size of fluctuations as a function of the potential, using the central limit theorem and the fact that the variance of fluctuations is given by the second derivative of the topological pressure. This is a further statistical test where the empirical variance can be easily measured and compared to the theoretical predictions.

Numerical stability of the method. Two factors limits the stability of the method, from a numerical point of view.

The first factor is that the matrix $\mathcal{L}(\psi)$ is a function of the *exponential* of the potential $\psi = \sum_l \lambda_l \, \phi_l$. As a consequence, positive or negative values of ψ yield huge or vanishing value of $\mathcal{L}(\psi)$, and numerical instabilities easily occurs.

However, though numerical instabilities are unavoidable, the good news is that they are easily detected, because we have introduced a rather large set of numerical tests in the code:

1. Negligible values (typically lower than 10^{-4}) are set to zero, implicitly assuming that they correspond to hidden transitions in the grammar.

2. Huge values (typically higher than 10^4) generate a warning in the code.

3. Several coherent tests regarding the calculation of the PF eigenelements are implemented: we test that the highest eigenvalue is positive (as expected from the PF theorem), and that the left and right PF related eigenvectors yield equal eigenvalues, as expected; we also detect that the power-method iterations converge in less than a maximal number of iteration (typically 2^{10}). We never found this spurious condition during our numerical tests. When computing the normalized potential (2.10), we verify that the right eigenvalue is 1 up to some precision, and check that the normal potential is numerically normalized (i.e. that the sum of probabilities is indeed 1, up to some "epsilon").

In other words, we have been able to use all what the piece of theory developed in the previous section makes available, to verify that the numerical estimation is valid.

The second factor of numerical imprecision is the fact that some terms $\lambda_l \, \phi_l$ may be negligible with respect to others, so that the numerical estimation of the smaller terms becomes unstable with respect to the imprecision

of the higher ones. This has been extensively experimented, as reported in the next section.

Relation with entropy estimation. The construction of a prefix-tree is also the basis of efficient entropy estimation methods [76, 167]. See [67] for a comparative about entropy estimation of one neuron spike train (binary time series). Authors numerically observed that the context-tree weighting methods [115] is seen to provide the most accurate results. This, because it partially avoids the fact that using small word-lengths fails to detect longer-range structure in the data, while with longer word-lengths the empirical distribution is severely under-sampled, leading to large biases. This statement is weaken by the fact that the method from [167] is not directly tested in [67], although a similar prefix-tree method has been investigated.

However the previous results are restrained to relative entropy estimation of "one neuron" whereas the analysis of entropy of a *group of neurons* is targeted if we want to better investigate the neural code. In this case [167] is directly generalizable to non-binary (thus multi-neurons) spike trains, whereas the context-tree methods seems intrinsically limited to binary spike-trains [115], and the numerical efficiency of these methods is still to be studied at this level.

Here, we can propose an estimation for the statistical entropy from eq. (2.30). Clearly, we compute here the entropy of a Gibbs statistical model μ_ψ while methods above try to compute this entropy from the raster plot. Thus, we do not solve this delicate problem, but instead, propose a method to benchmark these methods from raster plots (synthetic or real data) obeying a Gibbs statistics whose parametric form is already known.

Key aspects of the numerical implementation.

Unobserved blocks.

We make here the (unavoidable) approximation that unobserved blocks correspond to forbidden words (actually, our implementation allows to consider that a block is forbidden if it does not appear more than a certain threshold value). There is however, unless a *a priori* information about the distribution is available, no better choice. The present implementation allows us to take into account such a *a priori* information, for instance related to global time constraints on the network dynamics, such as the refractory period. See [26] for an extended discussion.

Potential values tabulation.

Since the implementation is anyway costly in terms of memory size, we have chosen to pay this cost but obtaining the maximal benefit of it and we

used as much as possible tabulation mechanisms (look-up tables) in order to minimize the calculation load. All tabulations are based on the following binary matrix:

$$\mathbf{Q} \in \{0,1\}^{L \times 2^{NR}},$$

with $\mathbf{Q}_{l,w} = \phi_l([\omega]_{-R}^{0})$, where w is given by (2.1). $\mathbf{Q}$ is the matrix of all monomial values, entirely defined by the choice of the parameter dimensions N, R and D. It corresponds to a "look-up table" of each monomial values where w encodes $[\omega]_{-R}^{0}$. Thus the potential (2.18) writes $\psi_w = (\mathbf{Q}\,\boldsymbol{\lambda})_w$. We thus store the potential exponential values as a vector and get values using a look-up table mechanism, speeding-up all subsequent computations.

This allows to minimize the number of operations in the potential eigen-elements calculation.

About other estimation alternatives.

Though what is proposed here corresponds, up to our best knowledge, to the best we can do to estimate a Gibbs parametric distribution in the present context [7], this is obviously not the only way to do it, and we have rejected a few other alternatives, which appeared less suitable. For the completeness of the presentation, it is important to briefly discuss these issues.

Avoiding PF right eigen-element's calculation. In the previous estimation, at each step, we have to calculate step 1 of the PF eigen-element's derivation for the criterion value calculation and step 2 of the PF eigen-element's derivation for the criterion gradient calculation. These are a costly $O(2^{N+NR})$ operations.

One idea is to avoid step 2 and compute the criterion gradient numerically. We have explored this track: we have calculated $\frac{\partial \tilde{h}}{\partial \lambda_l} \simeq \frac{\tilde{h}(\lambda_l+\epsilon)-\tilde{h}(\lambda_l-\epsilon)}{2\,\epsilon}$ for several order of magnitude, but always found a poorer convergence (more iterations and a biased result) compared to using the closed-form formula. In fact, each iteration is not faster, since we have to calculate $\tilde{h}$ at two points thus, to apply step 1, at least two times. This variant is thus to be rejected.

Another idea is to use a minimization method which does not require the calculation of the gradient: we have experimented this alternative using the simplex minimization method, instead of the conjugate gradient method, and have observed that both methods correctly converge towards a precise solution in most cases, while the conjugate gradient method is faster. However, there are some cases with large range potential, or at the limit of

[7] Additionally, without involving parallel computing methods and trying to maintain good portability

the numerical stability where the simplex method may still converge, while the other does not.

About analytical estimation of the PF eigen-element's. The costly part of the PF eigen-element's computation is the estimation of the highest eigenvalue. It is well-known that if the size of the potential is lower than five, there are closed-form solutions, because this problem corresponds to finding the root of the matrix characteristic polynomial. In fact, we are going to use this nice fact to cross-validate our method in the next section. However, except for toy's potentials (with $2^{NR} < 5 \Leftrightarrow NR \leq 2$!), there is no chance that we can not do better than *numerically* calculating the highest eigenvalue. In the general case, the power method is the most direct to compute it, although Krylov subspace methods are an interesting perspective for very large matrices [183].

Using other approximations of the KL-divergence criterion. Let us now discuss another class of variants: the proposed KL-divergence criterion in (2.31) and its empirical instantiation in (2.36) are not the only one numerical criterion that can be proposed in order to estimate the Gibbs distribution parameters. For instance, we have numerically explored approximation of the KL-divergence of the form:

$$d(\nu, \mu) \simeq \sum_{n=R}^{R'} \frac{\alpha_n}{n} \sum_{[\omega]_0^{n-1}} \nu\left([\omega]_0^{n-1}\right) \log \left[\frac{\nu\left([\omega]_0^{n-1}\right)}{\mu\left([\omega]_0^{n-1}\right)} \right],$$

and have obtained coherent results (for $\alpha_n = 1$), but not quantitatively better than what is observed by the basic estimation method, at least for the set of performed numerical tests.

All these variants correspond to taking into account the same kind of criterion, but some other weighted evaluations of the empirical average of the observable. There is no reason to use it unless some specific a *a priori* information on the empirical distribution is available.

Another interesting track is to use (2.10) which allows us to write a KL-divergence criterion, not on the probability block, but on the conditional probability block, as proposed in [29, 30] in a different context. We have considered this option. However a straightforward derivation allows one to verify, that this in fact corresponds the same class of criterion but with a different empirical observable average estimation. At the numerical level, we did not observe any noticeable improvement.

Estimation in the case of a normalized potential. In the case where the potential is normalized, the criterion (2.38) is a simple linear criterion, thus unbounded and its minimization is meaningless. In this singular case, its is obvious to propose another criterion for the estimation of the parameters. A simple choice is to simply propose that the theoretical likelihood of the measure matches the estimated one, in the *least square sense*. This has been integrated in the available code as follows.

In this singular case, we have to propose another criterion for the estimation of the parameters.

Following the likelihood choice it is straightforward to recover from this problem minimizing a modified, but well-defined criterion based on eq. (2.14):

$$\sum_{[\omega]_0^{n-1}} \left[\log\left(\pi^{(T)}\left([\omega]_0^{n-1}\right)\right) + Z_n - \sum_{t=0} n - R\boldsymbol{\psi}([\omega]_t^{t+R}) \right]^2, \tag{3.3}$$

taken here for $n = R$, where Z_n is now a simple normalization constant, estimated here in the least-square sense. Since we solve a quadratic problem we obtain linear equations with a unique solution in the general case and an infinity of solution in degenerated cases. In this last case we consider the solution with the smallest magnitude. At the implementation level, the optimal solution corresponds to a least-square problem with respect to $(\boldsymbol{\lambda}, Z)$, obvious to calculate using the singular value decomposition[8] of $\mathbf{A}$, which is known to provide either the unique minimal value of $|\mathbf{b} + \mathbf{A}\,\boldsymbol{\lambda}|^2$ if any, or the value of minimal amplitude $|\boldsymbol{\lambda}|$, with $\mathbf{b} + \mathbf{A}\,\boldsymbol{\lambda} = 0$ when there are several solutions. In the code, we have used the GSL[9] implementation of the singular value decomposition. In order to obtain an unbiased estimate for $\boldsymbol{\lambda}$, thus so that the related KL-divergence vanishes, the obtained result is reprojected

[8] The singular value decomposition (SVD) writes:

$$\mathbf{A} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T, \quad \begin{array}{l} \mathbf{U} \in \mathcal{R}^{L\times 2^{NR}}, \mathbf{D} \in (\mathcal{R}^+)^L, \mathbf{V} \in \mathcal{R}^{L\times L}, L \leq 2^{NR}, \\ \mathbf{U}^T\,\mathbf{U} = \mathbf{V}^T\,\mathbf{V} = \mathbf{V}\,\mathbf{V}^T = 1 \in \mathcal{R}^{L\times L} \end{array}$$

$\mathbf{U}$ and $\mathbf{V}$ being orthogonal, while is $\mathbf{D}$ is diagonal and positive. We obtain:

$$\boldsymbol{\lambda} = -\mathbf{V}\,\mathbf{D}^\dagger\,\mathbf{U}^T\,\mathbf{b}$$

writing $\mathbf{D}^\dagger$ the pseudo-inverse of the diagonal matrix $\mathbf{D}$, thus with:

$$\mathbf{D}^\dagger_{ll} = \text{ if } \mathbf{D}_{ll} > 0 \text{ then } 1/\mathbf{D}_{ll} \text{ else } 0.$$

[9] The GSL `http://www.gnu.org/software/gsl` SVD decomposition implements the most commonly-used Golub-Reinsch method of bi-diagonal matrix decomposition using Householder reflections followed by a QR decomposition, writing the matrix as a product of an orthogonal matrix and an upper triangular matrix, multiply the factors in the other order, and iterate. A modified Golub-Reinsch algorithm has also be tested in our case, since $2^{NR} \gg L$, but in fact without any noticeable improvement. The applicative code is available at `http://enas.gforge.inria.fr/classLinearSolver.html`.

onto the KL-divergence, i.e. the final estimate is the unique value whose distance to the quadratic criterion (3.3) optimum is minimal and for which the KL-divergence (a linear form in this case) vanishes. This estimation step is useful to both propose a direct estimation method for normalized potential and also an initial estimate for the non-linear iterative general method proposed now, if this can speed up the non-linear estimation.

3.1.5 Perspectives of ENAS development for Gibbs Distribution Framework

The following perspectives concern mainly how to improve the proposed framework of Gibbs distributions. Besides, benchmarking our method with respect to other alternative methods as MCMC, and further applications of the library to other computational neuroscience themes like input-coding/decoding and machine learning constitute other important perspectives by their own and are not explicitly treated.

Parallelization and optimal Eigenvalue solution As stated in section 2.3.3, a key aspect of Gibbs Distribution framework is that it conduces to a the change of paradigm for the parametric estimation from Markov chain sampling to matrix computations (see section (3.1.3). This enables to use parallel computation techniques that are not yet included in the current implementation and that become mandatory since the dimension of the matrices involved, increases exponentially fast with the range of the potential and the number of neurons. An second advantage in order to future optimization of the implementation is that since we are dealing with a *sparse matrix* there exist eigenvalue sparse-specific methods such as Krylov subspace iterations methods, see [52] and the detailed survey in [9].

Goodness-of-Fit test Several ready to use GOF implementations exists (see section 3.2). However in order to use them, I would be necessary to set-up a correct interface to introduce within those implementations the EnaS output. More precisely, to adequately present the estimated Gibbs distribution or vector at least a vector of marginal probabilities obtained from it. A point of attention needs to be done here since we recall that Gibbs distributions do not provide a continuous uni-variate distribution (see section 2.5.9) and whether or not it satisfies the validity criteria for the proposed test needs to be analyzed previously for each case.

Large deviations Integrating the results of section 2.5.7 to the current implementation will permit to estimate the size of fluctuations for each coefficient and hence better identify which ones of estimated parameters are

reliable, and which characteristics are intrinsic to the data and which ones are apparent constituting simply effects of the finite-size fluctuations

Neyman-Pearson criteria This criteria presented in section 2.5.9, offers a control of confidence intervals for model discrimination given the finite-sample length T. It could be very useful to integrate it to EnaS capabilities although its implementation is cumbersome.

Estimating the memory of Sequence Up to the author knowledge none of the methods discussed in section 2.5.6 are included on a freely available implementation. It would be interesting to integrate them to the EnaS library in order to pre-analyze the training sequence (the observed data) and take account of its results to help the choice of candidate models. However their implementation is not straightforward. For the first criteria, which consists in evaluating the block entropy difference, there are unavoidable finite sample biases in every block entropy estimators which need to be corrected by some choice of bias-corrector [88]. For the second criteria, the Morvai-Weiss technique requires a maximization within the space of sequence suffixes of different length which reflects a non trivial search with computational effort to be determined.

Optimizing model selection strategies In a complementary manner to the previous idea, there is the strategy of model "decimation" introduced in section 2.5.5. We recall that this consist in beginning the estimation from the largest possible potential with range $R+1$ (i.e.. with all the non-redundant monomials) and after parameter estimation to remove irrelevant monomials (their coefficient vanishes in the estimation). Then re-do the parameter estimation verifying that the KL-divergence error reduces or at least remains unchanged, and repeat the same procedure until obtaining a potential composed only of the relevant (non-zero). Since at the current stage this can be performed manually, it might be helpful for users to automatize it in the code.

Non-stationary higher-order Markov chain generation In order to validate numerically our parametric estimation method for time-varying parameters of Gibbs distributions with memory (see section **??**) we need to formally resolve and numerically implement the question of how to construct a Markov chain of higher-order with time-dependent transition matrix and implement.

Integrating Information theory aspects In the spirit of Ince et al.[85] it would be interesting to trying to investigate how to improve entropy and mutual-Information estimates thanks to the time-correlation structure of the system, more precisely through the knowledge of the value of the Gibbs potential parameters for the best suited (closest in the KL-divergence sense) Gibbs distribution.

3.2 Appendix: Other Useful Numerical software

This section is not intended to be an exhaustive nor a detailed listing of numerical software eventually useful for multi-cellular spike-train analysis mainly in the scope and perspective of Gibbs distributions. It concerns a set of existing sources -up to the authors knowledge- that might be complete or be useful for the future developments of ENAS tools. For additional software available for Neurophysiology data see the survey section *Open Source Computational Toolboxes for the Analysis of Neurophysiological Data* in [86].

3.2.1 Maximum entropy software

Up to our knowledge, none of this sources provide computations by the RPF method, hence they require sampling (constructing a sequence with the parametric model in order to perform averages over the constrains), and this for each training step.

Neuroscience-purpose tools by Ince et al. This work, originally coded In python (PyEntropy at `http://code.google.com/p/pyentropy/`), which is also available in MATLAB has been presented in a series of papers since 2009, by R. Inc and collaborators [88]. It is composed of two parts: an Information theory -purpose part and the other for Maximum entropy estimation. The information theory part contains implementation of several algorithms proposed to compute the entropy of spike-sequences and several finite-sample bias estimators for that computation. The maximum entropy estimator, is limited to memoryless models (range 1 observables). It is based in combinatorial analysis and the work about coordinate transformation properties presented in [177] for the memoryless potential. This implementation was used in the work of Montani et al. [87] to show the presence of high-order (range 1) interactions among somatosensory neurons.

The Information Breakdown Toolbox 5 (ibTB) Following [86] *"(Magri et al., 2009-`http://www.ibtb.org/`) is a MATLAB toolbox implementing*

several of the information estimates and bias corrections (...) Importantly, it does this via a novel algorithm to minimize the number of operations required during the direct entropy estimation, which results in extremely high speed of computation. It contains a number of algorithms which have been thoroughly tested and exemplified not only on spike train data (as for the above toolboxes), but also on data from analogue brain signals such as LFPs and EEGs."

Maximum Entropy Modeling Toolkit for Python and C++ This toolkit was developed by Zhang Le, between 2002-2003 and is based at `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`. It intends to be a general-purpose Maximum entropy tool for Natural Language Processing (NLP). This implementation do account for the conditional probability, however with two important remarks: First, here the events are "alphabetic words" and second, the conditioning event (the context) is limited to a "single word". Hence there are two main consequences: first is that if the only possible words were to be single time spiking pattern $\omega(0)$, this would produce a model of range 2, since its context (anteceding word) would be the precedent single time spiking pattern. Second, the parametric expansion can account for each possible block, or say in other words, the monomials here, take into account too the quiescent states (i.e.. the 0 within the pattern) because the context is a "full word", in contrast to our implementation were monomials are described merely by the firing states.

For the parametric estimation , this toolkit implements high performing minimization methods notably Limited-Memory-BGFS instead of Basic conjugated gradient (see a survey in [118]. This estimation method can be parallelized (See Robert Malouf's Maximum Entropy Parameter Estimation software at `http://tadm.sourceforge.net/`).

Other Maximum entropy special-purpose software We reproduce here most notable items appearing in the non exhaustive list provided by Zhang Le on his web page `http://homepages.inf.ed.ac.uk/lzhang10/maxent.html`., most of them related with NLP purposes, and refer readers to there for more details.

- **maxent.sf.net**. A great java maxent implementation with GIS training algorithm. Part of OpenNlp project.
- **Amis** A maximum entropy estimator for feature forests. A maximum entropy estimator with GIS, IIS and L-BFGS algorithms.

- **TADM**. Released in 2005, it is the R. Malouf's Maximum Entropy Parameter Estimation software, now available as Toolkit for Advanced Discriminative Modeling on sourceforge.net. It has GIS, IIS, L-BFGS and Gradient Descent training methods and parallel computation ability through PETSc numerical library.

- **Text Modeller** A python implementation of a joint Maximum Entropy model (aka. Whole Sentence Language Model) with sampling based training. Now seems to be part of scipy.

- **Maxent** A software for species habitat modeling by Robert E. Schapire et al.

3.2.2 Other relevant sources

The Spike Train Analysis Toolkit (Goldberg et al., 2009) This package coded in C is available at `http://neuroanalysis.org/toolkit/`. Following [86] it *"... is a MATLAB toolbox which implements several information-theoretic spike train analysis techniques. It is a comprehensive piece of software, covering a range of entropy and information bias correction methods. Particularly notable is the inclusion of the so-called metric space (Victor and Purpura, 1996) and binless (Victor, 2002) methods for estimating information theoretic quantities from spike trains, which to our knowledge are not available in any other package."*

Parallel computation and Scalable Eigenvalue software There exist a diversity of parallel computation libraries based on different programming languages such as Fortran, C, C++ , MATLAB. Since EnaS is a C++ library, we might like to focus our attention on C-based available alternatives. A notable example within this category (although certainly not the only) is PETSc (`http://www.mcs.anl.gov/petsc/petsc-as/`) which is a suite of data structures and routines for the scalable (parallel) solution of scientific applications, in for particular partial differential equations. In particular we find SLEPc, the Scalable Library for Eigenvalue Problem Computations (`http://www.grycap.upv.es/slepc/`), an extension of PETSc dedicated to the solution of large scale sparse eigenvalue problems on parallel computers. It is and can be used for either standard or generalized eigenproblems, with real or complex arithmetic. It can also be used for computing a partial SVD of a large, sparse, rectangular matrix, and to solve quadratic eigenvalue problems. Both packages might be completed thanks to the set of Large eigenvalue templates listed in [9].

Goodness-of-Fit (GOF) sources There are also a great number of statistical software/tool that provides ready-to-use routines to perform standard GOF test. A good C++ based example is the *statistical toolkit* developed but Donadio and collaborators. [37, 123] from INFN,Italy and the CERN, Switzerland and freely available at `http://www.ge.infn.it/statisticaltoolkit`. It is an open source software toolkit for statistical data analysis conceived with advanced programming techniques for efficiency and evolution freedom, and its first cycle focus on implementation of Goodness-of-Fit test. Besides, a second example specifically aimed for the Crammer-von-Misses test CvM, which corresponds to one of the most promising candidates to be used as complementary fitness test, it exist a 2006 C++ implementation [209] which source is available at `http://www.jstatsoft.org/v17/i08/supp/1`.

Chapter 4

Applications

Overview:

This chapter is organized as follows. The first section concerns the validation of the implementation described in chapter 3. We start with several elementary numerical test and showing its performance thanks to exact results known in the case of simple models. Then more complex examples are presented and discussed in the light of theoretical aspects discussed in chapter 2. Additionally, examples of finite sample size effects are shown. In the second section we focus on the analysis of biological data. It begins with a preliminary introduction about the theoretical/experimental concerns of multicellular recordings. Then, results about three different data sets issued from retina recordings are presented. Finally, the binning effect are discussed by means of a very basic numerical example.

Contents

4.1 Synthetic Data Results

4.1.1 Basic tests: validating the method

Method

Given a potential $\psi = \sum_{l=1}^{L} \lambda_l \, \phi_l$ it is easy to generate a spike train of length T distributed according to μ_ψ using (2.7). Thus, we have considered several examples of Gibbs potentials, where, starting from a sample raster plot $[\omega]_{-T}^{0}$ distributed according to μ_ψ, we use our algorithm to recover the right form of the generating potential ψ.

Given a potential of range-$R+1$ of the parametric form (2.18) and a number of neurons N we apply the following method:

1. Randomly choosing the parameter's values λ_l, $l = 1 \dots L$ of the Gibbs potential;
2. Generating a spike train realization of length T;
3. From these values re-estimating a Gibbs potential:
   (a) Counting the block occurrences, thus the probabilities $\pi^{(T)}$ from the prefix-tree,
   (b) Minimizing (2.38), given $\pi^{(T)}$, as implemented by the proposed algorithm.
   (c) Evaluating the precision of the estimation as discussed in the previous section.

We emphasize that in the previous method there is a way to simulate "infinite" ($T = +\infty$) sequences, by skipping step 2., and filling the prefix-tree in step 3.a directly by the exact probability of blocks. At first glance, this loop seems to be a "tautology" since we re-estimate the Gibbs potential parameters from a raster plot generated with a known Gibbs potential. However, the case $T = +\infty$ is a somewhat ideal case since no finite-sample statistical fluctuations are present and studying this case is useful since:

1. Using the same potential for the prefix-tree generation and for the parameters estimation, must yield the same result, but *up to the computer numerical precision*. This has to be controlled due to the non-linear minimization loop in huge dimension. This is obviously also a way to check that the code has no mistake.
2. The precision, rapidity and robustness of the method with respect to the number of parameters can be checked.

As an additional and mandatory test, one as then to generate rasters with a known potential where $T < +\infty$ is increasing in order to study the previous points in the realistic situation of finite size data set, providing quantitative estimations about the expected finite-sample effects as a function of T.

Some illustrative examples to understand what the algorithm calculates

Let us start with very simple examples, for which we can make explicit what the algorithm calculates thus helping the reader to understand in details what the output is, and then increase their complexity. In the first examples analytical expression for the topological pressure, entropy, eigenvectors and invariant measure are available. Thus we can check that we re-obtain, from the estimation method, the related values up to the numerical precision.

One neuron and range-2. Here $\psi(\omega) = \lambda_1 \omega_0(0) + \lambda_2 \omega_0(0)\omega_0(-1)$. We obtain analytically:

$$\begin{array}{lcl} s(\psi) & = & \frac{1+B+\sqrt{(1-B)^2+4A}}{2}, \\ P(\psi) & = & \log s(\psi), \\ \langle b & = & (1, s(\psi)-1, A, B(s(\psi)-1),) \\ b\rangle & = & (s(\psi)-B, s(\psi)-B, 1, 1)^{(t)}, \\ \mu_\psi & = & \frac{1}{s(\psi)^2+A-B}\left(s(\psi)-B, A, A, B(s(\psi)-1)\right), \\ h\left[\mu_\psi\right] & = & \log(s(\psi)) - \lambda_1 \frac{\partial s(\psi)}{\partial \lambda_1} - \lambda_2 \frac{\partial s(\psi)}{\partial \lambda_2} \\ r & = & \frac{A+B(s(\psi)-1)}{s^2(\psi)+A-B}, \\ C & = & \frac{B(s(\psi)-1)}{s^2(\psi)+A-B}, \end{array}$$

with $A = e^{\lambda_1} = e^{\psi_{10}}, B = e^{\lambda_1+\lambda_2} = e^{\psi_{11}}$ and where (t) denotes the transpose. We remind that the index vector encodes spike blocs by eq. (2.1). Thus, the first index (0) corresponds to the bloc 00, 1 to 01, 2 to 10 and 3 to 11. r is the firing rate, C the probability that the neuron fires two successive time steps. This is one among the few models for which a closed-form solution is available.

The following numerical verifications have been conducted. A simulated prefix-tree whose nodes and values has been generated using (2.18) with $\lambda_1 = \log(2), \lambda_2 = \log(2)/2$. We have run the estimation program of λ_l's and have obtained the right values with a precision better than 10^{-6}. This first test simply states that the code has no mistake.

A step further, we have used this simple potential to investigate to which extends we can detect if the model is of range-1 (i.e. with $\lambda_2 = 0$) or range-2 (i.e. with a non-negligible value of λ_2). To this purpose, we have generated

a range-2 potential and have performed its estimation using a range-1 and a range-2 potential, comparing the entropy difference (Fig. 4.1).

As expected the difference is zero for a range-2 model when $\lambda_2 = 0$, and this difference increases with λ_2. Less obvious is the fact that curves saturate for high values of λ_2. Increasing some λ_l's leads to an increase in the potential values for those blocks ω_{-R}^0 such that the monomial ϕ_l corresponding to λ_l is equal to 1. Consequently, the conditional probability $Prob\left[\omega(0)\,|\,\omega_{-R}^{-1}\right]$ increases. Since this probability is bounded by 1 the corresponding curve of $Prob\left[\omega(0)\,|\,\omega_{-R}^{-1}\right]$ and, likewise of the expectation of ϕ_l, saturates for high λ_l value. Now, the theoretical value for $\tilde{h}$ is given in the present case by $\tilde{h} = P(\psi_1) - \mu_{\psi_2}(\psi_1) = P(\psi_1) - \lambda_1\,\mu_{\psi_2}(\omega_1(0))$. As λ_2 increases $\mu_{\psi_2}(\omega_1(0))$ converges to 1 leading to the observed saturation effect.

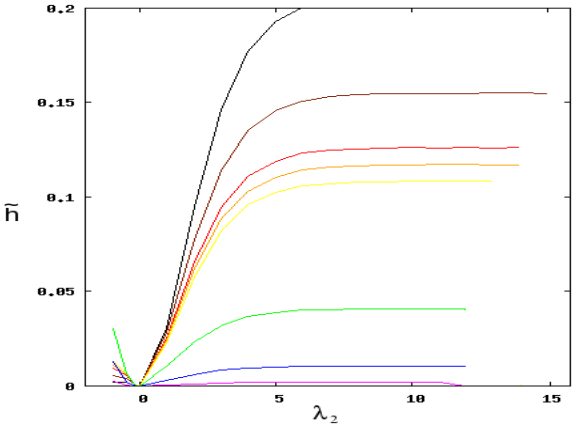

Figure 4.1: Entropy difference, using $\tilde{h}$, defined in (2.38), between the estimations of a range-1 and a range-2 model. The range-2 model reads $\psi_2 = \lambda_1\,\omega_0(0) + \lambda_2\omega_0(0)\,\omega_0(1)$ for $\lambda_1 = \{-1\ (black), -0.5\ (brown), -0.2\ (red), -0.1\ (orange), 0\ (green), 1\ (blue), 2\ (Magenta)\}$. λ_2 is a free parameter, in abscissa of this curve. The range-1 corresponds to $\lambda_2 = 0$.

We also have generated a range-1 potential and have performed its estimation, using a range-1 versus a range-2 model, and found always that using range-2 model is as good as using a model of range-1 (not shown).

Two neurons and range-1 (Ising). Here $\psi(\omega) = \lambda_1\,\omega_1(0) + \lambda_2\,\omega_2(0) + \lambda_3\,\omega_1(0)\,\omega_2(0)$. The largest eigenvalue of the $\mathcal{L}(\psi)$ matrix is $Z = s(\psi) = A+B+C+D$, with $A = 1, B = e^{\lambda_1}, C = e^{\lambda_2}, D = e^{\lambda_1+\lambda_2+\lambda_3}$ and the topological pressure is $\log s(\psi)$. We still obtain numerical precision better than 10^{-4}.

Two neurons and pattern of spikes. A step further, we have considered $\psi(\omega) = \lambda_1\,\omega_1(0) + \lambda_2\,\omega_2(0) + \lambda_3\,\omega_1(0)\,\omega_2(-1)\,\omega_1(-2)$, and $\psi(\omega) = \lambda_1\,\omega_1(0) + \lambda_2\,\omega_2(0) + \lambda_3\,\omega_1(0)\,\omega_2(-1)\,\omega_2(-2)\,\omega_3(-3)$, for random values drawn in $]-1, 0[$. We still obtain a numerical precision better

than 10^{-3} although the precision decreases with the number of degrees of freedom, while it increases with the observation time. This is investigated in more details in the remainder of this section.

When considering larger neuron N and range-$R+1$ the main obstacle toward analytical results is the Galois theorem which prevent a general method for the determination of the largest eigenvalue of the $\mathcal{L}(\psi)$ matrix. Therefore, we only provide numerical results obtained for more general potentials.

In all these numerical examples we have mainly considered $T = +\infty$ and used the same potential for the prefix-tree generation and for the parameters value estimation. However, we have also considered finite sequences with $T < +\infty$ and observed that for such simple models, the same numerical precision as the $T = +\infty$ case is obtained for $T \simeq 10^5$.

Gibbs potential precision paradigm: several neurons and various ranges.

In order to evaluate the numerical precision of the method, we have run the previous benchmark considering potentials with all monomial of degree less or equal to 1, and less or equal to 2, at a various ranges, with various numbers of neurons. Here we have chosen $T = +\infty$ and used the same potential for the prefix-tree generation and for the parameters value estimation. The computation time is reported in Table 4.1 and the numerical precision in Table 4.2, for $NR \leq 16$. This benchmark allows us to verify that there is no "surprise" at the implementation level: computation time increases in a supra-linear way with the potential size, but, thanks to the chosen estimation method, remains tractable in the size range compatible with available memory size. This is the best we can expect, considering the intrinsic numerical complexity of the method. Similarly, we observe that while the numerical precision decreases when considering large size potential, the method remains stable. Here tests have been conducted using the standard 64-bits arithmetic, while the present implementation can easily be recompiled using higher numerical resolution (e.g. "long double") if required.

This benchmark has also been used to explore the different variants of the estimation method discussed in the previous section (avoiding eigenvectors calculation, using other approximations of the KL-divergence criterion, ..).

Table 4.1: CPU-time order of magnitude in seconds (using Pentium M 750 1.86 GHz, 512Mo of memory), for the estimation of a potential with all monomial of degree less or equal to 1 for ψ_1, and less or equal to 2 for ψ_2, i.e., $\psi_1(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0)$, $\psi_2(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \sum_{\tau=-1}^{0} \lambda_{ij\tau} \omega_i(0)\omega_j(\tau)$, while the number N of neurons is increasing. Note that the present implementation is not bounded by the computation time, but simply by the rapid increase of the memory size.

ψ_1	R=1	R=2	R=4	R=8	R=16
N=1	2.0e-06	3.0e-06	8.0e-06	7.8e-05	2.9e-01
N=2	4.0e-06	1.0e-06	3.0e-05	6.7e-02	
N=4	1.3e-05	3.8e-05	8.3e-02		
N=8	2.4e-03	3.2e-01			

ψ_2	R=1	R=2	R=4	R=8	R=16
N=1	4.5e-16	4.0e-06	4.0e-06	7.2e-04	3.7e-02
N=2	3.0e-06	5.0e-06	4.0e-04	1.1e+00	
N=4	1.9e-05	1.2e-03	3.6e+00		
N=8	6.6e-03	6.2e-01			

4.1.2 Test of Finite sample size effects on the estimation precision

Test 1:

In order to test more general potentials for $N = 2$ neurons we have studied the forms (2.20), (2.23), (2.24), that we recall here:

$$\begin{aligned} \text{Ising} : \psi(\omega) &= \lambda_1\, \omega_1(0) + \lambda_2\, \omega_2(0) + \lambda_3\, \omega_1(0)\, \omega_2(0). \\ \text{RPTD} - k : \psi(\omega) &= \lambda_1\, \omega_1(0) + \lambda_2\, \omega_2(0) + \sum_{i=-k}^{i=k} \hat{\lambda}_i\, \omega_1(0)\omega_2(i). \\ \text{PTD} - k : \psi(\omega) &= \sum_{i=-k}^{i=k} \hat{\lambda}_i\, \omega_1(0)\omega_2(i). \end{aligned} \tag{4.1}$$

Protocol Given a generating potential ψ^* of the form (4.1) we choose randomly its coefficients λ_l^* from an uniform distribution on $[-2, 0]$ and we generate a spike-train of length $T = 4 \times 10^8$. Then we construct a prefix-tree from a sample of length $T_0 \ll T$ (typically $T_0 = 10^7$) taken from the generated spike-train. In this test we estimate the Gibbs potential knowing the monomials occurring in the generating potential ψ^* (i.e. only the λ_l's are to be determined). For each sample of length T_0 we propose a randomly chosen set of "initial guess" coefficients, used to start the estimation method, distributed according to $\tilde{\lambda}_l^{(0)} = \lambda_l^*(1 + (U[0,1] - 0.5)x/100)$, where x is the initial

Table 4.2: Numerical precision of the method in the same conditions as table 4.1. The Euclidean distance $|\bar{\lambda} - \tilde{\lambda}|$ between the estimated parameter's value $\tilde{\lambda}$ and the true parameter's value $\bar{\lambda}$ is reported here. We clearly observe and error increase but the method remains numerically stable.

ψ_1	R=1	R=2	R=4	R=8	R=16
N=1	5.0e-09	2.2e-02	6.3e-03	1.3e-02	6.9e-03
N=2	1.1e-08	1.3e-02	9.2e-03	5.2e-03	
N=4	8.0e-09	8.5e-03	6.8e-03		
N=8	3.8e-08	5.1e-03			

ψ_2	R=1	R=2	R=4	R=8	R=16
N=1	1.1e-10	1.9e-02	7.2e-03	4.8e-03	9.2e-02
N=2	1.1e-09	4.8e-03	3.7e-03	2.3e-03	
N=4	3.7e-08	2.6e-03	5.8e-02		
N=8	6.0e-06	2.4e-02			

percentage of bias from the original set of generating coefficients and $U[0,1]$ is a uniform random variable on $[0,1]$.

Test 1 Results Call $\tilde{\lambda}_l$ the values obtained after convergence of the algorithm. Our results show that:

(i) the error $E(|\tilde{\lambda}_l - \lambda_l^*|)$ increases with the range of the potential and it decreases with T_0; This error should scales as $\sim \frac{1}{\sqrt{(T_0)}}$ for each coefficient λ_l.

(ii) the error is independent of the initial bias percentage (see figs 4.1.2);

4.1.3 Test for comparison of Statistical Models

We select a potential ψ^* from (4.1); we choose randomly its coefficients λ_l^* from an uniform distribution in $[-2,0]$; we generate a spike-train of length $T = 1 \cdot 10^8$ and we construct the prefix-tree with the spike-train obtained. Using this prefix-tree we estimate the coefficients $\lambda_i^{(m)}$ that minimizes the KL divergence for several statistical models ψ_m proposed in (4.1). Therefore, in this test, the guess potentials have not necessarily the same parametric form as the generating potential: the monomials may be different as well as the number of monomials. The parametric coefficients $\lambda_l^{(m)}$ of potential ψ_m as well as $\tilde{h} = P[\psi_m] - \pi^{(T)}(\psi_m)$ are then averaged over 20 samples in order to compute error bars.

Our results show that :

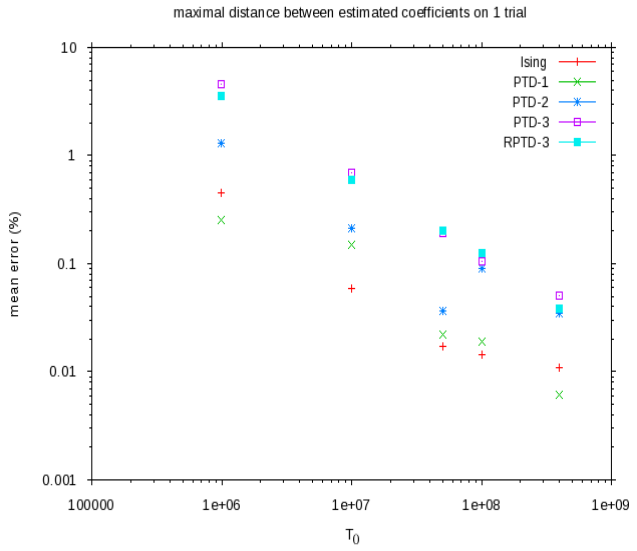

Figure 4.2: Mean error (in percentage) vs T_0 size. The line $400/\sqrt{(T_0)}$ is included as reference.

(i) The statistical models with lowest mean value KL divergence have the same monomials as ψ^*, plus possibly additional monomials, in agreement with section 2.5.5.

(ii) For all these models, the criterion $\tilde{h}\,[\psi]$ (2.38) averaged over trials, is fairly equal up to a difference of order $\delta \approx 10^{-6}$, while the difference with respect to other types of statistical models is at least of 4 orders of magnitude lower. We recall that, according to section 2.5.7, the deviation probability is of order to $\exp(-\delta T)$. After estimation from a raster generated with an Ising model, the ratio of the deviation probabilities (2.46) between an **Ising** and a **RPTD-**1 model is $\sim \eta = \exp(-0.0000115 \times 10^8)$, while between the **Ising** and the **PTD-**3 $\sim \eta = \exp(-0.00072 \times 10^8)$ meaning that the **PTD-**3 provide a worst estimation.

(iii) The value of the additional coefficients of an over-estimated model, corresponding to monomials absent in the parametric form of the generating potential, are null up to the numerical precision error. We call "best" model the one with the minimal number of coefficients. For example, as we checked, an **RPTD-**1 potential is as good as an **Ising** to approximate an **Ising**, but the additional coefficients are essentially null, so the "best" model to approximate an **Ising** is ... **Ising**.

(iv) The predicted probability of words matches the empirical value up to statistical errors induced by finite-sampling effects (fig. 3a, b; 4a, b).

The confidence region is estimated by $\mu_\psi(w) = \pi^{(T)}(w) \pm 3\sigma_w$ with σ_w given by eq.2.45.

In order to extend the model comparison we introduce the following notations: let w be a word (encoding a spiking pattern) of length R, $P_{est}(w)$ its mean probability over trials calculated with the estimated potential, $P_{emp}(w)$ its mean empirical average over trials (i.e average of form (2.33) including a time average $\pi^{(T)}$ and a sample average, where the samples are contiguous pieces of the raster of length $T_0 \ll T$), and $\sigma_{emp}(w)$ the standard deviation of $P_{emp}(w)$. We now describe the comparison methods.

We first use the box-plot method [65] which is intended to graphically depict groups of numerical data through their 'five-number summaries' namely: the smallest observation (sample minimum), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation (sample maximum)[1]. Figure 4.5 shows, in log-scale, the box-plot for the distribution of the quantity defined as:

$$\varepsilon(w) = \left| \left(P_{est}(w) - P_{emp}(w)\right) / \sigma_{emp}(w) \right| \tag{4.2}$$

that is taken as a weighted measure of the deviations. We have considered this distribution when it takes into account, either all the words up to a given size R_{max}, or only the words of that given size. There is no visual difference for $R_{max} = 7$. The results shows that only models containing the generating potential have the lower deviations value with very similar box. On the other hand a "bad" statistical model shows a much more extended error distribution .

Finally a χ^2 estimation is computed as $\chi^2 = \frac{1}{N_{\text{words}} - L} \sum_w \varepsilon(w)^2$ where $\varepsilon(w)$ is given by (4.2). Values are reported in tables 4.3, using all words or only those of size R_{max}. Since the number of words is high, it is clear that the lower the error, the lower the χ^2 estimated value. Note that χ^2 test assumes Gaussian fluctuations about the mean value, which are satisfied for finite-range Gibbs distributions, as can be easily seen by expanding the large deviations function I_l in (2.41) up to the second order in ϵ. However, when comparing two different Gibbs distributions it might be that the deviations from the expected value of one Gibbs distribution compared to the expected value

[1] The largest (smallest) observation is obtained using parameter dependent bounds, or "fences", to filter aberrant uninteresting deviations. Call $\beta = Q3 - Q1$ and let k denote the parameter value, usually between 1.0 and 2.0. Then the bound correspond to $Q3 + k\beta$ for the largest observation (and for the smallest one to $Q1 - k\beta$). A point x found above (below) is called "mild-outlier" if $Q3 + k < x < Q3 + 2k\beta$ (respectively, $Q1 - 2k\beta < x < Q3 - k\beta$) or extreme outlier if $x > Q3 + 2k\beta$ (respectively, $x < Q1 - 2k\beta$). We have used a fence coefficient $k = 2.0$ to look for outliers.

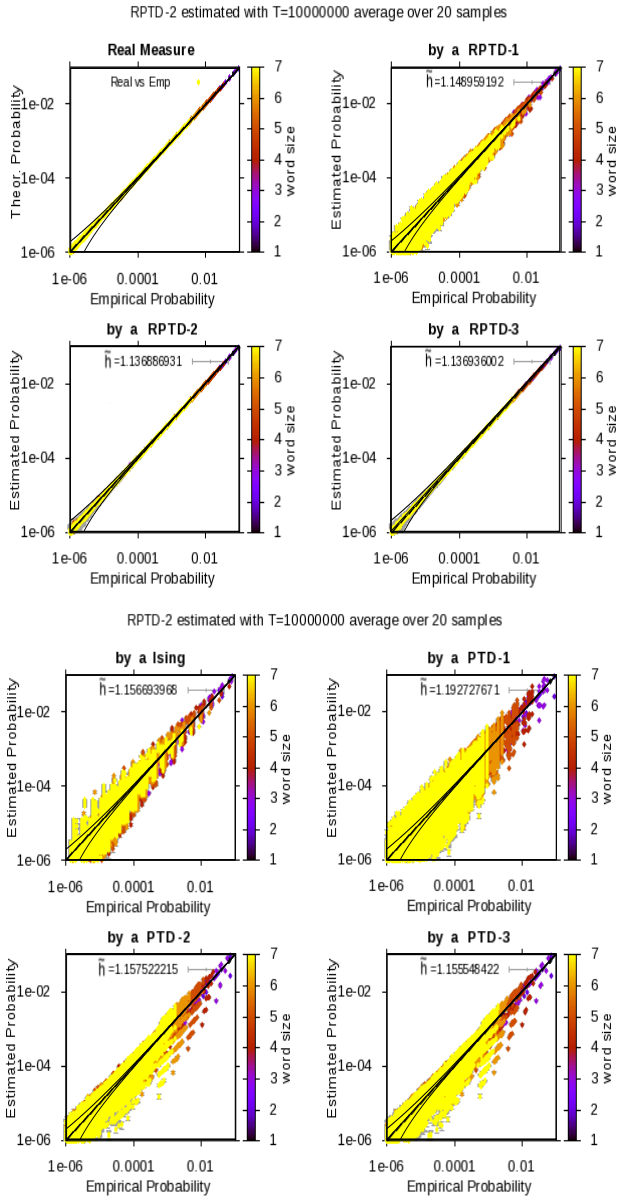

Figure 4.3: Figure 3a (top left) Expected probability μ_ψ versus empirical probability $\pi^{(T)}(w)$; Figure 3b (top right) to 8 (bottom right) Predicted probability versus empirical probability $\pi^{(T)}(w)$ for several guess potentials The generating potential is a **RPTD-2**. The confidence interval lines given by $\mu_\psi(w) = \pi^{(T)}(w) \pm 3\sigma_w$ with σ_w given by eq.2.45

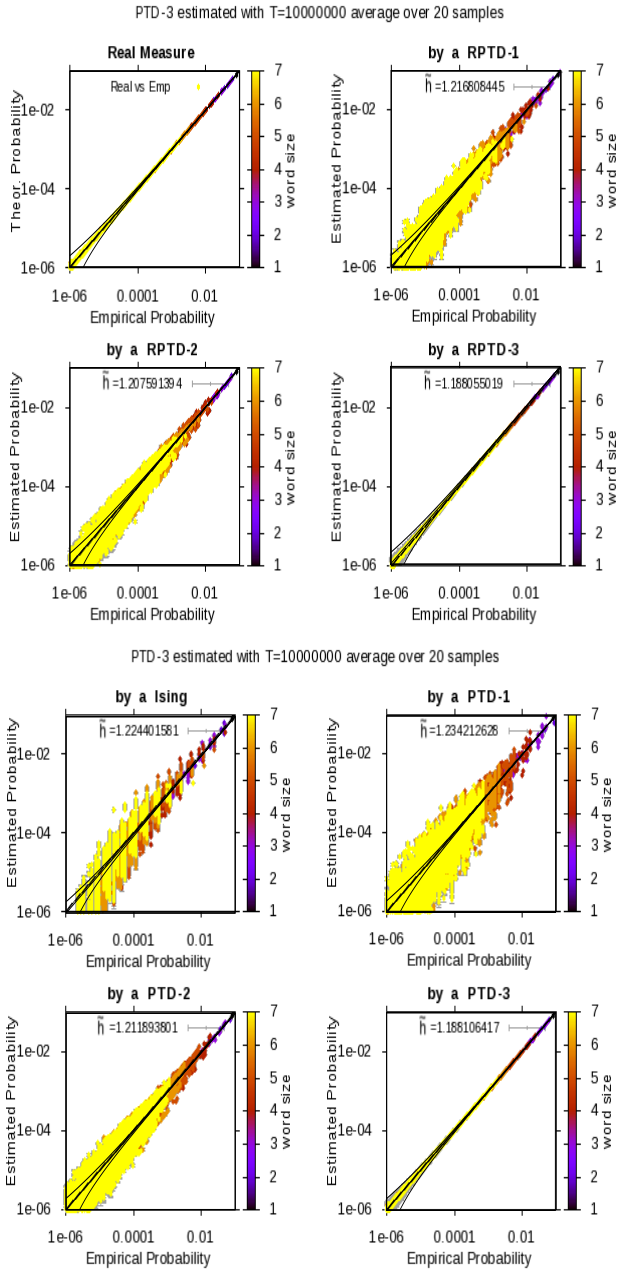

Figure 4.4: Same as previous figure where the generating potential is a **PTD-3**.

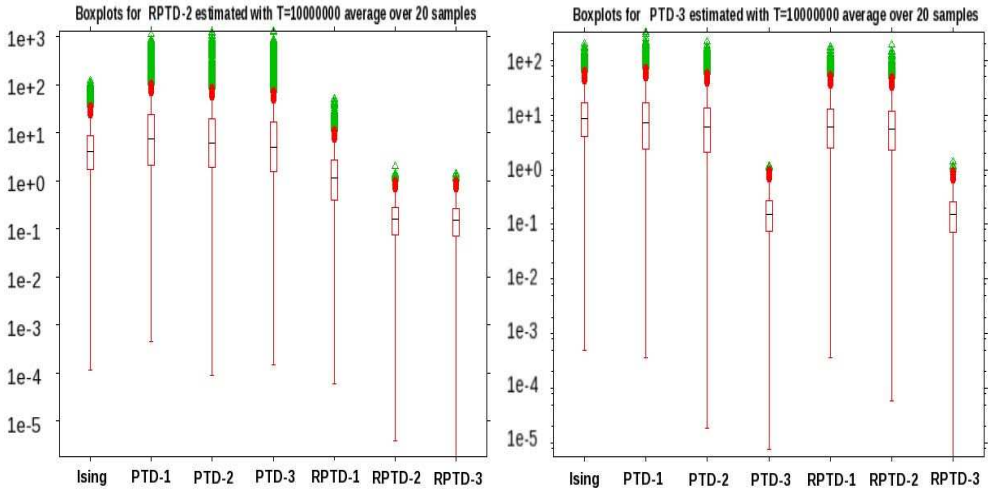

Figure 4.5: The box-plot (in log-scale) of the distributions of weighted deviations of word's probability versus their empirical probability, for several statistical models, using a generating potential of the form **(left) RPTD-**2 and **(right) PTD-**3. Midliers Outliers (see footnote 1) are shown by red dots and extreme outliers by green dots.

Table 4.3: χ^2 **coefficient calculated**: (left) with all words of size < 7; (right) with words of size 7 only. See text for details.

Estimating \Generating	RPTD-2	PTD-3
Ising	135.427	415.965
PTD-1	3146.17	564.396
PTD-2	3319.75	290.93
PTD-3	2533.35	0.0571905
RPTD-1	13.9287	274.773
RPTD-2	0.0607027	223.516
RPTD-3	0.0556114	0.0539691

Estimating \Generating	RPTD-2	PTD-3
Ising	121.825	347.502
PTD-1	2839.36	468.763
PTD-2	2537.39	229.255
PTD-3	2053.72	0.057065
RPTD-1	11.6167	218.458
RPTD-2	0.0605959	176.598
RPTD-3	0.0553242	0.0541206

of the other Gibbs distribution is well beyond the mean-square deviation of the Gaussian fluctuations distribution, giving rise to huge χ^2 coefficients, as we see in the tables 4.3.

4.1.4 Tests in the case of non-stationary parameters

Here we present results of the parameter estimation method applied to a spike train with statistics governed by a non-stationary statistical model of range 1, i.e. with time varying coefficients for rate or synchronization terms. Since the generation of spike-trains corresponding to more general higher time-order non-stationary process is not trivial, these potentials with higher range values will be analyzed in a forthcoming paper.

In the following we use an Ising potential form (4.1) with time-varying coefficients $\psi(t, [\omega]_0^0) = \lambda_1(t)\,\omega_1(0) + \lambda_2(t)\,\omega_2(0) + \lambda_3(t)\,\omega_1(0)\,\omega_2(0)$. The proce-

dure to generate a non stationary spike-train of length T is the following. We fix a time dependent form for the 3 coefficients $\lambda_i(t)$. From the initial value of the $\lambda_i' s$ (say at time t) we compute the invariant measure corresponding to the $\mathcal{L}(\psi)$ matrix. From this, we use equation (2.11) (with a time dependent matrix $\mathcal{L}$) computed using the next coefficient values $\lambda_i(t+1)$.

With the generated spike-train, we perform the parameter estimation, but computing the empirical average over a small fraction of the spike-train which means a time window of size $T_0 = \frac{T}{M} << T$. Then, we slide the observation window and after recalculating the empirical averages, we estimate again the coefficients value. We have verified that estimation procedure can recover correctly the coefficient values, for several types of time dependence, provided their variations be not too fast, and that the sliding window size be not too large with respect to T. In figure (4.6) We present the reconstruction of on of the parameters exhibiting a sinusoidal time-dependence given by $\lambda_0(t) = 0.4 + 0.3 \sin\left(\frac{4\pi t}{T - T_0}\right)$. The ability of the estimation scheme to provide such a good behavior respect time varying coefficients might outcomes from the fact that it is not ruled by a detailed balance assumption, although a deeper analysis of this properties is still to be done.

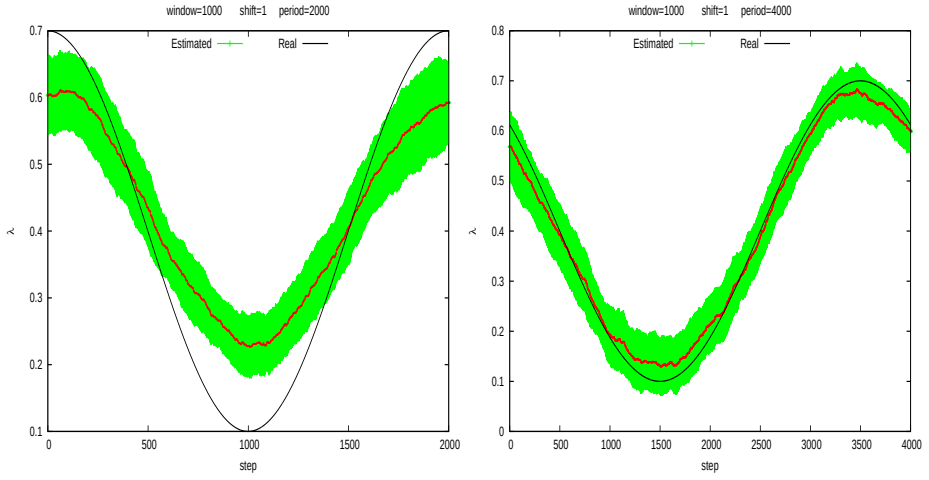

Figure 4.6: **Estimation of coefficients on a Non-Stationary process generated by an Ising model and sinusoidal time dependence**. Real value(black) and estimated parameter with its error bars (green) computed over 20 trials. The time shift is $\tau = 1$, Window size is fixed 1000, but oscillation period corresponds to 2000 (left) and 4000 (right).

4.2 Biological Data

In this section we will apply our framework of Gibbs distribution to analyze multicellular spike trains. Before presenting those results, we will comment briefly some delicate issues about pre-processing experimental raw data and

transform it into binary spike trains. However, we must state that we have not performed this preprocessing ourselves, and datases have been provided already as files of timestamps of the succesfull identified spike-trains after pre-processing the raw data issue from the electrophysiology recording. Finally,in the short future we will present a numerical test to show how changing the binning size (time snapshot delay) affects the information contained in data, from the point of view of Gibbs distributions with memory.

4.2.1 Preliminaries

The multicellular recordings

Pre-processing of recorded data is not free of procedural issues. A comparison of single units vs multi-units recording by [103] has shown that multielectrode recordings provides a reduced Signal-Noise ratio (SNR). However, it is worth to note that multi-electrode recordings in general might not avoid the single neighbor signals to mix up (electric potential effects) so the obtained signal needs to analyzed, spikes waveforms (templates) identified then ordered and separated) which is a delicate and highly computational consuming procedure known as Spike sorting [64]. The description of one of the state-of-the-art experimental setup (using tetrodes) and the spike sorting strategies is given in [56] with a discussion of the drawbacks found on previous multielectrode rigs and possible biases in older sorting methods. Finally, a non-exhaustive summary of Spike sorting techniques is available in [190].

The partition problem

In the preceding theoretical chapters and numerical applications we have assumed to have already obtained the multi-cell binary time sequences. However, the symbolic analysis of experimental time evolving data is far to be straightforward and we want to briefly comment on this point in the following.

The first aspect to consider is how to define the partition of the space domain into non overlapping regions and associated symbols so that any measurement value is assigned a single symbol s from a finite alphabet,representable as the integers $0, 1, \ldots, A-1$, which means how to discretise adequately the signal domain. The objective of this procedure is to obtain a symbolic sequence, to which one can apply methods of information theory, ergodic theory etc while hopefully preserving the invariants and main characteristics of the system dynamics (see section 5.1.2 for a detailed exposition).

Mathematically, the ideal partition is called “generating partition” and with it, the obtained symbolic orbits uniquely represent the continuous space orbit, and thus the symbolic and the continuous space dynamics are equivalent[22]. However, there is no standard way to do this when there is no available information about physical/topological properties of the system that helps one to make a choice. In fact, for the particular case of discretisation into binary sequences, where and arbitrary crossing threshold has been proposed, Bollt et al. in [16], has shown that systematic errors and misleading results might be produced by an inadequate choice of the threshold placement.

Fortunately, in the case of multicell membrane potential recordings, several evidences of the main role of action potentials lead us to focus on binary partitions denoting the event or absence of action potentials, and hence a natural way to perform signal discretisation arises. On the other hand, there may be other cases (such as analysis of fluorescence imaging data and/or recordings at other scales such as local field potentials) in which the partition problem is relevant. For such cases, some of the most promising strategies from a mathematical point of view, are based on obtaining successive refinements of the original partition. In some cases the equiprobability of symbols is intended, but it appears that a better choice for obtaining a more reliable sequence is to maximize the topological entropy of the obtained time sequence. In particular,[106] and several others have shown that estimation of orbits and parameter reconstruction improves for several deterministic systems under chaotic regime when using this protocol. However, up to our knowledge, which of that choices is more adequate for neural data is unknown, and further investigation of these strategies regarding neural experiments is still to be done.

The “Binning” problem

On previous paragraphs we have discussed as the problem of the signal domain partition as the first issue related to symbolic analysis of continuous systems. However, a second difficulty , which arises naturally, concerns the temporal dimension.

In neuroscience, focusing on spike events, provides a natural partition for the phase space. However the time domain problem about the choice of an adequate sampling rate remains still open. In order to create the symbolic sequence a sampling rate or equivalently a bin size needs to be chosen. However this choice is associated with the problem of symbolic redundancy which is similar to the problem of choosing an appropriate time interval for time-delay embedding ([47]). For instance, taking a very small bin increases biases due to oversampling and besides causes to unnecessarily enlarge

the range of time interactions, increasing the numerical effort to estimate a Gibbs distribution. On the other hand, taking a large bin size will cause over-smoothing or aliasing of the time-structure of data causing a lost of "information"[2]. for example, if a binary alphabet is maintained and more than a single spike of the observed neuron falls in the same bin.

In the framework of measure theory on symbolic sequences, called subshifts of finite type, some results are known in particular the Gibbs distributions . In this context, the binning process, is called a projection. It corresponds to a recoding of the original sequence into a new sequence defined on a new alphabet through the amalgamation of symbols. In the case when the invariant measure is a Markov measure, sufficient conditions were given for the projected measure (i.e the measure obtained from the Markovian measure for the binned sequence) to be also a Gibbs measure [32, 33]. Additionally, [104, 105] have stated rigorous conditions to deal with Gibbs measure over subshifts of finite type. However it remains unknown if the theorem conditions holds for experimental sequences and their usual binning procedure over binary spiking-time sequences. This appear difficult to verify because we would need to know before the binning procedure, whether the system dynamics is actually described by a Gibbs distribution and the explicit form of its Gibbs potential.

In the framework of symbolic analysis and dynamical systems theory, there is not a general consensus on how to choose the inter-symbol time interval and/or down-sampling the data. Most of current approaches are based on the auto-correlation function and/or the mutual-information function. Correlation functions are an indicator of linear dependencies between variables while mutual information measures and their generalizations are supposed to take into account all (linear and non-linear) interdependencies between variables. The behavior of the predicted underlying time-scale of dependence might differ between methods with highly complex sequences ([149]). A clear weakness of mutual information techniques is that they are highly limited by finite sample effects biases as they can be directly written as summations of entropies. The estimation of this entropies introduces several type of biases which has been shown to be sensitive to the underlying correlation structure itself ([111]). In [113], Li analyzed the relation between mutual information and correlation functions, where an exact relation for binary sequences is provided including and example for one-time step Markov chains. Li also included an analysis

[2]This need to be better specified, but for instance a coding strategy based in the inter-spike time delays.

of finite-size effects where he showed that mutual information true value is systematically overestimated. Later, [82] listed a complete comparison of weaknesses and advantages between correlation function and mutual information and extended Li's analysis of their relationship in sequences over alphabets of k symbols. It is shown that, statistical dependencies are measured by $(k-1)/2$ independent parameters.However, not all of them can be determined by auto-correlation functions. Appropriate sets of correlation functions (including cross-correlations) are introduced, which allow the detection of all dependencies. Their results are applied to analyze DNA sequences where long rage correlations are found. Finally, [77] have shown that the Jensen-Shanon Divergence can be used for segmentation of non stationary sequence into stationary sub-sequencies and apply the method to DNA sequences.

Within the perspective of the neuroscience community there is not a general consensus about how to choose the bin size for each type of experience and several strategies have been proposed depending of the specific use of data. Indeed, which most usually dictates the bin size is the (physiological) questions being asked and the time scale of behavioral events. For instance, during the analysis of spike train information from bowfly and expecting behavioral decisions in the 10-30 ms and a refractory period of 2ms, [131] are constrained to work with 2ms bins, while in [101] Kayser and collaborators analyze millisecond encoding precision of auditory cortex neurons which presents particularly high firing frequencies. Since refractory period in mammals is around 1ms, they stick to the use 1ms resolution windows. However, in past years, previously to the development of efficient multicells recording techniques, the most usual uses of binning methods in neuroscience was the construction of the Peri Stimulus Time-Histogram. This corresponds to a Non parametric kernel density estimation that requires the specification of smoothing parameters. Although a general maximum likelihood methods exists for this task (see [168]) the Mean Integrated Squared error has been the most current approach for time histogram analysis (see references in [174] [3]) although recently Bayesian and Gaussian kernel methods has been used ([44, 57]). In that paper it is shown that a bin size optimized for the unknown underlying spike rate may diverge for non-stationary cases so any time-histogram is likely to capture a spurious rate.

For the tasks of parametric model estimation for spike train statistics, mainly composed at the current state-of-the-art of Ising-models, an usual choice is to take a size equal to the mean single-neuron Inter-Spike Interval provided that it lies on the structured domain on the cross-correlation curve

[3] also `http://onlinestatbook.com/stat_sim/histogram/index.html`

[166]. Nevertheless, we remark that at the very beginning, the resolution sampling was usually used ([121]). A different choice which has been proposed is to take the size directly as the width of the central peak in cross-correlation histogram computed with the minimal sampling size (~ 1) ms (see [176]). Yet, in [191] authors has chosen the bin size as the average time interval between successive activations of two electrodes within a network burst when the inter-electrode spacing was $200\mu m$. As a result, for Maximum-entropy modeling the different binning practices have resulted in bin sizes around $4 - 20$ ms) and a small part of available spikes per neuron, usually ($\sim 1\%$), fall on the same bin while alphabet remaining binary[4]. In practice, it has been shown for the data sets used in literature that smaller bin sizes perform worst than bigger bin sizes ([191]) which can be explained by the sparse character of rasters and the fact that Ising models only take on account synchronization events which appear to be very sensitive to oversampling.

Cross-correlations and ISI-related binning proposal concern, by definition, only experimental information on pairs of neurons. However it would be interesting to obtain a method no restrained to pair of cells. In this perspective, the Gibbs distribution framework may provide an estimate of an adequate binning size that we briefly comment here but it will be developped. First we should estimate the parametric potential without binning, i.e, the minimal resolution sampling. Then we should need to compute the second largest eigenvalue of the associated Perron Frobenius matrix. The spectral gap, or difference between the first and second eigenvalues is related to the time-decay of correlations on the estimated probability distribution, providing an estimate of an acceptable bin size in order to preserve the temporal structure in data.

4.2.2 Analysis of Frog retina data

The following data set is courtesy of A. Palaces and it consist in a set of 5 neurons recorded from frog's retina. First, the experimental methods and technical details about data acquisition are briefly described. In the second part the our test results are reported. This results were part of an oral communication in the conference Neurocomp 2010.

[4]We remark this error add up to the spike-sorting process errors which will be described elsewhere

Methods

General Set-up description Adrian Palaces and collaborators are using multi-electrode arrays (MEAs) to record simultaneously from many retinal ganglion cells. Visual stimuli are presented to the retina using light emitting devices (LED) where all the stimuli parameters are controlled by computer. Palaces lab already has a working experimental rig consisting of a 64-channel PLEXON amplifier, a National Instruments A/D card (NI-6259, 16 bits, 1 MS/s), Labview homemade software (SpikeHunter) for data acquisition, (sampled at 20 KHz) control and exportation and visualization (SpikePlay). (PC), another computer to generate visual stimuli (PC), and external hard disks for data storage. High density recording arrays (HAD) (design, impendence and space between electrodes) are commercially available (Multichannel System). Each HAD array consist in 2 recording fields of 32 electrodes, separate by 500 μm, with electrodes of 10 μm diameter and separated by 30 μm. During an experiment, using standard protocols, a piece of retina is dissected out of the animal and mounted in a chamber on top of a MEA. Oxygenated AMES or L15 medium maintained at room temperature for amphibians and 35°C for rodents is perfused over the tissue to maintain its health. All experimental procedures were approved by the bioethics committee of the Universidad de Valparaiso in accordance with the bioethics regulation of the Chilean Research Council (CONICYT), and complied with International regulations (DHEW Publications, NIH 80-23).

Ganglion cell types identification was based on a MATLAB a spike-sorting module (Spike-It) used in complement with FIND `(http://find.bccn.uni-freiburg.de/)`.

Experimental information specific to data-set

- The animal is a Chilean Frog (Caudiverbera Caudiverbera).
- Stimuli was provided by a LED green (500nm) and consisted in flash emissions of 100ms of duration and with frequency of 0.5Hz. The total recording time was about 10.5s with a 1ms resolution.

Analysis of spike train-statistics

We have analyzed all possible pairs and triplets of neurons from the data set. We have considered **All-1,2,3** as statistical models and we have used bin sizes of $1, 2, 5$ms. Figure 4.7 shows, that the binning procedure has the apparent effect of increasing the pertinacy of long range interactions because for a statistical model of given fixed range, the $\tilde{h}$ shifts upwards with the

bin size (in the same fashion as data from of fishmovie, by O. Marre). Additionally is worth nothing that the $\tilde{h}$ value should not be compared between different rows as the related KL divergence lies on a different probability space (because of a different number of neurons). The best idea we can build about the quality of those estimations is provided by the points dispersion from the diagonal in the pattern rates diagram (see next figures).

The coefficients of the parametric estimation versus their monomial index in increasing order of lexical generation (i.e, the first terms are the rates, then the synchronization terms, then the 1-time step correlations and so on) are depicted for some sets of neurons in figures 4.8, 4.9. Results show that the binning reduces the relative value all monomials which signifies a lost of temporal structure. In general we could say that the shown sets of neurons do not present higher order/range correlations since most of the remaining monomials are practically zero, while those coefficients that remain are rather small to contribute strongly since in general, the larger the range and order of the observable the smaller its influence.

For the next result, we present the probability diagrams for the estimation of the **All-1,2,3** models with the same bin size for some pairs and some triplets. Several set of plots for bin sizes $1, 2, 5$ are presented in figures 4.10, 4.11, 4.12, 4.13, 4.14. Results shown that for pairs or neurons considered using bin size 1ms, a $\tilde{h} \approx 0.438$ is not too bad, and even range-1 models provide an acceptable good fit. For pairs with bin size 2ms, the acceptable $\tilde{h} \approx 0.69$. For triplets with bin size 1ms a $\tilde{h} \approx 0.62$ is a graphically a good. However, for certains ensembles a good fit is only achieved by the range-3 model. An example of a triplet with bin size 2ms where the range-1 model is enough for a good fit which is also shown in figure 4.12.

Finally, we remark that using a bin size of 5ms results in too short sampling size which appears insufficient to provide reliable estimates of longer words probabilities as indicated by the pronounced dispersion in the left bottom corner in figure 4.14.

4.2.3 A first analysis of Fishmovie in Salamander's retina

This data set was provided by courtesy of O. Marre and M. Berry II from Princeton University. The results here presented are a preliminary and unpublished and concerns only pair of neurons. Further analysis over larger groups of neurons are in development.

Methods

General description The recordings have been performed with high density multi-electrode arrays is build with 252 electrodes composed of ITO re-

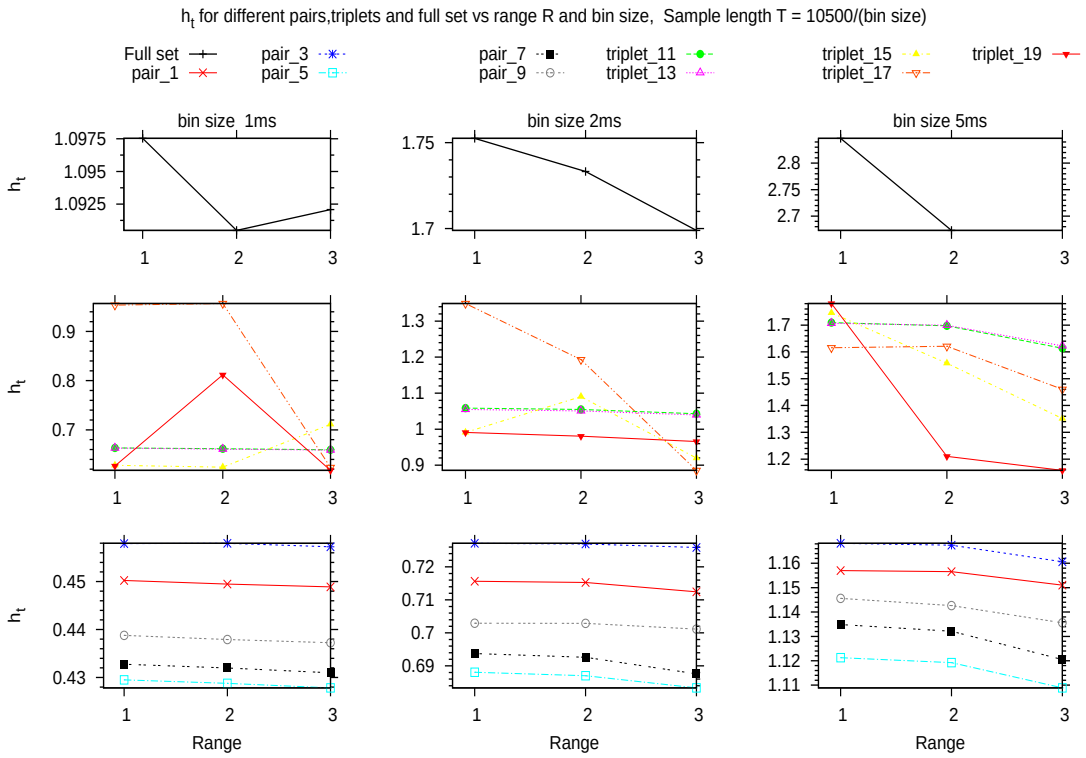

Figure 4.7: $\tilde{h}$ for several bin sizes, vs the model range. The top row corresponds to the full set of 5 neurons. The middle row for several pairs and the bottom row for triplets.

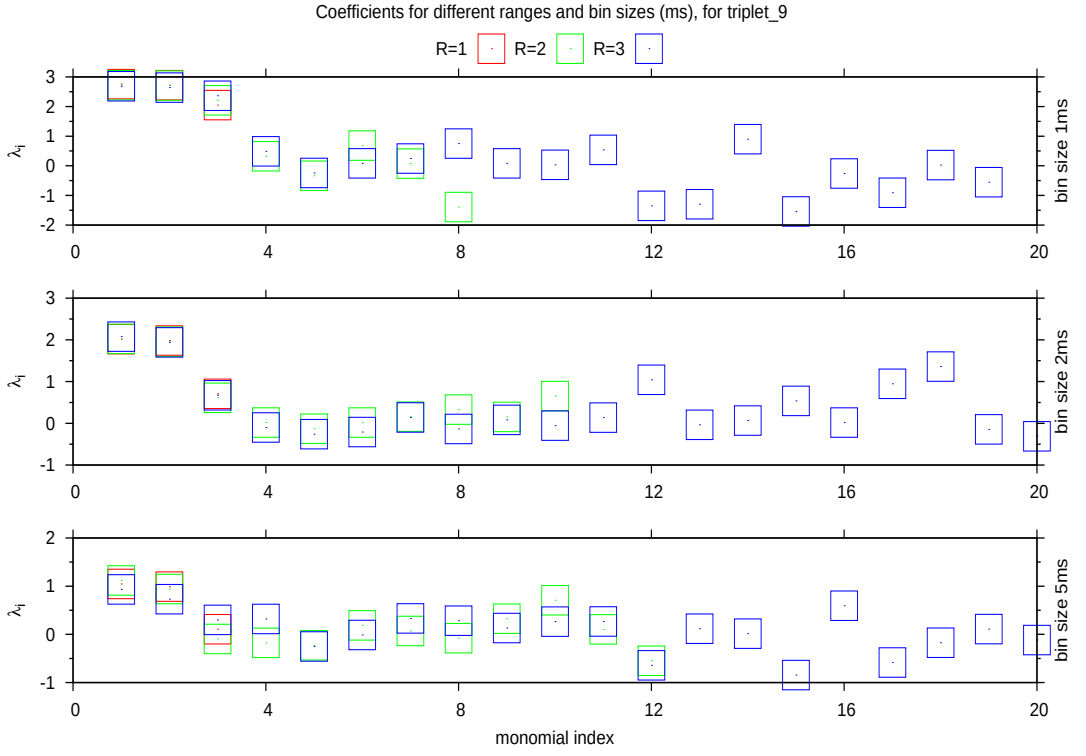

Figure 4.8: The Negative of Coefficient values for several bin sizes, and the model range for pair 9.

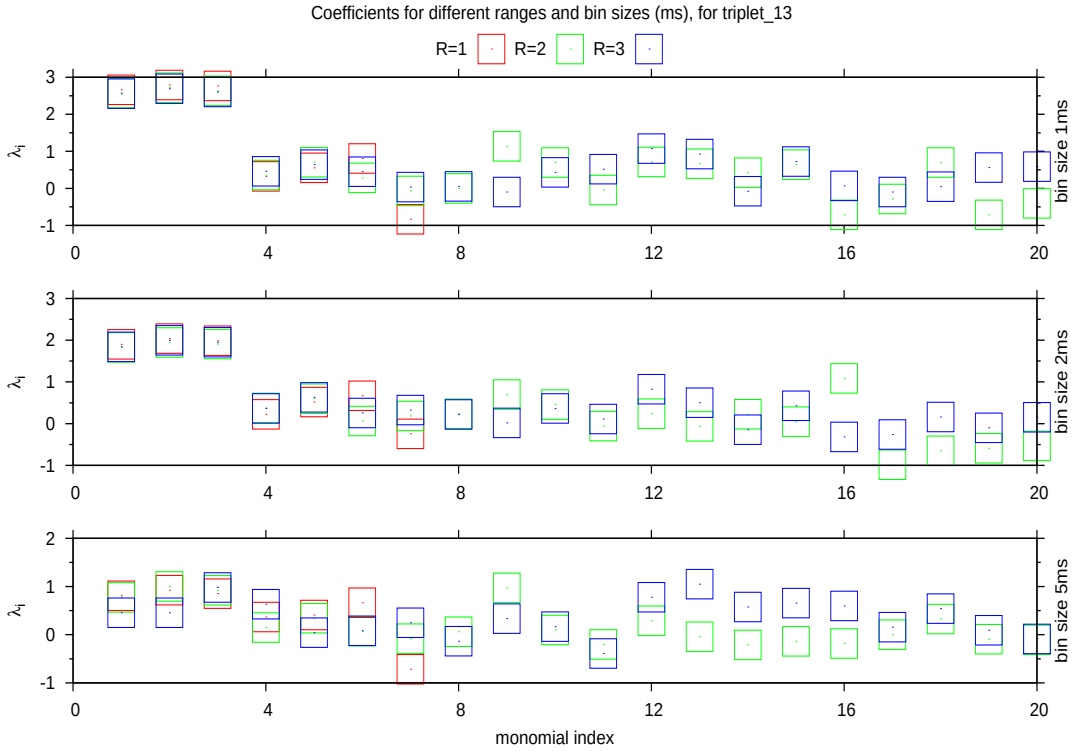

Figure 4.9: The negative of Coefficient values for several bin sizes, and the model range for triplet 13.

covered of platinum and isolated with nitride silicon. Each electrode has a diameter around 8 μm and is separated 30μm from the others. Because of their spacing, and given Ganglion cell density analysis (see methods in [170]), it might be assumed that each electrode registers spatially a single neuron. Reproducing the same protocol provided in [170], experiments were done on the larval tiger salamander (Ambystoma tigrinum) in accordance with institutional animal care standards. Retinas were isolated from the eye in darkness, placed with the ganglion cell layer facing a multi-electrode array (Multichannel Systems) and superfused with oxygenated (95% O_2/5% CO_2) Ringer's medium at room temperature (22 ÂřC). Extracellularly recorded signals were digitized at 10 kSamples/s and stored for offline analysis.

Spike-sorting Ganglion cell types identification and Spike sorting was performed with a new fast spike-sorting algorithm (-Publication in preparation-,personal communication with O. Marre), which is specially adapted to the recording of a large numbers of neurons. After spike-sorting process is performed, 154 good profiles were obtained with a recordings time scale is 0.1ms.

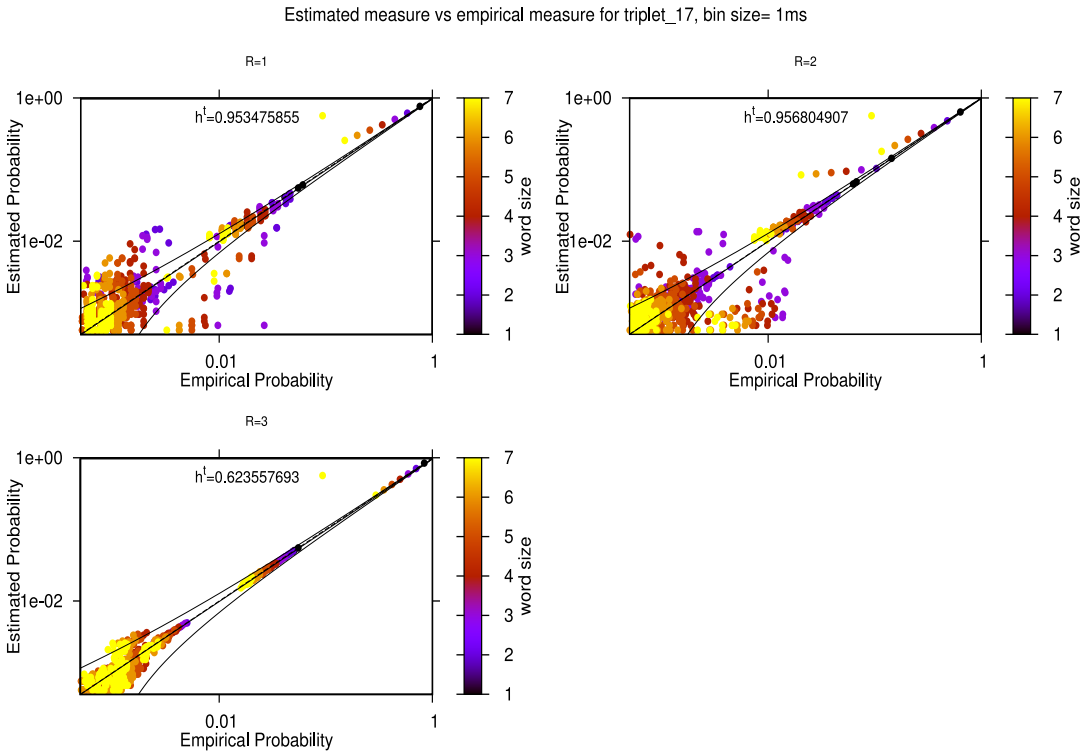

Figure 4.10: The estimated vs the empirical probabilities of each pattern, with different models **All-i** for triplet 13 using bin of 1ms. The error lines are $\mu_\psi(w) = \pi^{(T)}(w) \pm 3\sigma_w$ with σ_w given by eq.2.45

Stimuli The stimulus consist in a movie recorded in an aquarium where several fishes moving are observed. Therefore it is a natural film for an aquatic salamander. The stimulus alternates between the fishmovie played forward and then backward, and the same animated with a jitter mimicking eye movements. The eye movement sequence is always the same. Four different stimuli (films) were presented: Condition 1 is the fishmovie forward without eye movement, condition 2 the fishmovie backward without eye movement, condition 3 the fishmovie forward with eye movement, condition 4 the fishmovie backward with eye movement. The precise presentation scheme is depicted in figure 4.15, each presentation having the same time duration 120189ms and obtaining the following total presentation time (in ms) per film: film-1= 2163401.9 , film-2= 2043054.9 , film-3= 2043195.8, film-4= 2043198.4.

Preliminary spike-train analysis

In a first approach to data analysis, we will assume that dynamics is stationary during each different presentation, so we are allowed to break apart the whole raster plot recording, and concatenate all the presentations

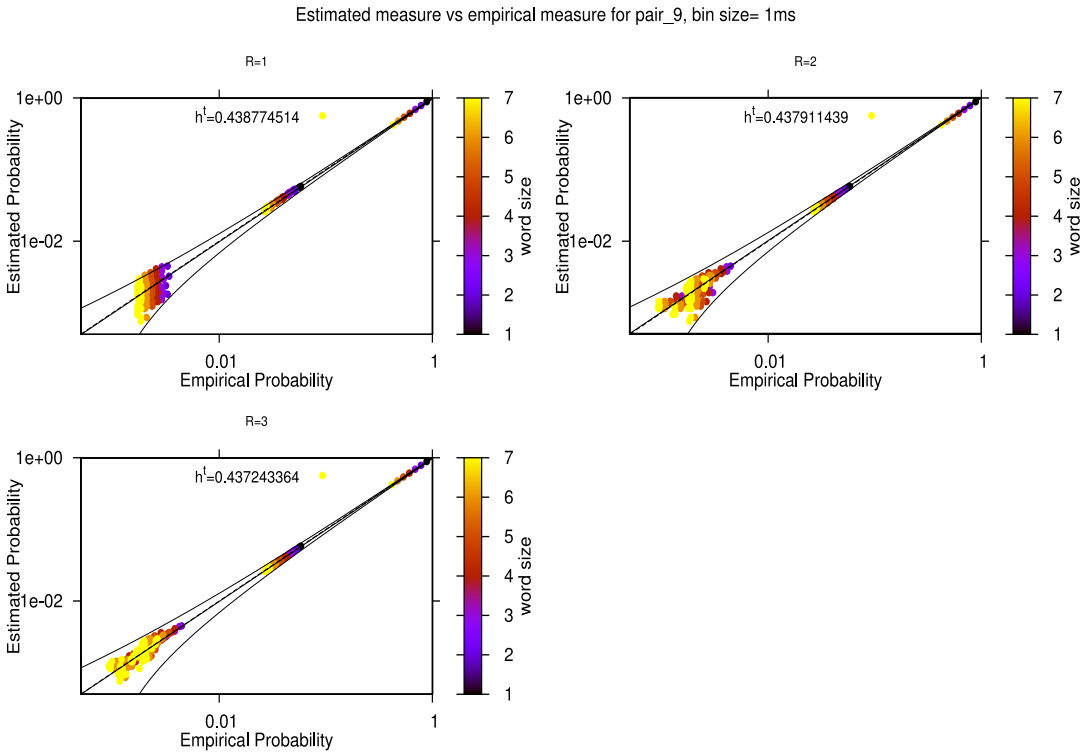

Figure 4.11: The estimated vs the empirical probabilities of each pattern, with different models **All-i** for pair 7 using bin of 1ms. The error lines are $\mu_\psi(w) = \pi^{(T)}(w) \pm 3\sigma_w$ with σ_w given by eq.2.45

belonging to the same film provided an appropriate recalibration of their spike-times to obtain a consecutive raster for the film, as if the films were presented in a single trial (the result is depicted in 4.16). If the assumption does not holds we need to perform statistics of observable over each film presentation and then to average over all the presentations of the film, which is not directly implemented in the current version of the EnaS library.

Our next step is to unravel first the general aspects of the neural activity during each film, especially the neural firing rates and its statistical distribution (see fig. 4.17). Assuming stationary firing rates, they lie between $[0.0015, 2.9]$Hz which means that the probability of a single firing in a time step scale of 1ms ranges between $[1.5 \cdot 10^{-6}, 0.0029]$. We observe that the firing rates distribution concentrate in the lower values. Let us to sort the neuron set by increasing order in their firing rate. In such manner we use their ordered position as a second way to refer such set, for instance as depicted in figure 4.19, where the lowest firing rate corresponds to ordered position 0 and neuron index 35, the second lowest firing rate in position 1 has index 113 ans so on, until the last neuron at position 154 is the neuron 70. Once ordered, we also observe that the subsets of most active and the

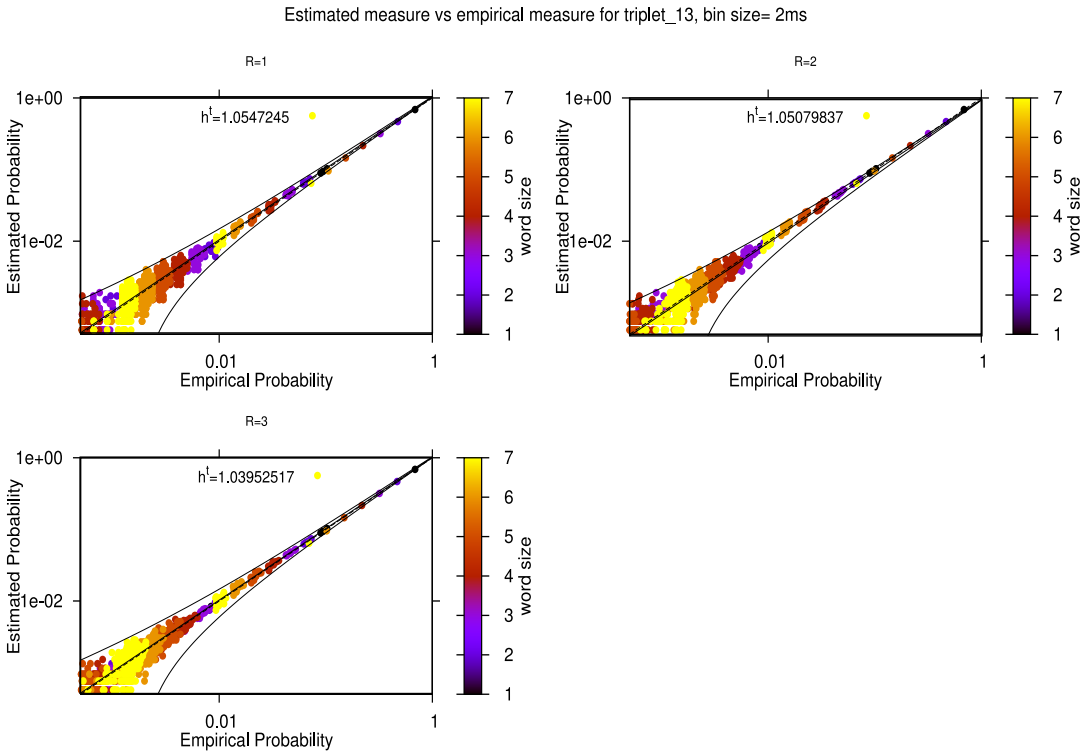

Figure 4.12: The estimated vs the empirical probabilities of each pattern, with different models **All-i** for triplet 13 using bin of 2ms. The error lines are $\mu_\psi(w) = \pi^{(T)}(w) \pm 3\sigma_w$ with σ_w

less active neurons are almost the same for all the films, although their ordered position might vary slightly between films (see fig. 4.19). On the other hand the intermediate-values set (ordered positions between $40 - 120$) may vary more between films, a single neuron shifting its ordered position more than 15 places, as it does neuron 5 between films 1 and 4. Finally, a remarkable difference exists between films $1, 3$ that have slightly lower firing probabilities that films $2, 4$ which is depicted in figure 4.20.

It is not straightforward to decide how to analyze neural activity with a so huge variety, going from few spikes to thousands of them in about the same recording duration around $2 \cdot 10^6$ms. In order to provide a simplified first approach analysis, we propose to focus on 4 different pairs of neurons per film, selected as follows:

- Pair 0 is composed by neurons with the 2 lowest firing rates (firing rate ordered indexes 0 and 1).
- Pair 1 is composed by neurons of the middle-low firing rate range (more precisely ordered indexes around 60 and 80).
- Pair 2 is composed by neurons of the middle-high firing rate range

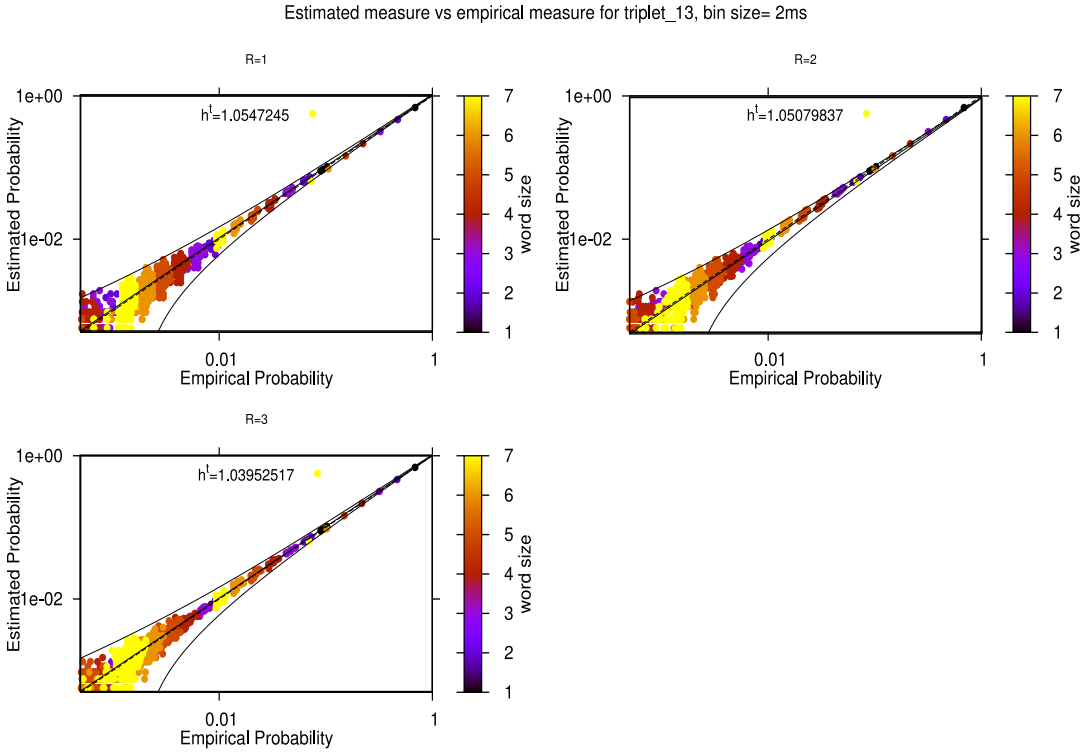

Figure 4.13: The estimated vs the empirical probabilities of each pattern, with different models **All-i** for pair 7 using bin of 2ms. The error lines are $\mu_\psi(w) = \pi^{(T)}(w) \pm 3\sigma_w$ with σ_w

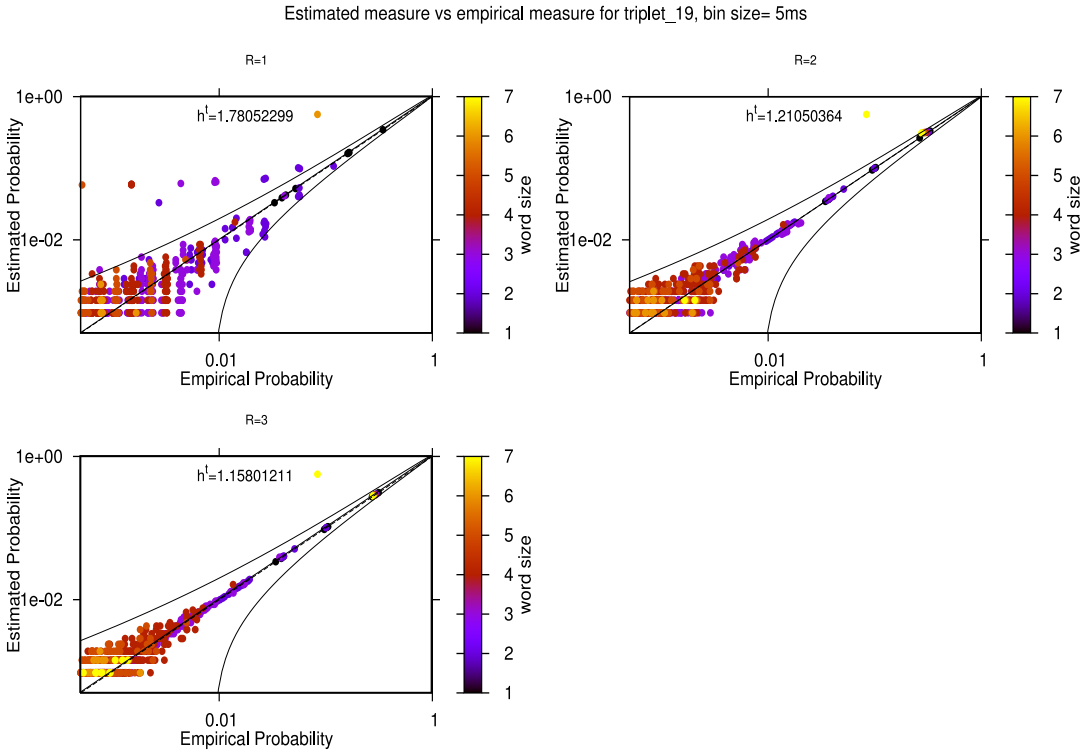

Figure 4.14: The estimated vs the empirical probabilities of each pattern, with different models **All-i** for triplet 19 using bin of 5ms.The error lines are $\mu_\psi(w) = \pi^{(T)}(w) \pm 3\sigma_w$

(more precisely ordered indexes around 100 and 120).

- Finally Pair 3 is composed by neurons with the 2 highest firing rates (ordered indexes 153 and 154).

We remark that with this choice of configuration, the real neurons corresponding to pairs 0 and 3 are the same for all films, but not for pairs 1 and 2. The overview of selected neurons activity is depicted in figures 4.21,4.22 (neurons are re indexed with labels 0 and 1, by its firing rate).

Additionally it is useful to examine the Inter-Spike Interval (ISI) distribution for pairs to be analyzed. We compute two different ISI distributions: the first one use fixed index (only Intervals on the same neuron are counted) while the second distribution, that we call ISI-mixed, is computed using the intervals between each spike and the latest spike on the other neuron. It is not difficult to show, that for not too high firing rates (as it is in this case), the ISI-mixed distribution behaves closely as the correlation between neurons. An step further, the ISI distribution gives us some qualitative information about how data sparseness will be affected by binning and a first key about more suitable and plausible monomials. Thanks to the ISI-mixed distribution taking account of both neurons we have a deeper qualitative insight of the binning effects, if not generally for all the neuron set, at least concerning the selected pairs. Indeed, in the short future we will present and include the crosscorrelograms for the selected pairs.

ISI-fixed results The total amount of ISI larger than 200ms for all pairs and films are shown together in figure 4.23. This results show that the middle-high firing rate neurons (corresponding to pair 2) have during films 1,2 and 4 the most clustered activity (most of produced inter-spikes interval between 65-75% are in the 200ms window), while the middle-low firing rate neurons (corresponding to pair 1) are a bit less clustered except during the film 3 where this pair appears with the lowest amount of ISI outside the 200ms window. The biggest difference in clustering between pairs 1 and 2 is found in film 2 where pair 1 has a clustered activity 24% lower than pair 2. On the other hand for the most active neurons (corresponding to pair 3), their clustering level appears almost constant during all films and being relative low (60% of ISI beyond 200ms). Finally and with no surprise as follows directly from raster plots, the less active neurons (pair 0) got the sparsest (less clustered) activity level with 80-90% of their ISI beyond 200ms.

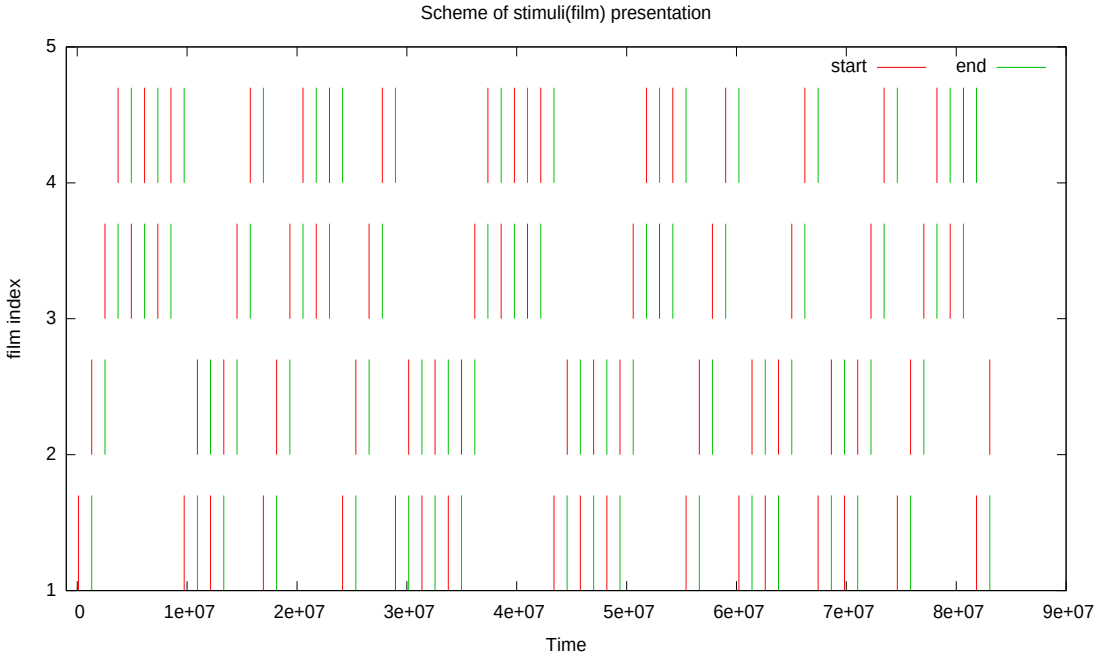

Figure 4.15: Stimuli(film) presentation scheme : Film index vs time (in 0.1ms). $film_{1,2,3,4}$. Each presentation begin time is depicted by red bars and the end time by green bars.

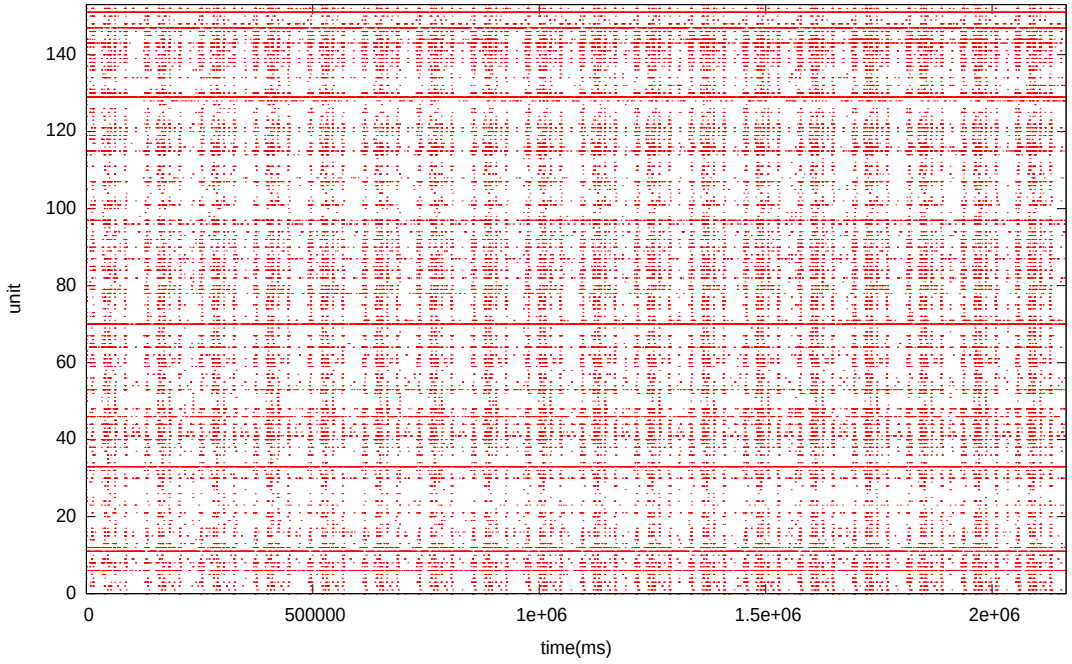

Figure 4.16: Raster plot for $film_1$.

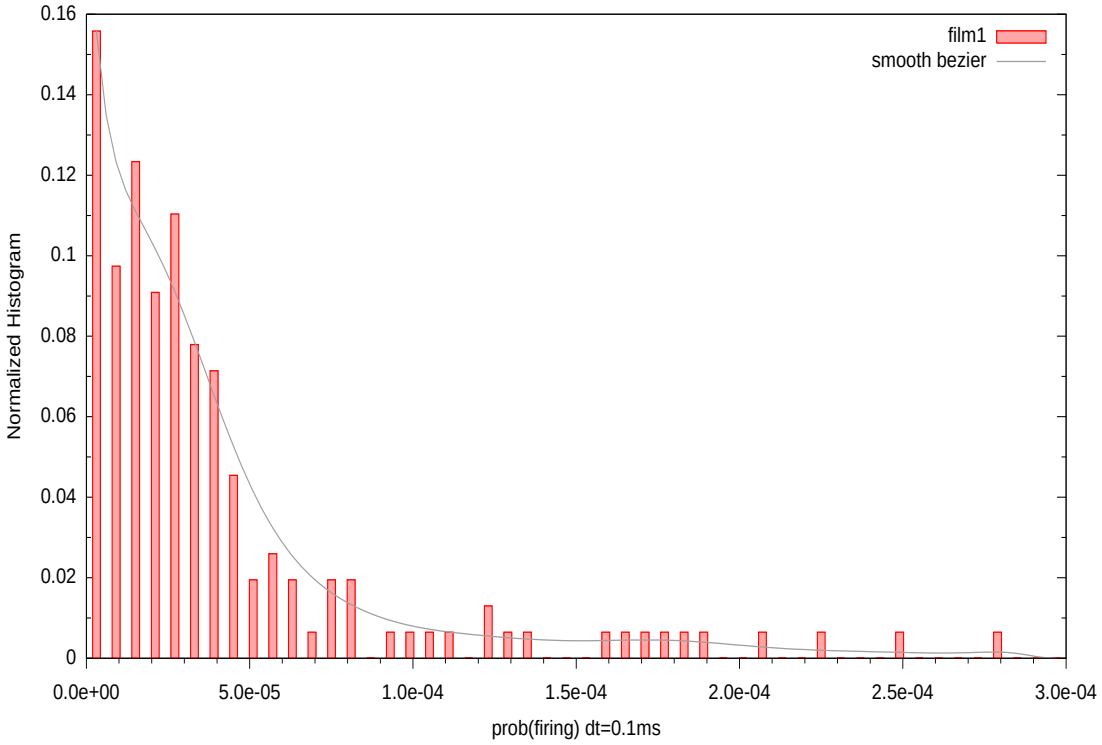

Figure 4.17: Normalized histogram vs firing probability at 0.1ms scale, for $film_1$ and smooth bezier curve.

ISI-mixed results While ISI-fixed index histograms tells us about the clustering of neural activity, it tells very few about the temporal lags and correlation between those activities. The total amount of ISI-mixed larger than 200ms for all pairs and films are shown together on a separated figure 4.24. It appears that for all selected pairs the number of inter-neuron ISI intervals larger than $200ms$ is at least of 60% (pair 3) and this percentage achieves indeed the 100% for pair 0. We also note that the lower percentage of ISI intervals larger than 200ms for both pairs 1 and 2 is achieved during film 1. While the ISI-mixed structure of high-rate neurons on pair 3 is almost between trials, pairs 1 and 2 show a diversity of distributions type, naturally of small percentage.

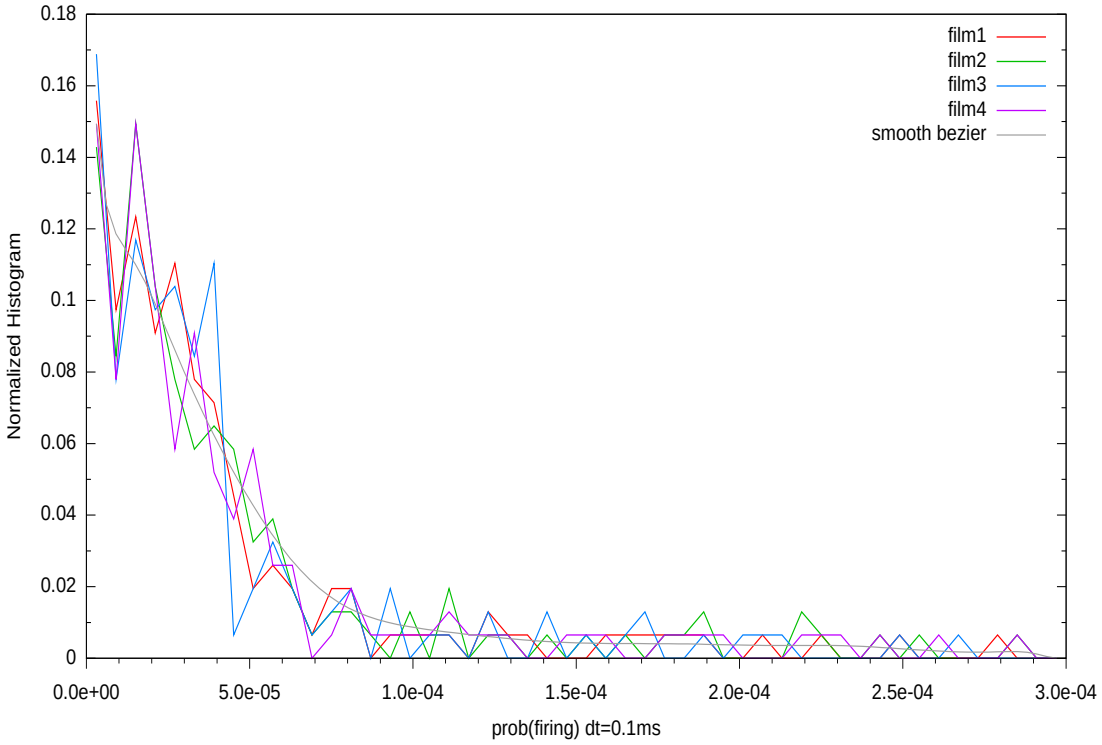

Figure 4.18: Normalized histogram vs firing probability at 0.1ms scale, for $film_{1,2,3,4}$ and smooth bezier curve for the latter.

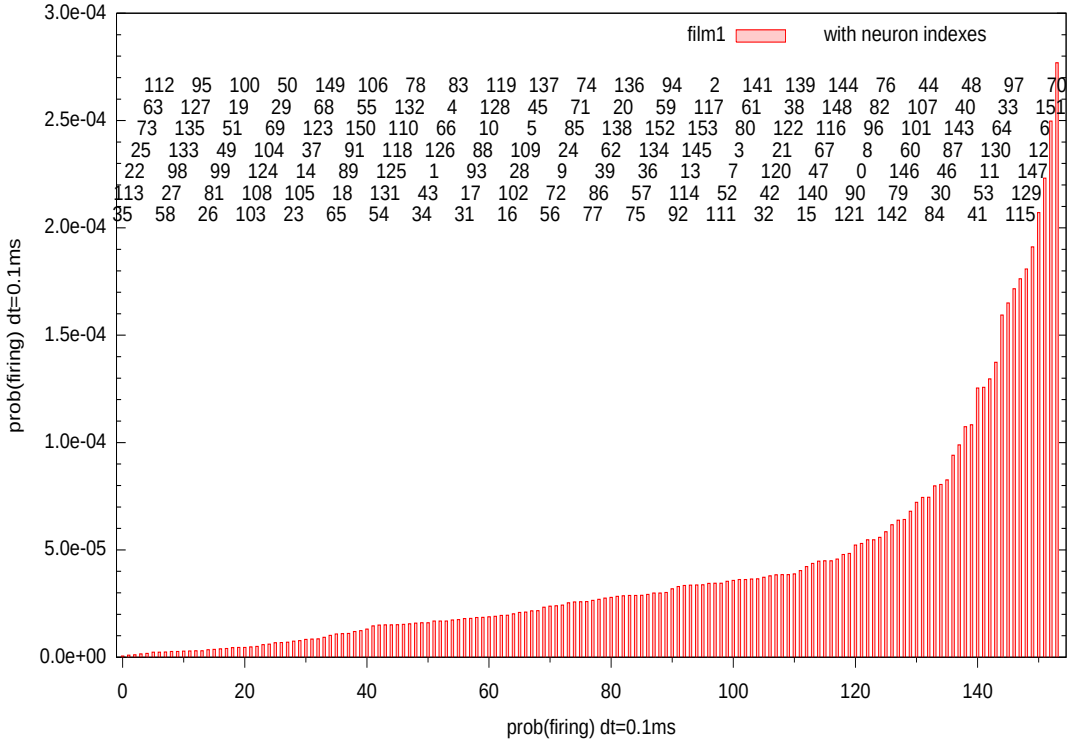

Figure 4.19: Firing probability within a 0.1ms bin vs ordered position for $film_1$ with indexes increasing always upwards and rightwards, i.e 35, 113,...112, 58,27,...,70.

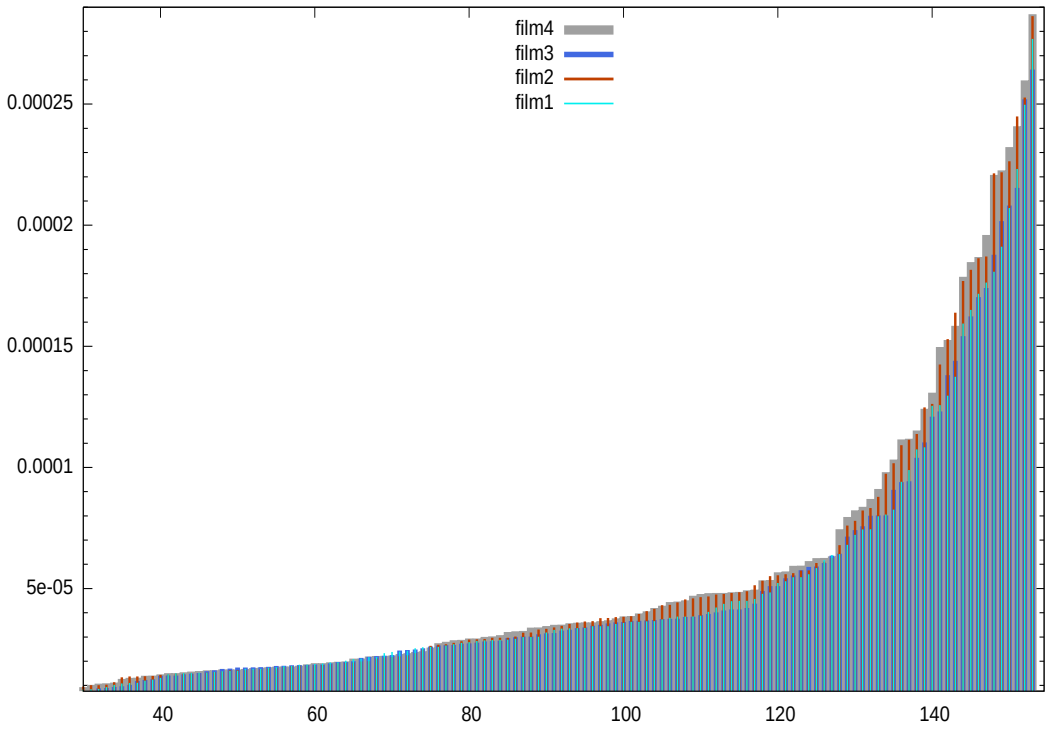

Figure 4.20: Firing probability within a 0.1ms bin vs ordered position for $film_{1,2,3,4}$

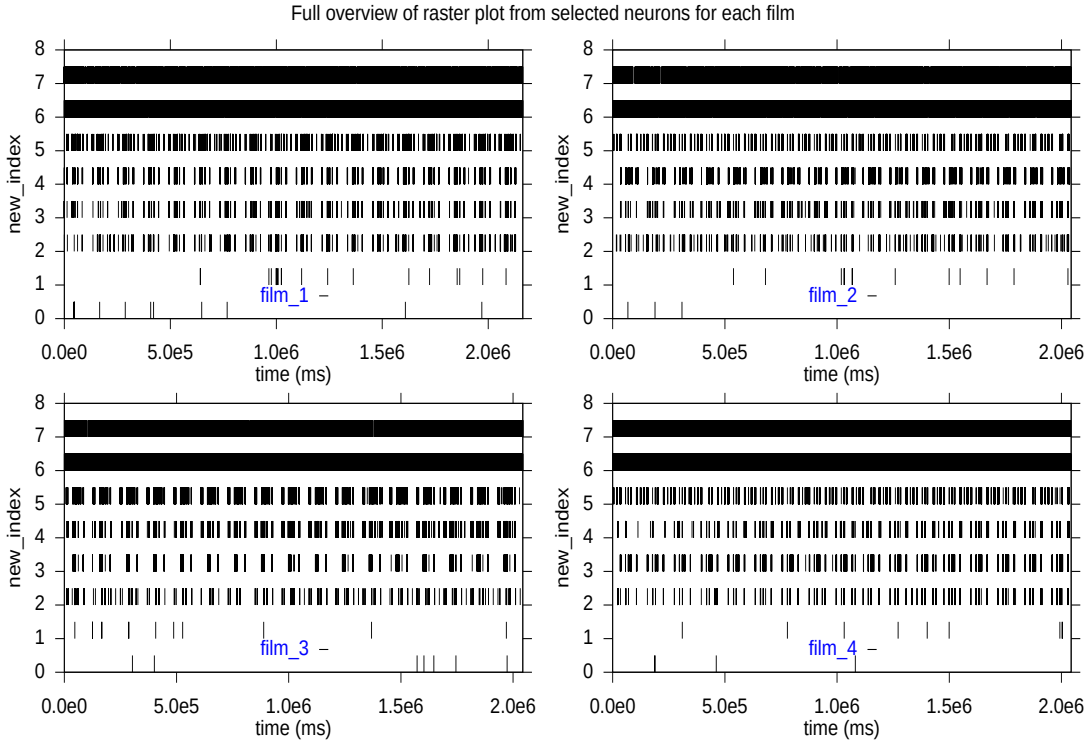

Figure 4.21: Full Raster plots for the selected 8 neurons New index vs time for $film_{1,2,3,4}$

What can be inferred from all this? We list our conclusions: Since there

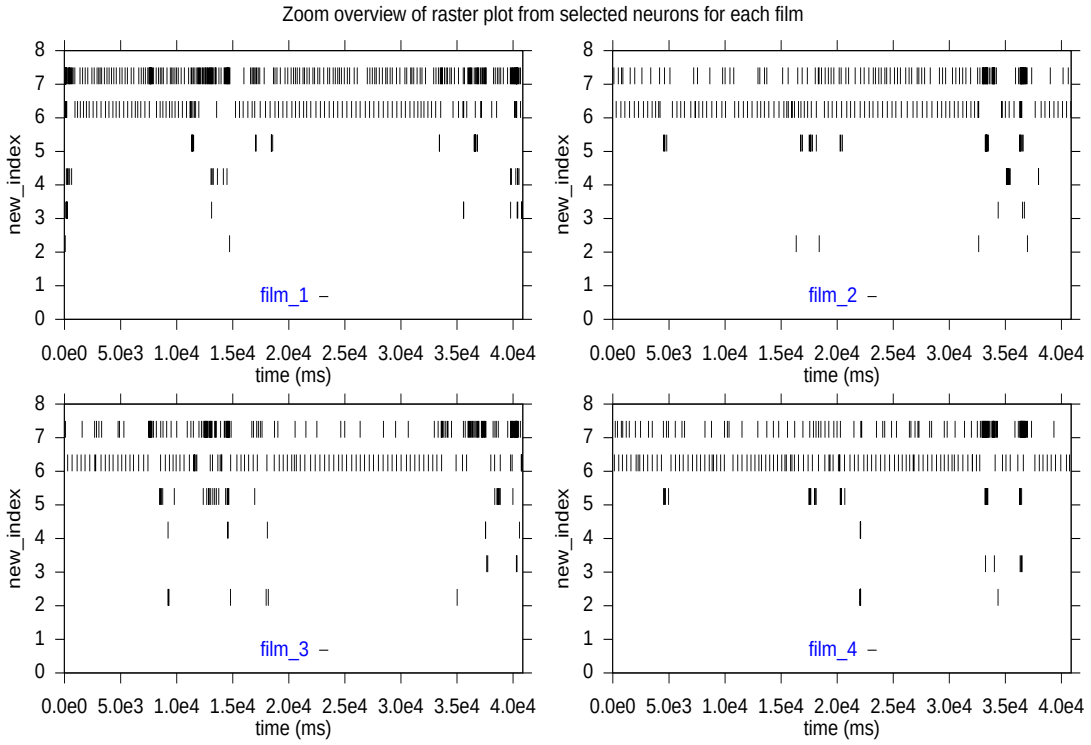

Figure 4.22: Zoom on Raster plots for the selected 8 neurons New index vs time for $film_{1,2,3,4}$

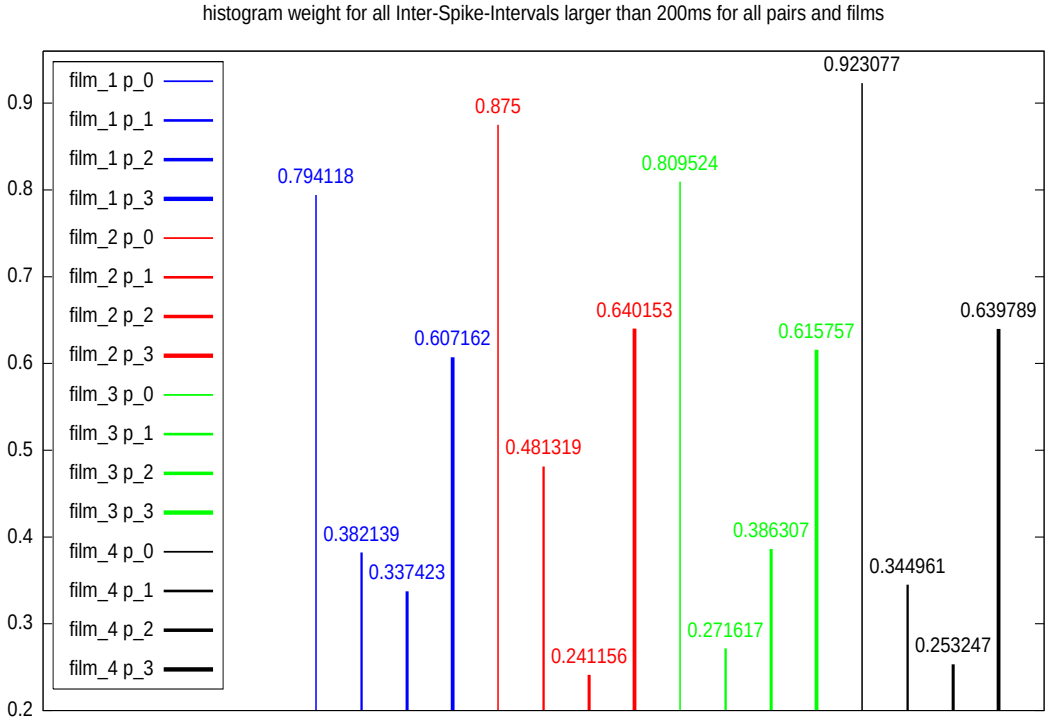

Figure 4.23: Weight corresponding to histograms with Inter-Spike-Intervals larger than 200ms, for all pairs and Films.

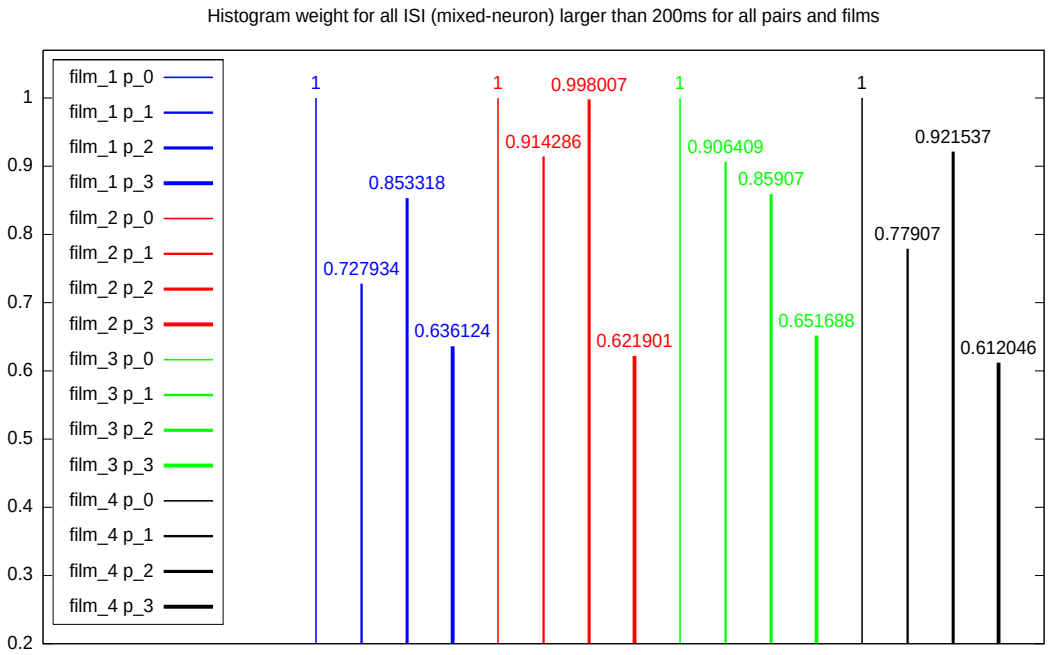

Figure 4.24: Weight corresponding to histograms with Mixed Inter-Spike-Intervals larger than 200ms, for all pairs and Films.

is a finite amount of data, we are restrained to treat small range Gibbs potential estimations, up to range 2 to maintain a small level of fluctuations. Moreover, a millisecond scale modeling seems useless because of the sparseness of data and its very small pairwise correlations in the 0-200ms range. Therefore, we will examine an **All-2** model (full monomial potential of range 2), explicitly given by equation

$$\begin{aligned}&\text{All-2 } (N=2, L=12)\\ \psi(\omega) = \omega_1(1) \;+\; &\omega_0(1) \;+\; \omega_0(1)\omega_1(1) \;+\; \omega_1(0)\omega_1(1) \;+\; \omega_0(0)\omega_1(1) \;+\; \omega_1(0)\omega_0(1)\\ &+\; \omega_0(0)\omega_0(1) \;+\; \omega_1(0)\omega_0(1)\omega_1(1) \;+\; \omega_0(0)\omega_0(1)\omega_1(1)\\ &+\; \omega_0(0)\omega_1(0)\omega_1(1) \;+\; \omega_0(0)\omega_1(0)\omega_0(1) \;+\; \omega_0(0)\omega_1(0)\omega_0(1)\omega_1(1)\end{aligned}$$

while binning the data using bin sizes of $10, 20, 30, 40$ ms.

The $\tilde{h}$ for all pairs and films, from different bin sizes are depicted in figure 4.25. Binning increases the KL error because, by reducing the sparseness of the spike-train then long memory effects become more representative. Hence a fixed low-range model becomes less adequate to describe the probabilities or larger blocks. Additionally it is natural that pair 0, which presents very sparse but asynchronous spikes, to be easily described by a simple model composed only of few monomials corresponding to firing rates terms and eventually monomials of consecutive spikes in silent state (pos-

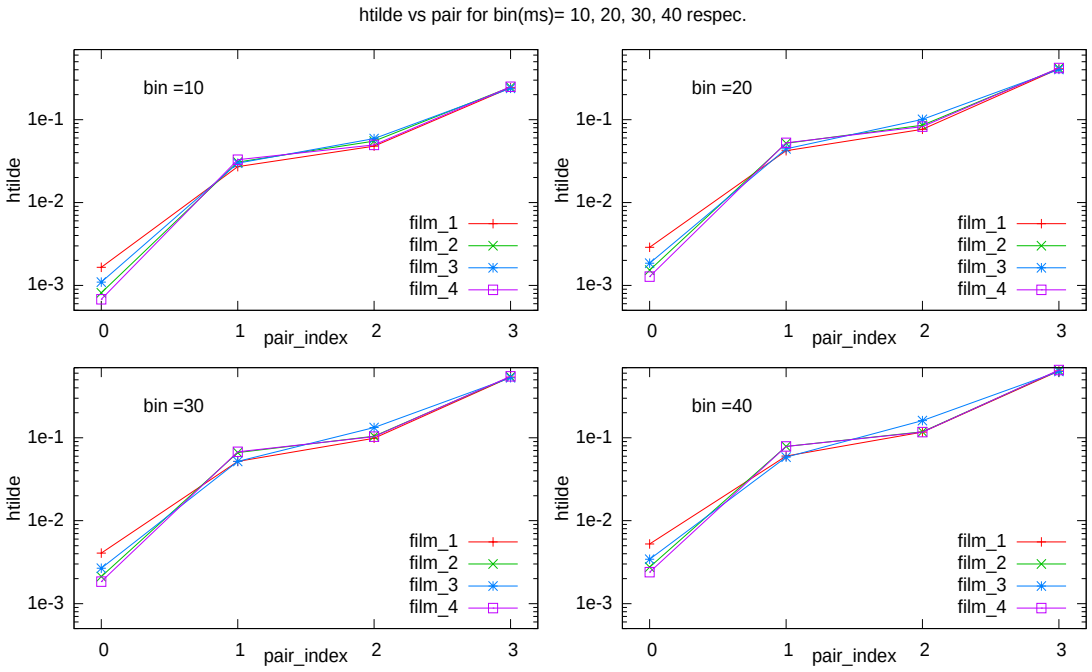

Figure 4.25: Show $\tilde{h}$ for all pairs and films for different bin sizes.

itive coefficients). This explains that it is pair 0 which has the lowest KL error. Moreover, the fact that the relative performance of each pair respect the others remains similar with several binning sizes might indicates that the KL error scales equally with the number of available spikes (i.e the firing rates of the neuron pair), since the set of available spikes convey the information defining (or enabling to estimate) the distribution. Finally, it is clear that given data sparseness, that the used binning sizes are still insufficient in order to a range 2 model be able to describe precisely the full temporal structure of the binned raster, in other words, that the effective memory of the underlying Markov chain after binning is still greater than 2. However, given a fixed range model, in the limit of an adequately large bin-size close the KL divergence respect that model should start instead to decrease.

The **negative value** of estimated parameters obtained for film 1 with several bin sizes is depicted in fig. 4.26. We remark that the coefficients values of the Gibbs distribution can be related with effective interaction structure and the retina structure. However this analysis remains out of the scope of this manuscript although some analytical results are known for theoretical neural dynamical models [25]. For the obtained results we list the following observations (we refer to the actual value of estimated parameters):

- Larger bin sizes imply a small decrease of firing rate coefficients.
- The pair 0, corresponding to the lowest and sparsest neural activity

presents their associated firing rate coefficients with the most positive values, and for larger bin sizes with the most negative coefficients for monomials corresponding to consecutive firings

- Pairwise coefficients reflects the anti-correlation as most of them are negative, while some are near from zero. Only for pair 3 with bin size 10ms the coefficients of one bin step self-correlation corresponding to index 3 and 6 are negative.

- Triplets monomial coefficients are diverse, from negative to positive values, but their small value added to their relative smaller monomial average indicates a small role in the potential characterization.

- For several pairs their coefficient values do not vary too much with binning. This might be also an indication that the memory scale of the underlying effective Markov chain is still larger than 2, in other words, for this binning size the spike trains statistics such that it is estimated by a range-2 approximation has not been altered.

Finally, we present the probability diagrams for the estimation of the **All-2** model for each pair during film 1, using bin sizes 10 and 20 ms (see figure 4.27, 4.28) . The probability diagram depicts the estimated probability of words as computed with the estimated Gibbs distribution versus the empirical probability of different words Since for the pair 0 the firing rates are so small, the probability of most of the possible words is zero. More precisely, for pair 0 only the words conformed by a single spike and several contiguous '0' symbols appears with almost the same frequency, up to several word lengths, so their corresponding position make them to overlie, which explains that only a few number of points appears in the pair 0 plot. It is evident that certain number of word probabilities are over estimated, but their empirical probabilities which are below $1 \cdot 10^{-5}$ are just in the limit of their statistical pertinacy since we recall the total sample duration is $2 \cdot 10^6$ms, that under 10ms bins provide an effective sampling length of $2 \cdot 10^5$.

4.2.4 Analysis on rat retina data

The following data set is courtesy of Bogdan Kolomiets in retina rat. The technical methods concerning this dataset have been described in [107]. The particularity of this dataset is that information about the recording channels correctly separated has been provided, so an analysis of results with different inter-cell distances. The recoding last about 30 sec. of spontaneous activity and 60 sec. of evoked potential.

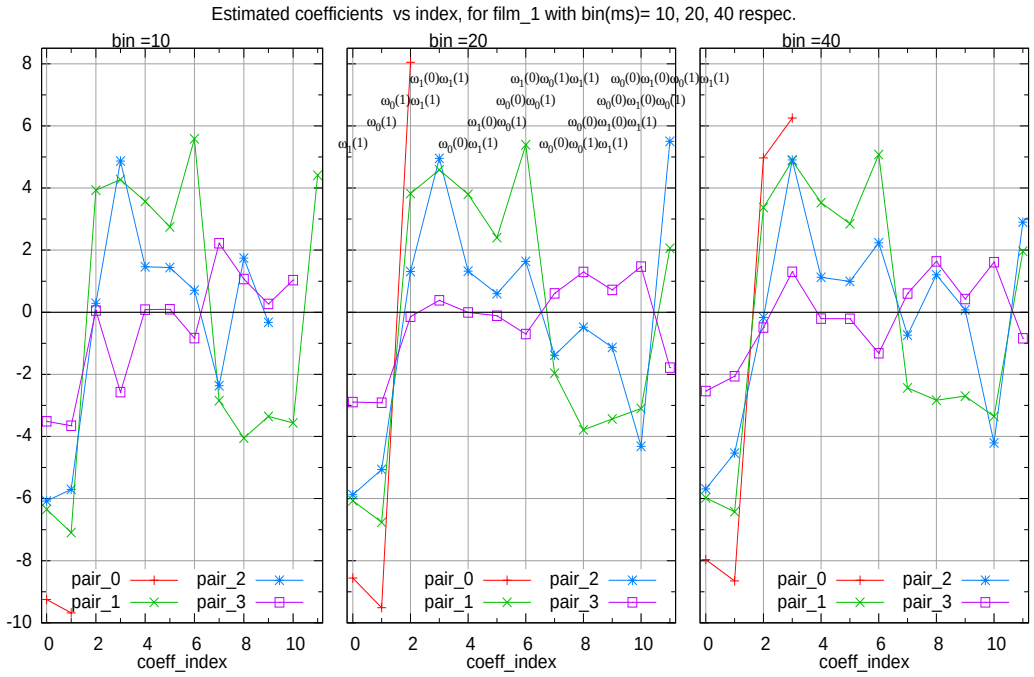

Figure 4.26: Negative value (i.e. $-\lambda$) of coefficients estimated for an All-2 potential for all pairs with different bin sizes on film 1.

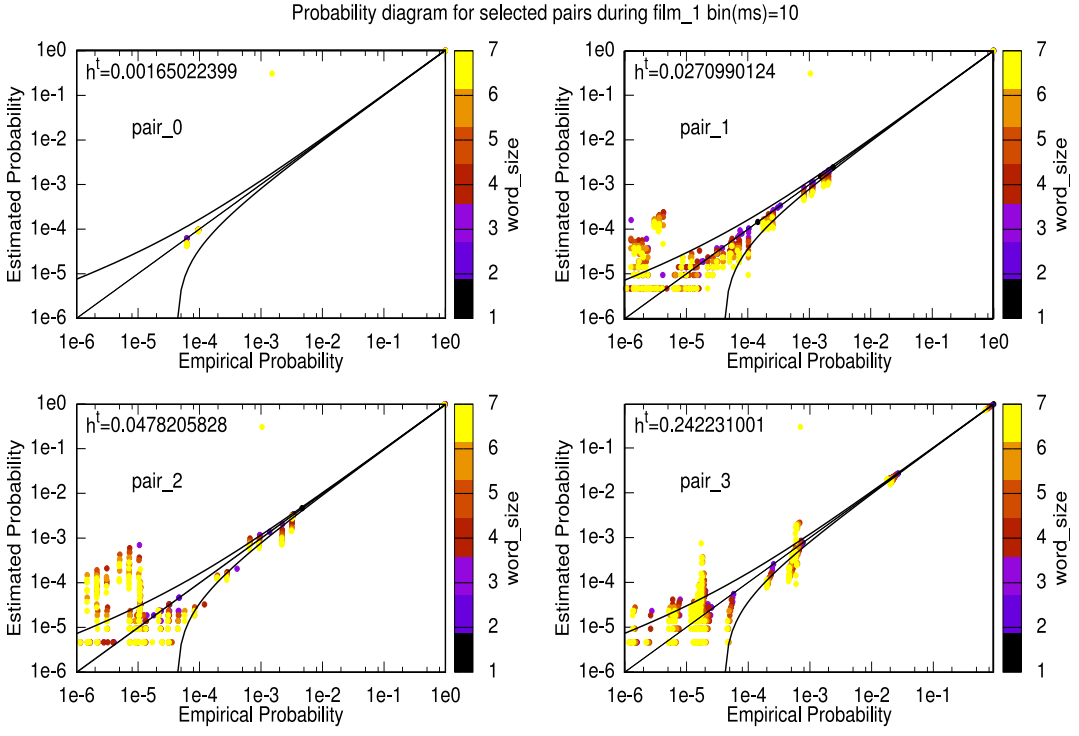

Figure 4.27: Probability of words vs Empirical Probability. The estimation is done with an All-2 potential for all pairs with bin size 10ms on film 1. The error lines are $\mu_\psi(w) = \pi^{(T)}(w) \pm 3\sigma_w$ with σ_w given by eq.2.45

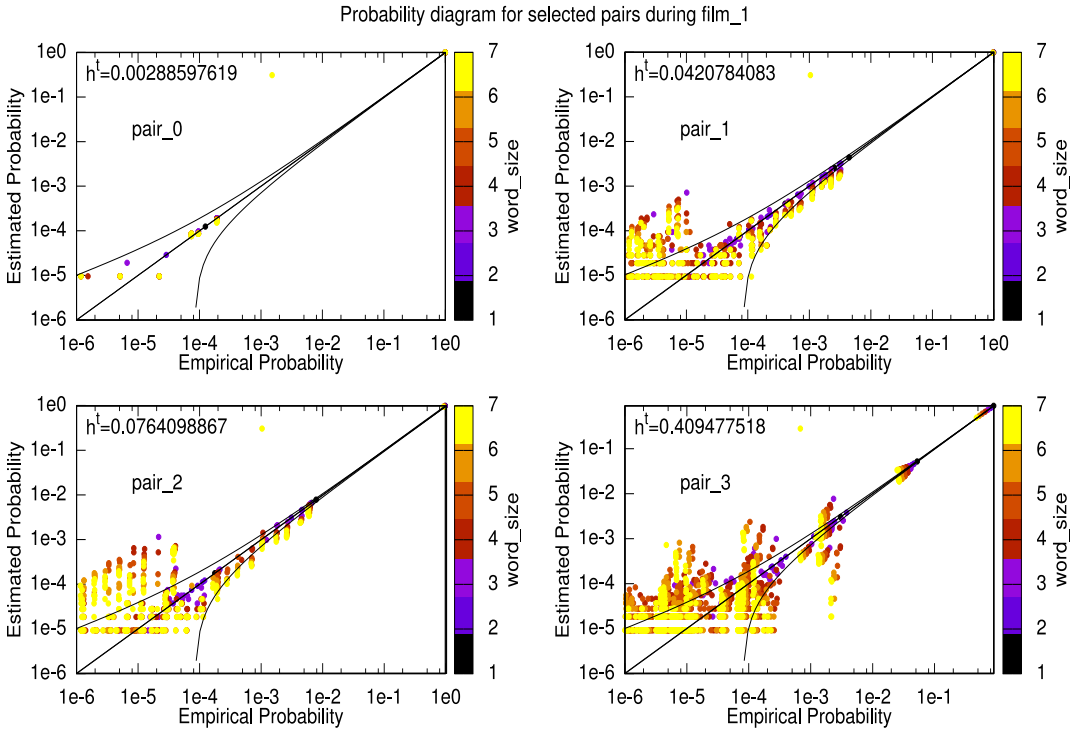

Figure 4.28: Probability of words vs Empirical Probability. The estimation is done with an All-2 potential for all pairs with bin size 20ms on film 1. The error lines are $\mu_\psi(w) = \pi^{(T)}(w) \pm 3\sigma_w$ with σ_w given by eq.2.45

Analysis of spike train-statistics

The following results correspond to an ongoing work, made in collaboration with Hassle Nasser using EnaS library. In order to have hints on the REC circuitry, we applied Ising model to analyze statistics of the data concerning groups of 7 neurons. The potential is given by:

$$\psi_{w'} = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \lambda_{ij} \omega_i(0) \omega_j(0).$$

Figure 4.29 shows the positions of the electrode composing the MEA array.

Figure 4.30 shows the $\boldsymbol{\lambda}$ coefficients for two different sets of 7 neurons (N0-N6). Knowing that RGC are connected with Gap Juction this sould imply a positive connectivity. However since the values obtained are negative it seems that this arise from larger order neural interactions.

Figure 4.31 shows the λ_{ij} for neurons at several distances. These units present different inter-cell separation. We have found that the coefficient value first increases and then decrease after a distance of about 300 μm. The interpretation of this behavior is still in progress.

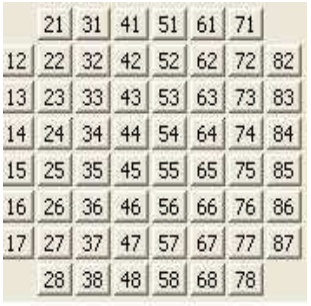

Figure 4.29: Scheme and positions of the electrodes composing the MEA array.

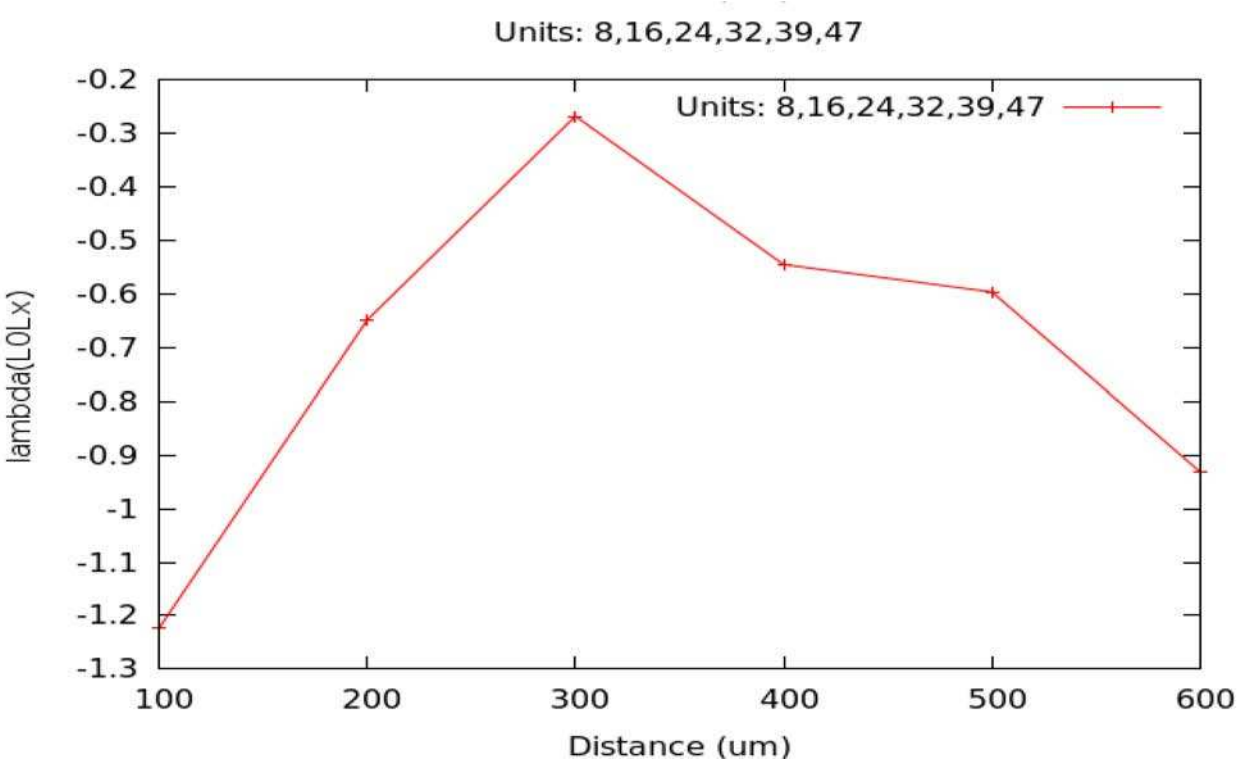

Figure 4.30: Figure shows the λ coefficients for two different sets of 7 neurons (N0-N6). The parameters L_i are associated to the firing rate and report positive values. The L_{ij} denotes the pairwise parameters and they give an idea about the RGC interaction.

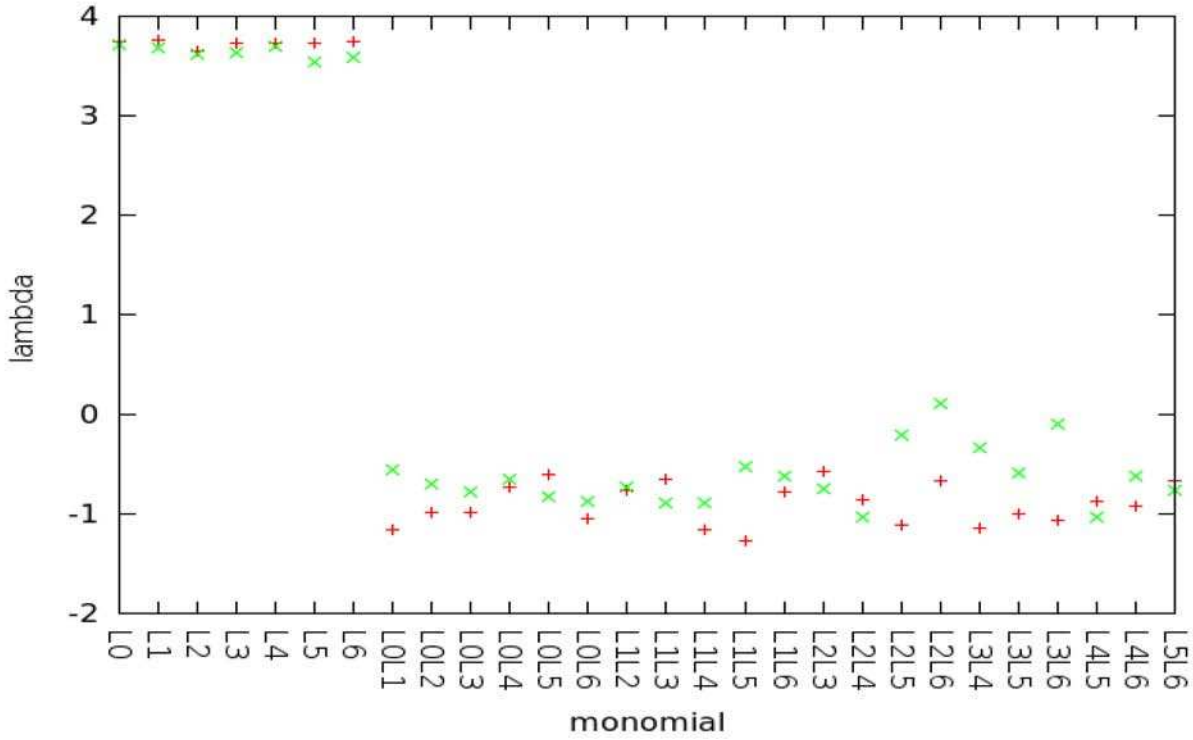

Figure 4.31: Figure shows the λ_{ij} for neurons at several distances. The units are the indexes of neurons in the MEA chip. We have renamed L0LX the coefficients where the index 0 corresponds to the unit 8, while the index X corresponds to units: 16, 24, 32, 39, 47. These units are located at various distances from the unit 8.

4.3 Binning effects: a simple case study with synthetic data

At the current stage we have made an analysis of binning effects on a toy model using ENAS Library. These results need to be further developped and extended to more general potentials.

Protocol The simplest case we analyze is through the spike-train statistics of a single neuron with potential of memory $\tau + 1$ given by $\boldsymbol{\psi}_\tau = -\lambda_0 w_0(\tau) + -\lambda_1 w_0(\tau) w_0(0)$. We generate a spike-train from this potential of lenght $T_{max} = 5 \cdot 10^7$ corresponding to time-sampling of 1. We fix a bin size $b = 2, 3, 4, 5$ and obtain the associated binned spike-train ($b = 1$ means that the original spike-train remains unchanged). Then we proceed to estimate the Gibbs dristribution for different guess potentials $\boldsymbol{\psi}_{\tau'}$ with $\tau' = 1, 2, 3, 4, 5$ using always a fixed length $T = 1 \cdot 10^7$ to avoid finite size effects to mix up with the effects especifically associated with the binning procedure.

Test results When binning, the best fitting potential (i.e. $\boldsymbol{\psi}_{\tau'}$ with the lowest KL-divergence) with respect the binned sequence appears to have lower memory than the original potential $\boldsymbol{\psi}_\tau$. However, this process is far from being a continuous one and a deeper understanding requires the analysis of amalgamation of symbols and grammar effects of binning. As our result shows, there are two generalities we may state:

- The best fitting potential $\boldsymbol{\psi}_{\tau'}$ is such that outperforms all the other candidates, i.e, for the difference of KL divergences $\eta = \tilde{h}(\boldsymbol{\psi}_{\tau'}, \pi^{(T)}) - \tilde{h}(\boldsymbol{\psi}_\tau, \pi^{(T)})$ between all the possible combinations of pair of potentials and the binned sequence the surface locates above zero. We have found that, if the original potential ($\boldsymbol{\psi}_\tau$) has a memory-lag value $\tau + 1$ which divides exactly by the bin size b, then the best suited potential of the form $\boldsymbol{\psi}_{\tau'}$ for the binned sequence will have a memory-lag equal to the ratio $\tau' + 1 = (\tau + 1)/b$. Figure 4.32 shows a $\boldsymbol{\psi}_\tau$ potential with this correct memory lag is shown outperforming another choice of potential. Their KL difference is always above zero positive at different regions of the surface (λ_0, λ_1) and , in particular it is of order $\sim 10^{-3}$.

- If the bin size b do not divides exactly $\tau + 1$, there would be not only one but two best suited potentials $\boldsymbol{\psi}_{\tau'}$, with $\tau' + 1$ corresponding to the two integer values closest to the obtained ratio $\tau' + 1 = \lfloor (\tau + 1)/b \rfloor$ and $\tau' + 1 = \lfloor (\tau + 1)/b \rfloor + 1$, where $\lfloor \rfloor$ denotes the integer part fonction. These two models will outperform each other at different regions of the surface (λ_0, λ_1), although the KL difference between both has magnitud

order $\sim 10^{-5}$, much smaller than their difference agaist other models, . This small difference is depicted in figure 4.33.

- Furthermore, if there is no positive τ which divides exactly by the bin size b, but still $b < \tau + 1$, it appears that by side effect, the best fitting potential is the ψ_1 but this is not very significatif since it is not too much better than the other models.

- Finally if $b \geq \tau+1$, the original memory structure has been fully hidden by the binning and therefore all ψ_τ models describe equivalently the spike-train statistics.

As a conclusion, for this simple family of time-correlated models we can bin the sequence hopefully without changing a lot the structure while eveidently the use of a very large bin size makes one to loose all the memory structure. However, in a more complex scenario the simultaneous presence of several monomials associated to different time-correlations might increase the sequence sensibility to amalgamation effects induced by binning procedures in a way that such "ideal bin size" does not exist. Finally, we recall through figure 4.34 that the best approximated statistical description of a binned sequence might can not be as good as the best approximated statistical description of the unbinned raster; evidently, in general binning causes a loose of information.

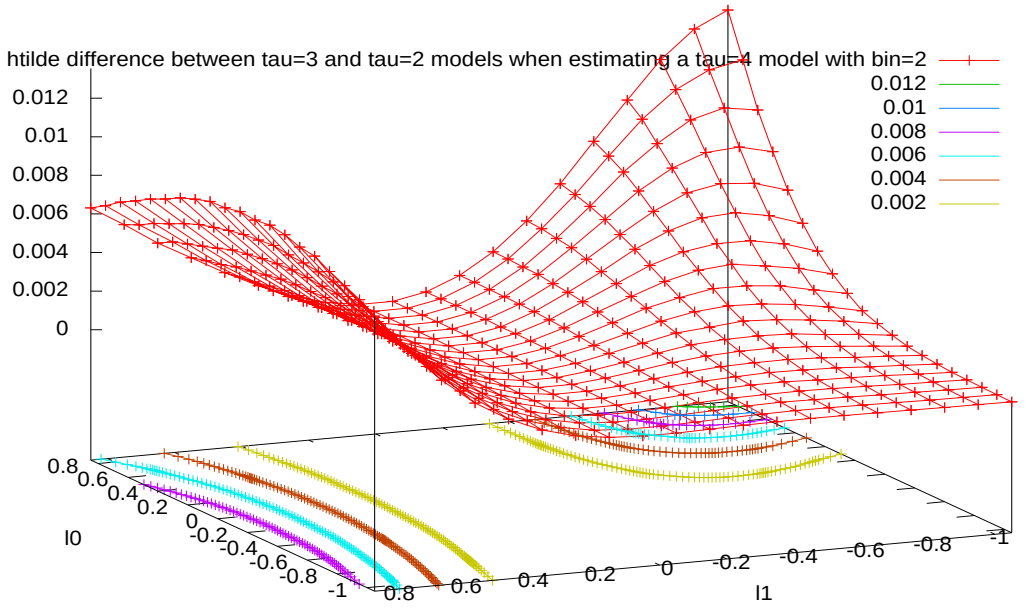

Figure 4.32: Example of the case when there is only one ideal "bin size" and hence a unique best-fitting-"after-binning" model: $\tilde{h}$ difference between two different ψ_τ models, estimated using a binned sequence as fonction of the coefficients of the original generating potential. Since the surface is always above zero. Indeed the second potential has always lower KL-divergence with respect to the binned sequence that othe choice of potential.

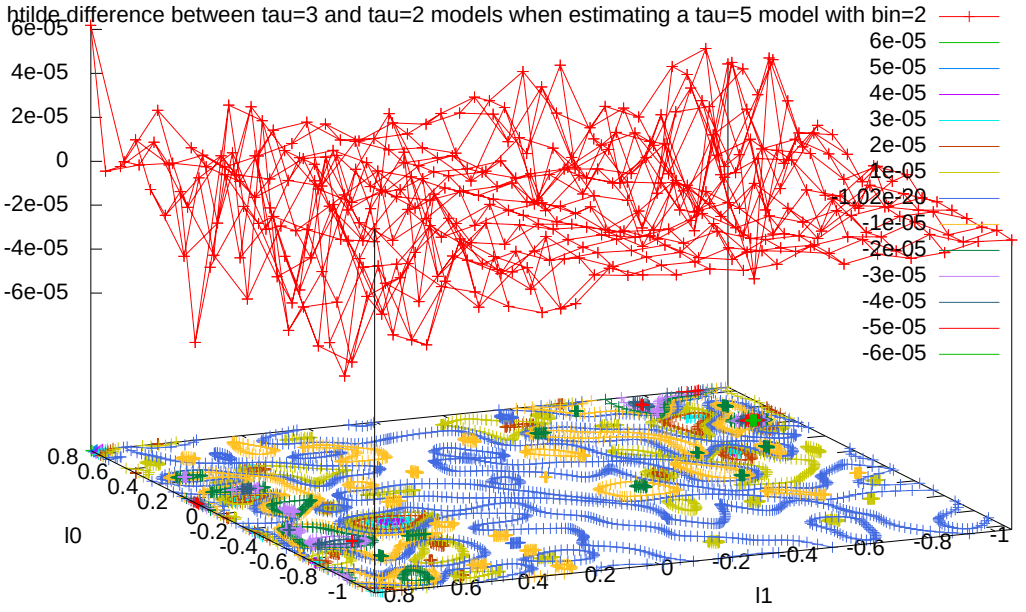

Figure 4.33: Example of the case when there are two ideal "bin size" and hence a unique best-fitting-"after-binning" model: $\tilde{h}$ difference between two different ψ_τ models, estimated using a binned sequence as fonction of the coefficients of the original generating potential. Since the surface is not always above zero, but its values are very small therefore both models performs equally with respect to the binned sequence. Furthermore, these two potentials have always lower KL-divergence with respect to the binned sequence that other choices of potential.

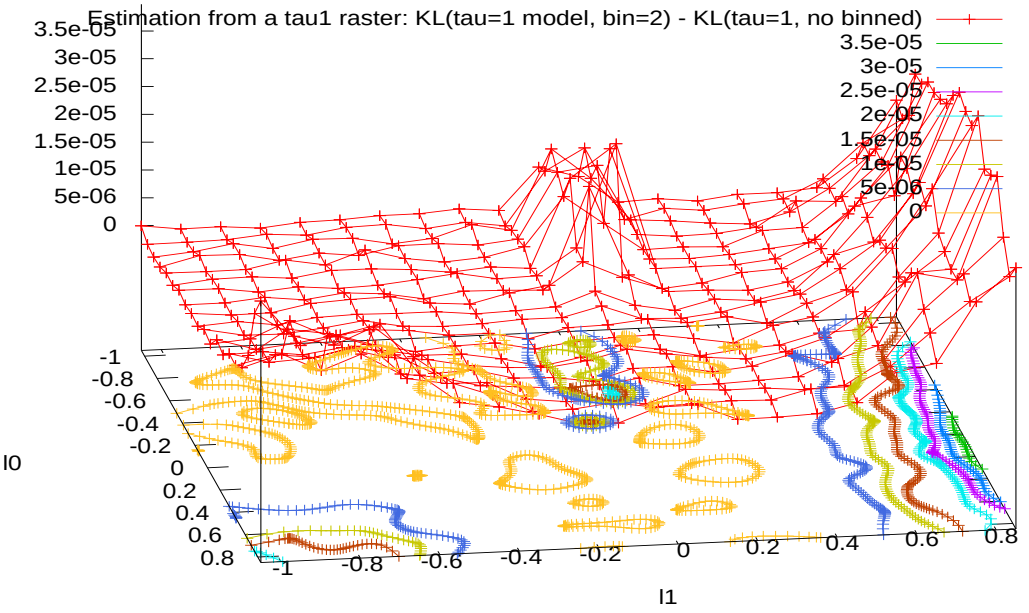

Figure 4.34: best-fitting-"after-binning" vs best fitting-unbinned models: KL-divergence difference between this pair of potentials as a fonction of the coefficients of the original generating potential. Estimation with the unbinned sequence gives always results as good as or even better than estimation over a binned sequence.

4.4 Conclusions

Thanks to the use of synthetic data we have validated our implementation and show its properties as for instance, its well behavior with respect to the over-estimation and also the finite size effects.

Regarding the biological data analysis, our results correspond to a first approach to the different difficulties one can encounter when applying our framework to a variety of biological data. In particular, we have shown that the analysis of biological data is not straightforward since our method present several limitations in the number of cells and the range of the approximating model we can use, in particular with respect of the total amount of data limited to NR=20 for the current EnaS version. We have seen that selecting and adequate binning size is a non trivial problem. For instance, it small bins on sparse data have shown to have small effect in approximating the underlying distribution. On the other hand incorrect binning might produce degenerate results with non-sparse data. Additionally, the identification of the pertinent neural subsets when a large number of neurons is available is not yet clear and information theory tools might help to making the good choice. A future step will be to compare reliable estimations with fixed range models over the most informative neurons, for several stimuli in order to unravel the connection between stimuli encoding and the obtained Gibbs distributions.

In summary the latter applications to the analysis of Biological data are preliminary, but they have greatly helped us to gain insights about the framework and how the implementation behaves facing real data sets and also they have confronted us with several other issues/difficulties arising when treating real data. Thanks to this recently acquired knowledge, we expect be able to extend our data analysis in the short future and provide more interesting results from the biological perspective.

As a final remark, we have shown an examples of binning effects over synthetical data generated from a simple statistical model. In particular memory structure can be collapsed in different ways, and this may cause the multiplicity of the best possible models choices. In this very simple example, we have show that if the memory depth is a integer factor of the bin size, the best fitting minimal potential is unique.

Chapter 5

Gibbs Distributions as Statistical Models for Neural Networks with Synaptic Plasticity

Overview

The aim of this chapter is to show the application of Gibbs Distributions framework to describe the activity of neural networks with synaptic plasticity mechanisms, or say differently, how Gibbs distributions may arise from synaptic adaptation mechanisms. This chapter is organized as follows. In the first part, we discuss some generalities about neural modelling and neural dynamics. We also introduce some basic models of neural activity for which some analytical results about their network dynamics are known. In particular, we can prove that there exist a one-to-one correspondence between spike train and membrane potential trajectories for such networks. Then, for the specific case of a simplified model consisting in a discrete-time Leaky-Intregrate-and-fire neural networks, we recall analytical results recently appeared that proves the existence of a Gibbs distribution describing its spike-train statistics. Using our ENAS implementation, presented in chapter 4, a numerical example of this result is presented.

Contents

5.1 From neuron dynamics to Gibbs Distributions

Essentially we discuss in this section how neural dynamics described as dynamical system can be related to spike statistics described by Gibbs distributions via ergodic theory. We will present more rigourously the thermodynamic formalism material used in the thesis and discuse some details about the validity of the required assumptions. Though all quantities defined below depend on γ, we have dropped this dependence, which is not central here, in order to alleviate notations. We follow [102, 139, 30, 29] in the presentation of Gibbs measures in ergodic theory.

5.1.1 A brief introduction to Modeling aspects

The most basic model of a neural network, considering neurons as logical gates, i.e. with only two possible internal states, was devised by Pitts and McCulloch in 1943. A neuron has a few entries provided by the outputs of other neurons. These inputs are summed up and then determine the state of the neuron by comparison to a certain threshold value. If the signal is larger than the threshold the neuron is active; otherwise it is inactive. After the Hodgkin-Huxley model (for details [84]), a more accurate description of each neuron was possible; specifically the mechanisms generating action potential where described, including as most important improvements the explicit inclusion of time and explicit time evolving value of voltage of membrane Potential and Inputs to the neuron.

Currently, there exists a diversity of neural dynamical models, analyses at different network scales and synaptic models with several degrees of detail. We could build a kind of general landscape of neural modeling. If we imagine the neuron description level as an axe going from binary state to the Hodgkin-Huxley model, there is another axe that we could call "neuron population" where the lowest level is a single isolated neuron, then a few neurons, then including weakly coupled populations and finally a large population and groups of populations. Moreover, considering differents models for the synapse, i.e. the types and detail of chemical interactions, one has a third axe. Obviously, for each system there may exist languages and methods more appropriated than others going from dynamical systems and bifurcation theory to Ergodic theory, probability theory and statistical mechanics, including graph theory and others. Depending on the parameters and structure, neural networks can exhibit complex collective behaviour arising from non-linearity of equations. although those effects can be often described by a few order parameters. In particular, neural networks can exhibit quite complex dynamics, like chaos, requiring a combination of methods from dynamical system theory, ergodic theory and statistical physics, to be studied

and analysed. We provide here an example of this.

5.1.2 Neuron dynamics.

Neural state.

We consider a set of N neurons. Each neuron i is characterized by its state, X_i, which belongs to some compact set $\mathcal{I} \in \mathbb{R}^M$. M is the number of variables characterizing the state of one neuron (we assume that all neurons are described by the same number of variables). A typical example is an integrate and fire model where $M = 1$ and $X_i = V_i$ is the membrane potential of neuron i and $\mathcal{I} = [V_{min}, V_{max}]$ (see examples of section 2.5 in [26]). Other examples are provided by conductances based models of Hodgkin-Huxley type[1] [84]. Then $X_i = (V_i, m_i, n_i, h_i)$ where m_i, n_i are respectively the activation variable for Sodium and Potassium channels and h_i is the inactivation variable for the Sodium channel.

We focus here on cases where the evolution of these N neurons is given by a deterministic discrete-time dynamical system of form:

$$\mathbf{X}(t+1) = \mathbf{F}_{\boldsymbol{\gamma}}\left[\mathbf{X}(t)\right] \tag{5.1}$$

where $\mathbf{X} = \{X_i\}_{i=1}^N$ represents the dynamical state of a network of N neurons at time t, while time is discrete (for a discussion on time discretisation in spiking neural networks see [28]). Thus $\mathbf{X} \in \mathcal{M} = \mathcal{I}^N$ where $\mathcal{M}$ is the phase space of (5.1), and $\mathbf{F}_{\boldsymbol{\gamma}}(\mathcal{M}) \subset \mathcal{M}$. The map $\mathbf{F}_{\boldsymbol{\gamma}} : \mathcal{M} \to \mathcal{M}$ depends on a set of parameters $\boldsymbol{\gamma} \in \mathbb{R}^P$. The typical case considered here is $\boldsymbol{\gamma} = \left(\mathcal{W}, \mathbf{I}^{(ext)}\right)$ where $\mathcal{W}$ is the matrix of synaptic weights and $\mathbf{I}^{(ext)}$ is some external current, assumed to be independent of time in the present chapter (Two explicit examples can be found in section 2.5 of [26]). Thus $\boldsymbol{\gamma}$ is a point in a $P = N^2 + N$ dimensional space of control parameters.

Natural partition.

Neurons are excitable systems. Namely, neuron i "fires" (emits a spike or action potential), whenever its state X_i belongs to some connected region $\mathcal{P}_1$ of its phase space. Otherwise, it is quiescent ($X \in \mathcal{P}_0 = \mathcal{I} \setminus \mathcal{P}_1$). In Integrate and Fire models neuron i fires whenever its membrane potential V_i exceeds some threshold θ. In this case, the corresponding region is $\mathcal{P}_1 = [\theta, V_{max}]$. In Fitzhugh-Nagumo [62, 63, 129] or Hodgkin-Huxley model [84], the firing corresponds to the crossing of a manifold called the threshold separatrix

[1] Note that Hodgkin-Huxley equations are differential equations, while (5.1) corresponds to a discrete time evolution. We assume that we have discretized time with a time scale that can be arbitrary small. A detailed discussion on these aspects for conductance based IF models has been presented in [28]. Some further comments are given below.

[42] which separates the phase space into two connected regions $\mathcal{P}_0$ and $\mathcal{P}_1$. For N identical neurons this leads to a "natural partition" $\mathcal{P}$ of the product phase space $\mathcal{M}$. Call $\Lambda = \{0,1\}^N$, $\boldsymbol{\omega} = (\omega_i)_{i=1}^N \in \Lambda$. Then, $\mathcal{P} = \{\mathcal{P}_{\boldsymbol{\omega}}\}_{\omega\in\Lambda}$, where $\mathcal{P}_{\boldsymbol{\omega}} = \mathcal{P}_{\omega_1} \times \mathcal{P}_{\omega_2} \times \cdots \times \mathcal{P}_{\omega_N}$. Equivalently, if $\mathbf{X} \in \mathcal{P}_{\boldsymbol{\omega}}$, all neurons such that $\omega_i = 1$ are firing while neurons such that $\omega_k = 0$ are quiescent. We call therefore $\boldsymbol{\omega}$ a "spiking pattern".

Raster plots.

To each initial condition $\mathbf{X} \in \mathcal{M}$ we associate a "raster plot" $\omega = \{\boldsymbol{\omega}(t)\}_{t=0}^{+\infty}$ such that $\mathbf{X}(t) \in \mathcal{P}_{\boldsymbol{\omega}(t)}, \forall t \geq 0$. We write $\mathbf{X} \rightharpoonup \omega$. Thus, ω is the sequence of spiking patterns displayed by the neural network when prepared with the initial condition $\mathbf{X}$. On the other way round, we say that an infinite sequence $\omega = \{\boldsymbol{\omega}(t)\}_{t=0}^{+\infty}$ is *an admissible raster plot* if there exists $\mathbf{X} \in \mathcal{M}$ such that $\mathbf{X} \rightharpoonup \omega$. We call $\Sigma_{\boldsymbol{\gamma}}$ the set of admissible raster plots for the set of parameters $\boldsymbol{\gamma}$.

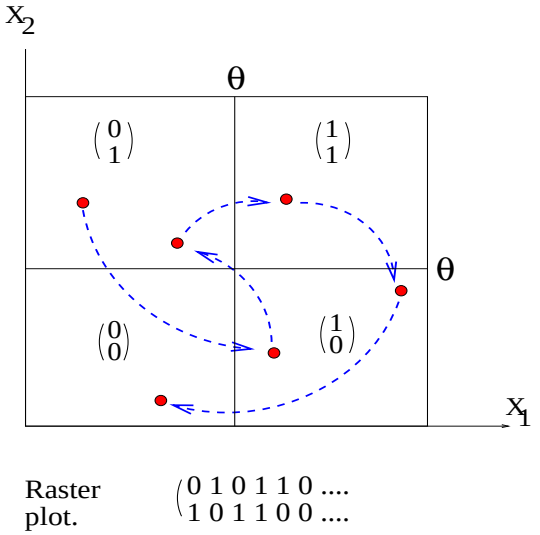

Figure 5.1: The phase space $\mathcal{M}$ (here represented for $N = 2$ neurons with constant threshold) is partitioned in 2^N parts $\mathcal{P}_{\boldsymbol{\omega}} = \prod_{i=1}^N \mathcal{P}_{\omega_i}$, $\boldsymbol{\omega}$ being a spiking pattern. In this way, one associates naturally to an orbit of (5.1) a raster plot.

The dynamics (5.1) induce a dynamics on the set of raster plot in the following way:

$$\begin{array}{ccc} \mathbf{X} & \xrightarrow{\mathbf{F}_{\boldsymbol{\gamma}}} & \mathbf{F}_{\boldsymbol{\gamma}}(\mathbf{X}) \\ \downarrow & & \downarrow \\ \omega & \xrightarrow{\sigma_{\boldsymbol{\gamma}}} & \sigma_{\boldsymbol{\gamma}}(\omega) \end{array}$$

In this construction $\sigma_{\boldsymbol{\gamma}}$, called the "left shift", shifts the raster plot leftwise at each time step of the dynamics. Thus, in some sense, raster plots

provide a code for the orbits of (5.1). But, the correspondence may not be one-to-one. That is why we use the notation $\rightharpoonup$ instead of $\rightarrow$.

5.1.3 Ergodic measures.

Fix ϕ a continuous function. The time average of ϕ along the orbit of $\mathbf{X}$ is given by

$$\pi_{\mathbf{X}}(\phi) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \phi(\mathbf{X}(t)).$$

If μ is an invariant measure[2], then, according to Birkhoff theorem, this limit exists for μ-almost every $\mathbf{X}$. Standard theorems in ergodic theory ensure that (5.1) has at least one invariant measure [100] but there are typically many. Among them, ergodic measures play a distinguished role. An invariant measure μ is ergodic if any real measurable invariant function is μ-almost surely constant. As a corollary of Birkhoff's theorem $\pi_{\mathbf{X}}(\phi) = \mu(\phi)$, for μ-almost every $\mathbf{X}$, where $\mu(\phi) = \int \phi d\mu$ is the expectation of ϕ with respect to μ. Namely, the empirical average $\pi_{\mathbf{X}}(\phi)$ is equal to the "ensemble" average $\mu(\phi)$ whatever the initial condition, provided it is chosen in the support of μ. Any invariant measure can be written as a convex decomposition of ergodic measures.

5.1.4 Raster plots statistics.

Fix $n > 0$, a set of times $t_1, \dots t_n$, and a set of prescribed spiking patterns $\boldsymbol{\omega}(t_1) \dots \boldsymbol{\omega}(t_n)$. The set of raster plots $\omega' \in \Sigma_{\boldsymbol{\gamma}}$ such that $\boldsymbol{\omega}'(t_k) = \boldsymbol{\omega}(t_k), k = 1 \dots n$ contains therefore raster plots where the spiking patterns at prescribed times $t_1 \dots t_n$ are imposed (cylinder set). By (countable) intersection and union of cylinder sets one can generate all possible *events* in $\Sigma_{\boldsymbol{\gamma}}$, such as "neuron i is firing at time t_1", or "neuron i_1 is firing at time t_1, and neuron i_2 is firing at time t_2, ... and neuron i_n is firing at time t_n", etc ....

The statistical properties of raster plots are inherited from the statistical properties of orbits of (5.1), via the correspondence $\nu[C] = \mu[\{\mathbf{X} \rightharpoonup \omega, \omega \in C\}]$, where C is a cylinder set. Thus, ν is an ergodic probability measure on the set of raster plots $\Sigma_{\boldsymbol{\gamma}}$. $\nu(C)$ can be estimated by the time-empirical average:

$$\pi^{(T)}(C) = \frac{1}{T} \sum_{t=1}^{T} \chi(\sigma_{\boldsymbol{\gamma}}^t \omega \in C) \tag{5.2}$$

where χ is the indicatrix function. Namely, call:

$$\pi_\omega(.) = \lim_{T\to\infty} \pi^{(T)}(.), \tag{5.3}$$

[2]Namely $\mu({\mathbf{F}_\gamma}^{-1} A) = \mathbf{F}_\gamma(A)$ for any measurable set $A \subset \mathcal{M}$.

the *empirical measure* for the raster plot ω, then:

$$\nu = \pi_\omega \tag{5.4}$$

for ν-almost every ω. Equivalently, the time average of some function $\phi : \Sigma_\gamma \to \mathbb{R}$ is given by:

$$\pi_{\tilde{\omega}}^{(T)}(\phi) = \frac{1}{T}\sum_{t=1}^{T} \phi(\sigma^t \tilde{\omega}) \tag{5.5}$$

and, as $T \to \infty$:

$$\pi_{\tilde{\omega}}(\phi) = \lim_{T\to\infty} \pi_{\tilde{\omega}}^{(T)}(\phi) = \nu(\phi), \tag{5.6}$$

for ν-almost every ω. Thus, ν characterizes the averages obtained from experimental raster plots.

5.1.5 Thermodynamic formalism.

Potentials.

This definition of potential extends beyond the Markov case. Fix $0 < \Theta < 1$. We define a metric on Σ by $d_\Theta(\omega, \omega') = \Theta^p$, where p is the largest integer such that $\boldsymbol{\omega}(t) = \boldsymbol{\omega}(t'), 1 \le t \le p-1$. (The related topology thus structures the fact that raster plots coincide until $t = p-1$).

For a continuous function $\psi : \Sigma \to \mathbb{R}$ and $n \ge 1$ define

$$var_n\psi = \sup\left\{|\psi(\omega) - \psi(\omega')| : \boldsymbol{\omega}(t) = \boldsymbol{\omega}'(t), 1 \le i \le n\right\}.$$

We denote by $C(\Sigma)$ the space of continuous real functions ψ such that $var_n\psi \to 0$ as $n \to \infty$.

$\psi \in C(\Sigma)$ is called a *potential* if $var_n(\psi) \le C\,\Theta^n, n = 1, 2\ldots$, where C is some positive constant[3]. Equivalently, ψ is Holder. For a positive integer R a *range-R* potential is a potential such that $\psi(\omega) = \psi(\omega')$, if $\boldsymbol{\omega}(t) = \boldsymbol{\omega}'(t), 0 \le t \le R$. That is, ψ depends only on the r first spiking patterns $\boldsymbol{\omega}(1), \ldots, \boldsymbol{\omega}(R)$. Examples are $\omega_{i_1}(0)$ (range 1 potential), $\omega_{i_1}(0)\omega_{i_2}(k)$ (range k potential), $\omega_{i_1}(0)\omega_{i_2}(k)\omega_{i_3}(l)$ (range $\max(k,l)$ potential), etc, ... We remark that we can integrate the grammar into a potential ψ such that $\psi(\omega) = 0$ is ω is allowed, and $\psi(\omega) = -\infty$ otherwise.

[3]This condition is analogous to the potential decay ensuring the existence of a thermodynamic limit in statistical mechanics [161, 124]. In the present context it ensures that Gibbs measures coincides with equilibrium states.

Gibbs measure.

According to [17] which is the more general definition that we now, although equivalent definitions exist (see e.g. [164]).

Given a regular potential ψ and n a positive integer, set $S^{(n)}\boldsymbol{\psi}(\omega) = \sum_{t=0}^{n-1} \psi(\sigma^t\omega)$. Assume that the transition matrix of the shift σ is primitive. Then, there is a unique ergodic measure $\mu_{\boldsymbol{\psi}}$ for which one can find some constants $P(\boldsymbol{\psi}), c_1, c_2$ with $0 < c_1 \leq 1 \leq c_2$ such that for all $n \geq 1$ and for all $\omega \in \Sigma$:

$$c_1 \leq \frac{\mu_{\boldsymbol{\psi}}\left([\omega]_0^{n-1}\right)}{exp(-nP(\boldsymbol{\psi}) + S^{(n)}\boldsymbol{\psi}(\omega))} \leq c_2. \tag{5.7}$$

Basically, the condition (5.7) expresses that the measure of the cylinder $[\omega]_0^{n-1}$ behaves like

$$\mu_{\boldsymbol{\psi}}\left([\omega]_0^{n-1}\right) \sim \frac{\exp S^{(n)}\boldsymbol{\psi}(\omega)}{Z^{(n)}(\psi)}, \tag{5.8}$$

where the the normalisation factor $Z^{(n)}(\psi)$ is defined below in eq. 5.9. We retrieve therefore the classical form of Gibbs distribution where spin chains are replaced by sequences of spiking patterns.

Topological pressure. Set

$$Z^{(n)}(\psi) = \sum_{\omega \in \Sigma^{(n)}} \exp S^{(n)}\boldsymbol{\psi}(\omega), \tag{5.9}$$

where $\Sigma^{(n)}$ is the set of admissible cylinders of length n, corresponds to the partition function in statistical physics. As a further link to statistical physics one has:

$$P(\boldsymbol{\psi}) = \lim_{n\to\infty} \frac{1}{n} \log\left[Z^{(n)}(\psi)\right] \tag{5.10}$$

which is analogous to a free energy.

Ruelle-Perron-Frobenius operator.

The Ruelle-Perron-Frobenius (RPF) operator for the potential ψ, denoted by L_ψ, acts on functions $g \in C(\Sigma)$ as $L_\psi g(\omega) = \sum_{\omega'=\sigma\omega} e^{\psi(\omega')} g(\omega')$. This operator has a unique maximal eigenvalue $s_\psi = e^{P[\psi]}$ associated to a right eigenfunction b_ψ and a left eigenfunction ρ_ψ (probability measure) such that $L_\psi b_\psi(\omega) = s b_\psi(\omega)$ and $\int L_\psi v d\rho_\psi = s \int v d\rho_\psi$, for all $v \in C(\Sigma)$. The remaining part of the spectrum is located in a disk in the complex plane, of radius strictly lower than s. Finally, for all $v \in C(\Sigma)$

$$\frac{1}{s^n} L_\psi^n v \to b_\psi \int v d\rho_\psi. \tag{5.11}$$

The Gibbs measure is $\nu_\psi = b_\psi \rho_\psi$.

Given ψ and knowing its right eigenfunction b_ψ and pressure $P[\psi]$ one can construct a new potential:

$$\Psi(\omega) = \psi(\omega) - \log(b_\psi(\sigma\omega)) + \log(b_\psi(\omega)) - P[\psi], \tag{5.12}$$

such that $L_\Psi \mathbf{1} = 1$, where $\mathbf{1}$ is the constant function $\mathbf{1}(\omega) = 1$. Such a potential is called "normalised". Its pressure is zero.

If ψ has a finite range R then the RPF operator reduces to a matrix, b_ψ has range $R-1$ and is the right eigenvector of this matrix with largest eigenvalue $e^{P[\psi]}$. Thus, b_ψ and $P[\psi]$ can be easily determined. The matrix corresponding to the normalised potential Ψ is the stochastic matrix of a Markov chain. Then, the convergence property (5.11), which reduces to classical properties of Markov chains [171, 66], can be used to generate a raster plot distributed according to a Gibbs distribution with a potential ψ. Moreover, for ν_ψ-almost-every raster plot ω:

$$\lim_{T\to+\infty} \frac{\pi^{(T)}[j\boldsymbol{\omega}(1)\dots\boldsymbol{\omega}(r-1)]}{\pi^{(T)}[\boldsymbol{\omega}(1)\dots\boldsymbol{\omega}(r)]} = e^{\Psi(j\boldsymbol{\omega}(1)\dots\boldsymbol{\omega}(r-1))} \tag{5.13}$$

This allows to obtain of Ψ numerically.

Properties of the topological pressure.

The topological pressure is a convex functional of the potentials i.e.

$$P[\alpha\psi_1 + (1-\alpha)\psi_2] \leq \alpha P[\psi_1] + (1-\alpha)P[\psi_2]; \quad \alpha \in [0,1]. \tag{5.14}$$

Like the free energy, the topological pressure is a generating functional. For example,

$$\left.\frac{\partial P[\psi + \alpha\phi]}{\partial\alpha}\right|_{\alpha=0} = \mu_{\boldsymbol{\psi}}[\phi], \tag{5.15}$$

where $\mu_{\boldsymbol{\psi}}[\phi]$ is the average value of ϕ with respect to $\mu_{\boldsymbol{\psi}}$, and:

$$\left.\frac{\partial^2 P[\psi + \alpha_1\phi_1 + \alpha_2\phi_2]}{\partial\alpha_1\partial\alpha_2}\right|_{\alpha_1=\alpha_2=0} = \sum_{t=-\infty}^{+\infty} C_{\phi_1\phi_2}(t), \tag{5.16}$$

where:

$$C_{\phi_1\phi_2}(t) = \left[\mu_{\boldsymbol{\psi}}\left(\phi_1 \circ \sigma^t \phi_2\right) - \mu_{\boldsymbol{\psi}}(\phi_1)\mu_{\boldsymbol{\psi}}(\phi_2)\right], \tag{5.17}$$

is the correlation function of ϕ_1, ϕ_2 at time t. Note that the equality (5.16) is known in statistical physics under the name of fluctuation-dissipation theorem.

Gibbs measure and variational principle.

The Gibbs measure obeys *a variational principle*. Let ν be a σ-invariant measure. Call:

$$h^{(n)}[\nu] = -\sum_{\omega\in\Sigma^{(n)}} \nu\left([\omega]_0^{n-1}\right)\log\nu\left([\omega]_0^{n-1}\right). \tag{5.18}$$

Then,

$$h[\nu] = \lim_{n\to\infty}\frac{h^{(n)}[\nu]}{n} \tag{5.19}$$

is the entropy of ν. Let $\mathcal{M}^{inv}$ be the set of invariant measures for σ, then:

$$P(\boldsymbol{\psi}) = \sup_{\nu\in\mathcal{M}^{inv}}\left(h[\nu]+\nu[\psi]\right) = h[\mu_{\boldsymbol{\psi}}]+\mu_{\boldsymbol{\psi}}[\psi]. \tag{5.20}$$

$\mu_{\boldsymbol{\psi}}$ is called an equilibrium state, as it maximises some version of the (neg) free energy.

5.2 Spike train statistics in a simulated Neural Network

5.2.1 A basic Neural Network Model: the BMS Model.

The Leaky-Integrate-and-Fire (LIF) model is probably the best-known example of a formal spiking neuron model. Additionally, 'IF models are good enough" to approximate biological neurons spike trains [96]. See [156] for a recent illustration of this. Here we use discrete-time LIF model introduced by Soula et al. (BMS) in [185] which was mathematically analysed in [24, 28].

From a LIF model to a discrete-time spiking neuron model.

In this section we present the correct discretization of a LIF model, which has been introduced in [185] and called BMS. The discretization process is carried out thanks to the well-done analogy presented in [70] between a RC circuit and a biological neuron.

From this work, it is proposed the neuron membrane potential $V_i(t)$, during the quiescient state, to be described by a differential equation:

$$\frac{dV_i(t)}{dt} = -\frac{V_i(t)}{\tau}+\frac{I_i(t)}{C}$$

where $\tau = RC$ is the time required for the voltage to fall to V_0 and is called the time constant.

Then, we consider that the spiking neuron model will be used in numerical simulations, thus the Euler method is applied in order to solve the

differential equation. Then we fix some constant values, such as the sampling time scale $dt = 1$ and the capacitance $C = 1$. Finally, we consider that $\gamma = 1 - \frac{1}{\tau}$, where $\tau \geq 1$, thus $\gamma \in [0, 1[$, the equation reads:

$$\frac{V_i(t+dt) - V_i(t)}{dt} = -\frac{V_i(t)}{\tau} + \frac{I_i(t)}{C}$$

and setting $dt = 1$,

$$V_i[t+1] = V_i[k] - \frac{V_i[k]}{\tau} + \frac{I_i(t)}{C}$$

$$V_i[t+1] = V_i[k]\Big(1 - \frac{1}{\tau}\Big) + \frac{I_i[k]}{C}$$

$$V_i[t+1] = \gamma V_i[k] + \frac{I_i[k]}{C}$$

$$V_i[k+1] = \gamma V_i[k] + I_i[k]$$

Additionally, we may include in this equation the after-firing reset mechanism. Let us denote by $Z(V[k])$ the indicator function $\chi_{[\theta,+\infty[}(V[k])$, where θ is the membrane spiking threshold and $\chi_A(x) = 1$ if $x \in A$ and 0 otherwise. More precisely $Z(V)$ defines the firing state of the neuron. Then we introduce a reset term $(1 - Z(V))$ to act once that it neuron has reached the fixed firing threshold θ. Thus, the discrete-time neuron model reads [24]:

$$V_i[k] = \gamma V_i[k-1](1 - Z(V_i[k-1])) + I_i[k] \tag{5.21}$$

The network model

The model is defined as follows. Denote by W_{ij} the synaptic weight of neuron j over neuron i, I_i^{ext} an external input on neuron i. Each neuron is submitted to noise, modeled by an additional input, $\sigma_B B_i(t)$, with $\sigma_B > 0$ and where the $B_i(t)$'s are Gaussian, independent, centered random variable with variance 1. The network dynamics is given by:

$$V_i(t+1) = \gamma\, V_i\,(1 - Z[V_i(t)]) + \sum_{j=1}^{N} W_{ij} Z[V_j(t)] + I_i^{ext} + \sigma_B B_i(t); \qquad i = 1 \ldots N, \tag{5.22}$$

where $\gamma \in [0, 1[$ is the leak in this discrete time model ($\gamma = 1 - \frac{dt}{\tau}$). The equation (5.22) implements both the integrate and firing regime. It turns out that this time-discretization of the standard integrate-and-Fire neuron model, which as discussed in e.g. [89], provides a rough but realistic approximation of biological neurons behaviors. Its dynamics has been fully characterized for $\sigma_B = 0$ in [24] while the dynamics with noise is investigated in [25]. Its links to more elaborated models closer to biology is discussed in [28].

5.2.2 Exact spike trains statistics.

The spike train statistics of model (5.22) is known explicitly [25]. We describe this results here. For $\sigma_B > 0$ in eq (5.22), it has been proved that there is a unique Gibbs distribution in this model, in the sense given in section 5.1, whose potential is given by:

$$\psi([\omega]_{-\infty}^{0}) = \sum_{i=1}^{N} \left[\omega_i(0) \log \left(\pi \left(\frac{\theta - C_i(\underline{\omega})}{\sigma_i(\underline{\omega})} \right) \right) + (1 - \omega_i(0)) \log \left(1 - \pi \left(\frac{\theta - C_i(\underline{\omega})}{\sigma_i(\underline{\omega})} \right) \right) \right], \tag{5.23}$$

where $\pi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-\frac{u^2}{2}} du$, $\underline{\omega} = [\omega]_{-\infty}^{-1}$, $C_i(\underline{\omega}) = \sum_{j=1}^{N} W_{ij} x_{ij}(\underline{\omega}) + I_i^{ext} \frac{1-\gamma^{t+1-\tau_i(\underline{\omega})}}{1-\gamma}$, $x_{ij}(\underline{\omega}) = \sum_{l=\tau_i(\underline{\omega})}^{t} \gamma^{t-l} \omega_j(l)$, $\sigma_i^2(\underline{\omega}) = \sigma_B^2 \frac{1-\gamma^{2(t+1-\tau_i(\underline{\omega}))}}{1-\gamma^2}$. Finally, $\tau_i(\underline{\omega})$ is the last time, before $t = -1$, where neuron i has fired, in the sequence $\underline{\omega}$ (with the convention that $\tau_i(\underline{\omega}) = -\infty$ for the sequences such that $\omega_i(n) = 0, \forall n < 0$).

This potential has infinite range, however range $R \geq 1$ approximations exist, that consist in replacing $\underline{\omega} = [\omega]_{-\infty}^{-1}$ by $[\omega]_{-R}^{-1}$ in (5.23). The KL divergence between the Gibbs measure of the approximated potential and the exact measure decays like γ^R. Recall that finite range potentials admit a polynomial expansion of form (2.17).

This is the only example that we know of explicit Gibbs distributions and explicit justification of the polynomial expansion. In particular. the Ising model is the worst after the Bernoulli, since it does not consider the memory terms arising in eq.5.23 [8]

5.2.3 Numerical estimation of spike train statistics

Here we have considered only one example of model (5.22) (more extended simulations and results will be provided elsewhere). It consists of 4 neurons, with a *sparse* connectivity matrix so that there are neurons without synaptic interactions. The synaptic weights matrix is:

$$\mathcal{W} = \begin{pmatrix} 0 & -0.568 & 1.77 & 0 \\ 1.6 & 0 & -0.174 & 0 \\ 0 & 0.332 & 0 & -0.351 \\ 0 & 1.41 & -0.0602 & 0 \end{pmatrix},$$

while $\gamma = 0.1, \sigma_B = 0.25, I_i^{ext} = 0.5$.

First, one can compute directly the theoretical entropy of the model using the results exposed in the previous section: the entropy of the range-R approximation, that can be computed with our formalism, converges exponentially fast with R to the entropy of the infinite range potential. For these parameters, the asymptotic value is $h = 0.57$.

Then, we generate a raster of length $T = 10^7$ for the 4 neurons and we compute the KL divergence between the empirical measure and several potentials including:

- (i) The range-R approximation of (5.23), denoted $\psi^{(R)}$. Note that $\psi^{(R)}$ does not contain all monomials. In particular, *it does not have the Ising term (the corresponding coefficient is zero)*.
- (ii) A Bernoulli model ψ^{Ber};
- (iii) An Ising model ψ^{Is};
- (iv) A one-time step Ising Markov model (as proposed in [120]) ψ^{MEDF} [4] ;
- (v) A range-R model containing all monomials ψ^{all}.

Here we can compute the KL divergence since we known the theoretical entropy. The results are presented in the table (5.1). Note that the estimated KL divergence of range-1 potentials slightly depend on R since the $\mathcal{L}(\psi)$ matrix, and thus the pressure, depend on R.

Table 5.1: Kullback-Leibler divergence between the empirical measure of a raster generated by (5.22) (See text for the parameters value) and the Gibbs distribution, for several statistical models.

	$\psi^{(R)}$	ψ^{Ber}	ψ^{Is}	ψ^{MEDF}	ψ^{all}
R=1	0.379	0.379	0.312	1.211	0.309
R=2	0.00883	0.299871	0.256671	0.257068	0.0075
R=3	-0.001	0.250736	0.215422	0.200534	0.0001

We observe that our procedure recovers the fact that the range-R potential $\psi^{(R)}$ is the best to approximate the empirical measure, in the sense that it minimizes the KL divergence and that it has the minimal number of terms (ψ^{all} does as good as $\psi^{(R)}$ for the KL divergence but it contains more monomials whose coefficient (almost) vanish in the estimation).

5.3 Synaptic Plasticity in neural networks

5.3.1 General overview

The notion of neural code and information cannot be separated from the capacity of neuronal networks to evolve and adapt by *plasticity* mechanisms,

[4] or equivalently, a **RPTD-**1 from (4.1)

and especially *synaptic plasticity*. The latter occurs at many levels of organization and time scales in the nervous system [13]. It is of course involved in memory and learning mechanisms, but it also alters excitability of brain area and regulates behavioural states (e.g. transition between sleep and wakeful activity). Therefore, understanding the effects of synaptic plasticity on neurons dynamics is a crucial challenge. On experimental grounds, different synaptic plasticity mechanisms have been exhibited from the Hebbian's ones [80] to Long Term Potentiation (LTP) and Long Term Depression (LTD), and more recently to Spike Time Dependent Plasticity (STDP) [119, 12] (see [48, 71, 40] for a review). Modeling these mechanisms requires both a bottom-up and top-down approach.

This issue is tackled, on theoretical grounds, by inferring "synaptic updates rules" or "learning rules" from biological observations [207, 13, 127] and extrapolating, by theoretical or numerical investigations, the effects of such synaptic rule on such neural network *model*. This bottom-up approach relies on the belief that there are "canonical neural models" and "canonical plasticity rules" capturing the most essential features of biology. Unfortunately, this results in a plethora of canonical "candidates" and a huge number of papers and controversies. In an attempt to clarify and unify the overall vision, some researchers have proposed to associate learning rules and their dynamical effects to general principles, and especially to"variational" or "optimality" principles, where some functional has to be maximised or minimised [49, 150, 151, 15, 34, 200, 201]. Therefore, in these "top-down" approaches, plasticity rules "emerge" from first principles. Unfortunately, in most examples, the validations of these theories has been restricted to considering isolated neurons submitted to input spike trains with ad hoc statistics (typically, Poisson distributed with independent spikes [200, 201]).

However, addressing the effect of synaptic plasticity in neuronal networks where dynamics is *emerging* from collective effects and where spikes statistics are *constrained* by this dynamics seems to be of central importance. This is the point of view raised in the present chapter the thesis, where, we focus the BMS model, which is a simplification of real neurons. Even in this case, this issue is subject to two main difficulties. On one hand, one must identify the generic collective dynamical regimes displayed by the model for different choices of parameters (including synaptic weights). On the other hand, one must analyse the effects of varying synaptic weights when applying plasticity rules. This requires to handle a complex interwoven evolution where neurons dynamics depends on synapses and synapses evolution depends on neuron dynamics. The first aspect has been addressed by several authors using mean-field approaches (see e.g. [163] and references therein), "Markovian approaches" [186], or dynamical system theory

(see [27] and references therein). The second aspect has, up to our knowledge, been investigated theoretically in only a few examples with Hebbian learning [46, 179, 180] or discrete time Integrate and Fire models with an STDP like rule [184, 185] and is further addressed here.

5.3.2 Mathematical Formulation of Synaptic plasticity.

Synaptic plasticity corresponds to the evolution of synaptic efficacy (synaptic weights). More precisely, in our notations, W_{ij} essentially provides the maximal amplitude of the post-synaptic potential induced, at the synapse linking j to i, when neuron j fires a spike. At the present moment no simple model can account by itself the full diversity of experimental results. However, it exists a separation naturally imposed whether the model accounts or not directly non-linear contributions of spikes. Following [72], in order to get a generic STDP model the first step is to assume that

$$\frac{d}{dt}W_{ij}(t) = g\left(W_{ij}(t); V_i^{post}(t - T_s < t), V_j^{pre}(t - T_s < t)\right) \tag{5.24}$$

where $V(\bar{t} < t)$ denotes the membrane potential history from a time $t - T$ to t, and indexes $post$ et pre recall that neuron i is the post-synaptic neuron and the neuron j the presynaptic neuron. The next step will consist in propose a Volterra series expansion of this functional. In the following, we will assume that we can neglect the information given by the membrane potential evolution and use instead the 'spike-time history' or more formally the raster code on the past time T

Synapses update as an integration over spikes trains.

In this approximation, synaptic weights evolve in time according to the spikes emitted by the pre- and post- synaptic neuron. In other words, the variation of W_{ij} , (as introduced in 5.22,at time t is a function of the spiking sequences of neurons i and j from time $t - T$ to time t, where T is time scale characterizing the width of the spike trains influencing the synaptic change. Here we consider a discrete-time synaptic change of type eq. 5.24, thus the synapse update writes:

$$\delta W_{ij}(t+1) = g\left(W_{ij}(t), [\omega_i]_{t-T}^{t}, [\omega_j]_{t-T}^{t}\right),\ t > T, \tag{5.25}$$

with $[\omega_i]_{t-T}^{t} = [\omega_i(t - T) \ldots \omega_i(t)]$, thus denoting a single unit of the whole block $[\omega]_{t-T}^{t}$. Thus, typically, synaptic adaptation results from an integration of spikes over the time scale T.

Spikes responses of synapses.

The synaptic variation δW_{ij} is the integrated response of the synapse from neuron j to neuron i when neuron j sends a spike sequence $[\omega_j]_{t-T}^{t}$ and neuron i fires according to $[\omega_i]_{t-T}^{t}$. This response is not a deterministic function, while g is. Thus (5.25) is an approximation. As a matter of fact, the explicit form of g is usually derived from phenomenological considerations as well as experimental results where synaptic changes can be induced by *specific* simulations conditions, defined through the firing frequency of pre- and post-synaptic neurons [14, 55], the membrane potential of the post-synaptic neuron [6], or spike timing [112, 119, 12] (see [117] for a review). Thus, these results are usually based on a repetition of experiments involving the excitation of pre- and post-synaptic neurons by specific spike trains. The phenomenological plasticity rules derived from these experiments are therefore of *statistical* nature. Namely, they do not tell us what will be the exact changes induced on synapses when this or this spike train is applied to pre- and post-synaptic neuron. Instead, they provide us the average synaptic change. Thus, the function $g(W_{ij}, [\omega_i]_{t-T}^{t}, [\omega_j]_{t-T}^{t})$ in (5.25) is typically a statistical average of the synaptic response when the spike train of neuron j (resp. i) is $[\omega_j]_{t-T}^{t}$ (resp. $[\omega_i]_{t-T}^{t}$), and the actual synaptic weight value is W_{ij}.

In this thesis we have investigated the effects of synaptic plasticity rules where the characteristic time scale T_s is quite a bit larger than the time scale of evolution of the neurons. Namely, we consider *slow* adaptation rules. In this situation the synaptic weights update can be written as a function of the empirical average (5.2) over the spike train. We therefore consider adaptation rules of form:

$$g\left(W_{ij},\ \right) = \epsilon \pi^{(T)}\left[\phi_{ij}(W_{ij}, [\omega]_{t-T_s}^{t})\right]. \tag{5.26}$$

where ϵ is a parameter that will be typically small, $\pi^{(T)}$ is the empirical average as defined in (5.2) over a time window T, T_s is the interspike interval relevant for the synaptic plasticity[5] (seethe examples below), and finally $\phi_{ij}(W_{ij}, [\omega]_{t-T_s}^{t})$ is a function that we explicit in the next paragraph, corresponding to the monomial expansion of the interspike plasticity interaction.

Following [71] canonical form of adaptation rules use an expansion in singlet, pairs, triplets etc of spikes. Formally, one may write a generic form for these rules. Let us denote $L \equiv \mathcal{P}(1, T_s)$, L i.e. be the finite set of finite ordered

[5]This value depends on the internal synaptic mechanisms like the chemical interaction time-scale

lists of integers $\in \{-T_s, -T_s+1, \ldots, 0\}$ with the form $l(n) = (t_1, t_2, \ldots, t_n)$ with $-T_s \leq t_1 < t_2 < \cdots < t_n \leq 0$ and $0 < n < T_s$. Let us consider all monomials (see chapter 2) $m_{i,l(n)}$ containing only activities of neuron i and determined by $l(n)$, i.e., a function $m_{i,l(n)}([\omega]_{-T_s}^{0}) = \omega_i(t_1)\ldots\omega_i(t_n)$. Additionally, we fix $T_s < \infty$, the interspike plasticity interaction as a positive integer. Then, we can write the functional $\phi_{ij}(W_{ij}, [\omega]_{t-Ts}^{t})$ as a linear combination of products of $m_{i,l(n)}m_{j,l'(n)}$:

$$\phi_{ij}(W_{ij}, [\omega]_{t-Ts}^{t}) = \sum_{l_1,l_2 \in L} h_{ijl_1l_2}(W_{ij})\, m_{i,l_1}([\omega_i]_{t-T}^{t})\, m_{j,l_2}([\omega_j]_{t-T}^{t}), \qquad (5.27)$$

where $h_{ijl_1l_2}(W_{ij})$ are the parametric coefficients ($\boldsymbol{\lambda}$) as introduced in chapter 2, and corresponds to smooth functions of W_{ij}. They can be constant, linear functions of W_{ij} ("multiplicative" rules) or nonlinear functions allowing for example to constrain W_{ij} within bounded values (e.g. hard bound or soft bounds rules). The form (5.27) is the most general form of synaptic adaptation rules considered in this manuscript

Examples of adaptation rule instantiation.

Hebbian learning uses firing rates. A typical example corresponds to $g_{ij}(W_{ij}, [\omega_i]_{t-T}^{t}, [\omega_j]_{t-T}^{t}) = \epsilon\frac{1}{T}\sum_{s_1,s_2=t-T}^{t}(\omega_i(s_1) - r_i(s_1))(\omega_j(s_2) - r_j(s_2)$ (correlation rule) where $r_i(t) = \frac{1}{T}\sum_{s=t-T}^{t}\omega_i(s)$ is the frequency rate of neuron i in the raster plot ω, computed in the time windows $[t-T, t]$. For this case we do not explicit the form of ϕ_{ij} but only remark that it is independent of W_{ij}.
Spike-Time Dependent Plasticity as derived from Bi and Poo [12] provides the average amount of synaptic variation given the delay between the pre- and post-synaptic spike. Thus, "classical" STDP writes [71, 91]:

$$g\left(W_{ij}, [\omega_i]_{t-T}^{t}, [\omega_j]_{t-T}^{t}\right) = \frac{\epsilon}{T}\sum_{s_1,s_2=t-T}^{t} f(s_1 - s_2)\,\omega_i(s_1)\,\omega_j(s_2) \qquad (5.28)$$

with:

$$f(x) = \begin{cases} A_- e^{\frac{x}{\tau_-}},\ x < 0; \\ A_+ e^{-\frac{x}{\tau_+}},\ x > 0; \\ 0,\ x = 0; \end{cases} \qquad (5.29)$$

where the shape of f has been obtained from statistical extrapolations of experimental data. Hence STDP is based on a second order statistics (spikes correlations). There is, in this case, an evident time scale for the interspike interaction $T_s = \max(\tau_-, \tau_+)$, beyond which f is essentially zero. In this rule ϕ_{ij} is also independent of W_{ij}, and more precisely it is given by

$$\phi_{ij} = \omega_j(0)\sum_{u=-T_s}^{T_s} f(u)\,\omega_i(u),$$

.

"Nearest neighbors" STDP (according to Izhikevich terminology [91]) writes:

$$g\left(W_{ij}, [\omega_i]_{t-T}^{t}, [\omega_j]_{t-T}^{t}\right) = \frac{\epsilon}{T} \sum_{s=t-T}^{t} f(\tau(s) - s)\, \omega_i(\tau(s))\, \omega_j(s), \tag{5.30}$$

with $\tau(s) = \min_{t, \omega_j(t)=1} |t - s|$.

Generalized STDP As a last example, [71] propose a rule which corresponds to :

$$g(W_{ij}, [\omega_i]_{t-T}^{t}, [\omega_j]_{t-T}^{t}) = \epsilon \left[a_1^{pre} \sum_{s=t-T}^{t} \omega_j(s) + a_1^{post} \sum_{s=t-T}^{t} \omega_i(s) + \sum_{s_1,s_2=t-T}^{t} f(s_1 - s_2)\omega_i(s_1)\omega_j(s_2) \right]. \tag{5.31}$$

We capture thus the main standard synaptic adaptation rules with (5.27).

Coupled dynamics.

We consider now the following coupled dynamics. Neurons are evolving according to (5.1). We focus here on *slow* synapses dynamics. Namely, synaptic weights are constant for $T \geq T_s$ consecutive dynamics steps, where T is large. This defines an "adaptation epoch". At the end of the adaptation epoch, synaptic weights are updated according to (5.25). This has the consequence of modifying neurons dynamics and possibly spike trains. The weights are then updated and a new adaptation epoch begins. We denote by t the update index of neuron states (neuron dynamics) inside an adaptation epoch, while τ indicates the update index of synaptic weights (synaptic plasticity). Call $\mathbf{X}^{(\tau)}(t)$ (see section 5.1.2) the state of the neurons at time t within the adaptation epoch τ. Let $W_{ij}^{(\tau)}$ be the synaptic weights from neuron j to neuron at i in the τ-th adaptation epoch. At the end of each adaptation epoch, the neuron dynamics indexes are reset, and $x_i^{(\tau+1)}(0) = x_i^{(\tau)}(T), i = 1 \ldots N$. The coupled dynamics writes:

$$\left\{ \begin{array}{ccc} \mathbf{X}^{(\tau)}(t+1) & = & \mathbf{F}_{\boldsymbol{\gamma}^{(\tau)}}(\mathbf{X}^{(\tau)}(t)) \\ \delta W_{ij}^{(\tau)} & \stackrel{\text{def}}{=} & W_{ij}^{(\tau+1)} - W_{ij}^{(\tau)} = g\left(W_{ij}^{(\tau)}, [\omega_i]_{t-T}^{t}, [\omega_j]_{t-T}^{t}\right) \end{array} \right. \tag{5.32}$$

Recall that $\boldsymbol{\gamma} = (\mathcal{W}, \mathbf{I}^{(ext)})$ (see section 5.1.2) and $\boldsymbol{\gamma}^{(\tau)}$ is the set of parameters at adaptation epoch τ. In the present setting the external current $\mathbf{I}^{(ext)}$ is kept fixed and only synaptic weights are evolving. Basically, $\mathbf{I}^{(ext)}$ is used as an *external stimulus*.

Statistical effects of synaptic plasticity.

The coupled dynamics (5.32) has several prominent effects. First, obviously, a change in the synaptic weights changes the orbits of (5.1).Thus, the attractors and their attraction basin are modified.Furthermore, the spikes train statistics are also modified. Typically, the time-empirical average of the raster plot changes ($\pi^{(T)}_{\omega^{(\tau)}} \to \pi^{(T)}_{\omega^{(\tau+1)}}$) and the corresponding statistical model evolves.

Static synaptic weights. Let us first consider the situation where $\delta\mathcal{W} = 0$, corresponding to synaptic weights that do not evolve. Typically, this is the case if synaptic weights matrix converges to an asymptotic value $\mathcal{W}^*$. From eq. (5.26) this corresponds to :

$$\pi^{(T)}\left[\phi_{ij}\right] = 0,\ \forall i,j \in \{1,\dots N\}\,. \tag{5.33}$$

This imposes a constraint on the average value of ϕ_{ij}. Therefore, from section 5.1.5 *this suggest to propose a statistical model is a Gibbs measure ν with a potential of form*:

$$\boldsymbol{\psi}^* = \Phi + \boldsymbol{\lambda}^*.\boldsymbol{\phi}, \tag{5.34}$$

where $\boldsymbol{\psi}^* = \left(\psi^*_{ij}\right)_{i,j=1}^{N}$, $\boldsymbol{\phi} = (\phi_{ij})_{i,j=1}^{N}$, $\boldsymbol{\lambda}^* = \left(\lambda^*_{ij}\right)_{i,j=1}^{N}$ and $\boldsymbol{\lambda}^*.\boldsymbol{\phi} = \sum_{i,j=1}^{N}\lambda^*_{ij}\phi_{ij}$. The potential Φ in (5.34) is such that $\Phi(\omega) = 0$ is ω is admissible and $\Phi(\omega) = -\infty$ if it is forbidden, so that forbidden raster plots have zero probability. This is a way to include the grammar (i.e, the constraints on symbolic sequences that can be produced by the dynamical system 5.1) on the in the potential. The statistical parameters λ^*_{ij}, are given by eq. (5.15) in section 5.1.5, and making ϕ_{ij} explicit (eq. (5.27)):

$$\frac{\partial P}{\partial \lambda_{ij}}(\boldsymbol{\lambda}^*) = \nu(\phi_{ij}) = \sum_{l_1,l_2\in L} h_{ijl_1l_2}(W^*_{ij})\nu\left[m_{i,l_1}m_{j,l_2}\right] = 0. \tag{5.35}$$

Since $m_{i,l_1}m_{j,l_2}$ are monomials, this equation thus imposes *constraints* on the probability

$$\nu\left[\omega_i(t_1)\dots\omega_i(t_{l_1})\omega_j(s_1)\dots\omega_j(s_{l_2})\right]$$

of spikes n-uplets $\omega_i(t_1)\dots\omega_i(t_{l_1})\omega_j(s_1)\dots\omega_j(s_{l_2})$. Note that condition (5.35) corresponds to an *extremum* for the topological pressure $P\left[\boldsymbol{\lambda}\right]$ in 5.10, as a function of $\boldsymbol{\lambda}$ and consequently of the parameters W_{ij}.

Let us emphasize what we have obtained. The statistical model that fits with the condition (5.33) is a Gibbs distribution such that the probability of a spin block R of length n is given by :

$$\nu\left[R|S\right] = \frac{1}{Z_n\left[\boldsymbol{\lambda}^*(S)\right]} \sum_{\omega \subset [R]} \exp\left[\sum_{t=1}^{n} \boldsymbol{\psi}^*(\sigma_{\boldsymbol{\gamma}}^t \omega)\right]. \tag{5.36}$$

where $\boldsymbol{\psi}^*$ depends on S via the statistical parameters $\boldsymbol{\lambda}^*$ and via Φ (grammar).

When the situation $\delta\mathcal{W} = 0$ corresponds to the asymptotic state for a synaptic adaptation process, this potential provides us the form of the statistical model *after adaptation*, and *integrates all past changes in the synaptic weights*. We now discuss this process within details.

Variational formulation of synaptic plasticity.

Synaptic weights update and related potentials. Let us now formalize the coupled evolution (5.32) in the context of thermodynamic formalism. From section 5.1.2, each adaptation step, $\pi^{(T)}_{\omega^{(\tau)}}$ can be approximated by a Gibbs measure $\nu_{\boldsymbol{\psi}^{(\tau)}}$ with potential $\boldsymbol{\psi}^{(\tau)}$ and topological pressure $P\left[\boldsymbol{\psi}^{(\tau)}\right]$. Therefore, when T is large, synaptic adaptation writes :

$$\delta W_{ij}^{(\tau)} = \epsilon \nu_{\boldsymbol{\psi}^{(\tau)}}\left[\phi_{ij}(W_{ij}^{(\tau)}, .)\right]. \tag{5.37}$$

The synaptic update results in a change of parameters $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}^{(\tau+1)} = \boldsymbol{\gamma}^{(\tau)} + \delta\boldsymbol{\gamma}^{(\tau)}$. This induces a variation of the potential $\boldsymbol{\psi}^{(\tau+1)} = \boldsymbol{\psi}^{(\tau)} + \delta\boldsymbol{\psi}^{(\tau)}$ and of the pressure $P\left[\boldsymbol{\psi}^{(\tau+1)}\right] = P\left[\boldsymbol{\psi}^{(\tau)}\right] + \delta P^{(\tau)}$. We now distinguish two situations both arising when synaptic weights evolve.

Smooth variations. Let us first assume that these variations are smooth, that the topological pressure is differentiable with respect to the variation $\delta\boldsymbol{\gamma}^{(\tau)}$ Let us define:

$$\mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}(\mathcal{W}) = P\left[\boldsymbol{\psi}^{(\tau)} + (\mathcal{W} - \mathcal{W}^{(\tau)}).\boldsymbol{\phi}(\mathcal{W}^{(\tau)})\right] - P\left[\boldsymbol{\psi}^{(\tau)}\right], \tag{5.38}$$

so that $\mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}(\mathcal{W}^{(\tau)}) = 0$. Note that $\mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}(\mathcal{W})$ is convex, due to the convexity of the topological pressure (eq. (5.14)).

Then, using eq. (5.15), the adaptation rule (5.37) can be written in the form:

$$\delta\mathcal{W}^{(\tau)} = \epsilon \nabla_{\mathcal{W}=\mathcal{W}^{(\tau)}} \mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}(\mathcal{W}). \tag{5.39}$$

Since, in this section, pressure P is assumed to be smooth, one has, using (5.15),(5.16):

$$\epsilon\left(P\left[\boldsymbol{\psi}^{(\tau)}+\delta\mathcal{W}^{(\tau)}.\boldsymbol{\phi}(\mathcal{W}^{(\tau)})\right]-P\left[\boldsymbol{\psi}^{(\tau)}\right]\right)=\delta\mathcal{W}^{(\tau)}.\left[I+\frac{\epsilon}{2}\kappa^{(\tau)}\right].\delta\mathcal{W}^{(\tau)}+O(\delta\mathcal{W}^{(\tau)^3}),\tag{5.40}$$

where $\kappa^{(\tau)}$ is the tensor with entries:

$$\kappa^{(\tau)}_{i_1,j_1,i_2,j_2}=C^{(\tau)}_{\phi_{i_1,j_1}\phi_{i_2,j_2}}(0)+2\sum_{t=1}^{+\infty}C^{(\tau)}_{\phi_{i_1,j_1}\phi_{i_2,j_2}}(t);\qquad i_1,j_1,i_2,j_2=1\dots N,\tag{5.41}$$

and $C^{(\tau)}_{\phi_{i_1,j_1}\phi_{i_2,j_2}}(t)=\nu_{\boldsymbol{\psi}^{(\tau)}}\left[\phi_{i_1,j_1}\circ\sigma^t_{\boldsymbol{\gamma}^{(\tau)}}\phi_{i_2,j_2}\right]-\nu_{\boldsymbol{\psi}^{(\tau)}}\left[\phi_{i_1,j_1}\right]\nu_{\boldsymbol{\psi}^{(\tau)}}\left[\phi_{i_2,j_2}\right]$ is the correlation function of $\phi_{i_1,j_1},\phi_{i_2,j_2}$ for the measure $\nu_{\boldsymbol{\psi}^{(\tau)}}$. Using the explicit form (5.27) of ϕ_{ij} one can see that $\kappa^{(\tau)}$ is a sum of time correlations between uplets of spikes. This is a version of the fluctuation-dissipation theorem where the response to a smooth variation of the potential $\boldsymbol{\psi}^{(\tau)}$ is given in terms of a series involving the time correlations of the perturbation [161, 162]. This series converges provided dynamics is hyperbolic as it is the case for eq. 5.22.

For sufficiently small ϵ, the matrix $I+\frac{\epsilon}{2}\kappa^{(\tau)}$ is positive and:

$$P\left[\boldsymbol{\psi}^{(\tau)}+\delta\mathcal{W}^{(\tau)}.\boldsymbol{\phi}(\mathcal{W}^{(\tau)})\right]\geq P\left[\boldsymbol{\psi}^{(\tau)}\right].\tag{5.42}$$

It follows that the variation $\delta^{(\tau)}\mathcal{F}_\phi=\mathcal{F}^{(\tau+1)}_{\boldsymbol{\phi}}(\mathcal{W}^{(\tau+1)})-\mathcal{F}^{(\tau)}_{\boldsymbol{\phi}}(\mathcal{W}^{(\tau+1)})$ is given by:

$$\delta^{(\tau)}\mathcal{F}_{\boldsymbol{\phi}}=P\left[\boldsymbol{\psi}^{(\tau)}\right]-P\left[\boldsymbol{\psi}^{(\tau)}+\delta\mathcal{W}^{(\tau)}.\boldsymbol{\phi}(\mathcal{W}^{(\tau)})\right]=-\frac{1}{\epsilon}\delta\mathcal{W}^{(\tau)}.\left[I+\frac{\epsilon}{2}\kappa^{(\tau)}\right].\delta\mathcal{W}^{(\tau)}-O(\delta\mathcal{W}^{(\tau)^3})\tag{5.43}$$

This variation is therefore *negative* when ϵ is sufficiently small.

We come therefore to the following important conclusion. The adaptation rule (5.26) is a *gradient* system where the function $\mathcal{F}^{(\tau)}_{\boldsymbol{\phi}}$ decreases when iterating synaptic adaptation rules. If the transition $\tau\to\tau+1$ is smooth for all τ, $\mathcal{F}^{(\tau)}_{\boldsymbol{\phi}}$ reaches a minimum[6] at some $\mathcal{W}^*$ as $\tau\to\infty$. Such a minimum corresponds to $\nabla_{\mathcal{W}^*}\mathcal{F}^{(\tau)}_{\boldsymbol{\phi}}=0$, thus to $\delta\mathcal{W}=0$ according to eq. (5.39). Hence, this minimum corresponds to a *static distribution* for the synaptic weights. Therefore, according to section 5.3.2 the potential $\boldsymbol{\psi}^{(\tau)}$ converges to the potential (5.34) as $\tau\to\infty$. In general, we cannot expect these transitions to be smooth for all τ, however in this case it is guarantee in system of eq. 5.22 by the exponential decay of potential given in 5.23 with the memory depth

[6]Additional constraints are required ensuring that $\mathcal{F}^{(\tau)}_{\boldsymbol{\phi}}$ does not tend to $-\infty$. Typically, such constraints amount to bound the synaptic weights variation (soft or hard-bounds rules) by a suitable choice of functions $h_{ijl_1l_2}$ in eq. (5.27).

(absence of phase transition). Furthermore, as $\sigma_B \to 0$ on goes to sort of a "spin glass“ phase (coexistence of many stable periodic orbits [24], and for finite observations, σ_B small irregularities in the statistics variations with $\delta\mathcal{W}^{(\tau)}$ may be observed)

5.4 A numerical example.

5.4.1 Model.

Adaptation rule.

As an example we consider an adaptation rule inspired from (5.28) with an additional term $r_d W_{ij}^{(\tau)}$, $-1 < r_d < 0$, corresponding to passive LTD.

$$\delta W_{ij}^{(\tau)} = \epsilon \left[r_d W_{ij}^{(\tau)} + \frac{1}{T} \sum_{t=T_s}^{T+T_s} \omega_j^{(\tau)}(t) \sum_{u=-T_s}^{T_s} f(u)\, \omega_i^{(\tau)}(t+u) \right], \tag{5.44}$$

where $f(x)$ is given by (5.29) and with:

$$T_s \stackrel{\text{def}}{=} 2\max(\tau_+, \tau_-).$$

Set :

$$S_i(\omega) = \sum_{u=-T_s}^{T_s} f(u)\, \omega_i(u), \tag{5.45}$$

$$H_{ij}(\omega) = \omega_j(0) S_i(\omega),$$

$\mathbf{H} = \{H_{ij}\}_{i,j=1}^N$, and:

$$\phi_{ij}(W_{ij}, \omega) = r_d W_{ij} + H_{ij}(\omega),$$

with $\boldsymbol{\phi} = \{\phi_{ij}\}_{i,j=1}^N$, where ϕ_{ij} is a finite range potential with range $2T_s$. Then (5.44) has the form (5.26), $\delta W_{ij}^{(\tau)} = \epsilon \pi_{\omega^{(\tau)}}^{(T)}[\phi_{ij}(W_{ij}, .)]$.

Effects of the adaptation rule on the synaptic weights distribution.

The term $S_i(\omega)$ (eq. 5.45) can be either negative, inducing Long Term Depression, or positive inducing Long Term Potentiation. In particular, its average with respect to the empirical measure $\pi_{\omega^{(\tau)}}^{(T)}$ reads:

$$\pi_{\omega^{(\tau)}}^{(T)}(S_i) = \eta r_i(\tau) \tag{5.46}$$

where:

$$\eta = \left[A_- e^{-\frac{1}{\tau_-}} \frac{1 - e^{-\frac{T_s}{\tau_-}}}{1 - e^{-\frac{1}{\tau_-}}} + A_+ e^{-\frac{1}{\tau_+}} \frac{1 - e^{-\frac{T_s}{\tau_+}}}{1 - e^{-\frac{1}{\tau_+}}}. \right] \tag{5.47}$$

and where $r_i(\tau) = \pi^{(T)}_{\omega^{(\tau)}}(\omega_i)$ is the frequency rate of neuron i in the τ-th adaptation epoch.

The term η neither depend on ω nor on τ, but only on the adaptation rule parameters $A_-, A_+, \tau_-, \tau_+, T_s$. Equation (5.46) makes explicit 3 regimes.

- **Cooperative regime.** If $\eta > 0$ then $\pi^{(T)}_{\omega^{(\tau)}}(S_i) > 0$. Then synaptic weights have a tendency to become more positive. This corresponds to a cooperative system [83]. When iterating adaptation, dynamics become trivial with neurons firing at each time step or remaining quiescent forever.

- **Competitive regime.** On the opposite if $\eta < 0$ synaptic weights become negative. This corresponds to a competitive system [83].

- **Intermediate regime.** The intermediate regime corresponds to $\eta \sim 0$. Here no clear cut tendency can be distinguished from the average value of S_i and spikes correlations have to be considered as well.

Static weights.

Thanks to the soft bound term $r_d W_{ij}$ the synaptic adaptation rule admits a static solution given by:

$$W_{ij} = -\frac{\pi^{(T)}_{\omega^{(\tau)}}\left[\omega_j(0) S_i(\omega)\right]}{r_d}. \tag{5.48}$$

Note that this equation can have several solutions.

Using the same decomposition as the one leading to (5.46) we obtain:

$$\pi^{(T)}_{\omega^{(\tau)}}\left[\omega_j(0) S_i(\omega)\right] = A_- \sum_{u=-T_s}^{-1} e^{\frac{u}{\tau_-}} \pi^{(T)}_{\omega^{(\tau)}}\left[\omega_j(0)\, \omega_i(u)\right] + A_+ \sum_{u=1}^{T_s} e^{-\frac{u}{\tau_+}} \pi^{(T)}_{\omega^{(\tau)}}\left[\omega_j(0)\, \omega_i(u)\right].$$

Note that $\omega_j(0)\, \omega_i(u) \geq 0$, thus the first term is negative and the second one is positive (see eq. (5.29)). The sign of W_{ij} depend on the parameters A_-, A_+, T_s, but also on the relative strength of the terms $\pi^{(T)}_{\omega^{(\tau)}}\left[\omega_j(0)\, \omega_i(u)\right]$.

Convergence to the static weights solution.

The synaptic adaptation rule (5.44) defines a mapping $\mathcal{Q}$ on the set of synaptic weights:

$$W_{ij}^{(\tau+1)} = W_{ij}^{(\tau)}(1 + \epsilon r_d) + \epsilon \pi^{(T)}_{\omega^{(\tau)}}\left[H_{ij}\right] = \mathcal{Q}_{ij}(\mathcal{W}^{(\tau)})$$

with $|1 + \epsilon r_d| < 1$, where $\pi^{(T)}_{\omega^{(\tau)}}$ depends on $\mathcal{W}^{(\tau)}$. Thus,

$$\frac{\partial \mathcal{Q}_{ij}}{\partial W_{kl}} = (1+\epsilon r_d)\delta_{ij,kl} + \epsilon \frac{\partial \pi^{(T)}_{\omega^{(\tau)}}[H_{ij}]}{\partial W_{kl}},$$

where $\delta_{ij,kl} = 1$ if $i = k, j = l$ and 0 otherwise. The second term is a linear response characterizing the variation of $\pi^{(T)}_{\omega^{(\tau)}}[H_{ij}]$ with respect to small variations of W_{kl}. Approximating $\pi^{(T)}_{\omega^{(\tau)}}$ by a Gibbs distribution with a potential $\psi^{(\tau)}$, it writes $C_{\psi^{(\tau)},H_{ij}}(0) + 2\sum_{t=0}^{+\infty} C_{\psi^{(\tau)},H_{ij}}(t)$ where $C_{\psi^{(\tau)},H_{ij}}(t)$ is the time-t correlation function between $\psi^{(\tau)}$ and H_{ij}. When $\nu_{\psi^{(\tau)}}$ is smooth with respect to $\mathcal{W}$, the derivative

The contraction property ensures the convergence to a static solution for any initial condition in this ball (Brouwer theorem).

In this thesis we postulate the convergence and check it numerically.

Spike train statistics in a static weights regime.

As emphasized in section 5.3.2, when the synaptic adaptation rule converges to a fixed point, the corresponding statistical model is approximated by a Gibbs measure with a potential:

$$\psi^* = \Phi + \boldsymbol{\lambda}^*.\boldsymbol{\phi},$$

where Φ contains the grammar and $\boldsymbol{\lambda}$ are free statistical parameters. The value $\boldsymbol{\lambda}^*$ of these parameters in the potential ψ^* is determined by the relation:

$$\left.\frac{\partial P[\psi]}{\partial \lambda_{ij}}\right|_{\boldsymbol{\lambda}^*} = r_d W^*_{ij} + \nu_{\psi^*}[H_{ij}] = 0,\ \forall i,j\,, \tag{5.49}$$

where the pressure is given by:

$$P[\psi] = r_d \boldsymbol{\lambda}.\mathcal{W} + \lim_{T\to\infty} \frac{1}{T} \log \sum_{\omega\in\Sigma^{(T)}} e^{\boldsymbol{\lambda}.S_T\mathbf{H}(\omega)}.$$

This procedure provides us the explicit form of the raster plot probability distribution when the adaptation rule converges. But the price to pay is that we have to determine *simultaneously* the N^2 parameters λ^*_{ij} on which the Gibbs measure depends. Focusing on the joint probability of a small set of neurons (pairs, triplets) this constraint can be relaxed in order to be numerically tractable.

5.4.2 Numerical checks.

The main goal of section 5.4 is to provide an example of the theoretical concepts developed above. We focus here on numerical simulations for BMS model for the case of η value corresponding to the intermediate regime defined in section 5.4.1.

Implementation.

Neurons dynamics. Previous numerical explorations have shown that a BMS- network of N neurons, with synapses taken randomly from a distribution $\mathcal{N}(0, \frac{C^2}{N})$, where C is a control parameter, exhibits a dynamics with very large periods in specific regions of values of the space (ρ, C) [24, 28]. On this basis, we choose $N = 100,\ \rho = 0.95,\ C = 4.0, N = 100$. The external current $\mathbf{I}^{(ext)}$ in eq. (5.22) is given by $I_i^{ext} = 0.06 + 0.01\mathcal{N}(0, 1)$. Note that fixing a sufficiently large average value for this current avoids a situation where neurons stops firing after a certain time ("neural death").

STDP implementation. An efficient implementation of a STDP rule does not only depend on the analytic form of the rule but also on the respective time scales characterizing neurons and synapses evolutions. In "on-line-protocols", where the synaptic changes occur at the same time scale as neurons dynamics the so-called recursive approach appears to be the more appropriated (see [214] and references therein). In our case, we use instead an offline protocol, where we register the dynamics of the system on a long time windows, then compute the STDP modification, then let the system evolve again to its attractor (see section 5.3.2). The offline protocols are very expensive in machine-time, especially when using long spike trains to get a reliable time average ($\pi^{(T)}_{\omega^{(\tau)}}(\phi_{ij})$). For this reason we need to add some words about our implementation.

We register spike trains in a binary code. Indeed, this is the cheapest way in memory requirements, though it might be expensive for accessing specific spike times. Also, bit-wise operations are faster than their equivalents on other types of data. Finally, there exist very fast methods for computing the number of bits on any specific variable of type of integer (The faster for large number of iterations is a look-up table of precomputed values, but in-line methods - using parallel methods based on masks- are not so far in terms of performances. For details and speed comparison see the G. Manku website http://infolab.stanford.edu/ manku/bitcount/bitcount.html). Numerical comparison of this method with the direct one that records the dynamics in Boolean arrays and compute STDP spike by spike shows enormous performance difference growing exponentially as the length of trains increases.

Computation of $\delta^{(\tau)}\mathcal{F}_\phi$. To check the result in section 5.3.2, stating that $\mathcal{F}^{(\tau)}_\phi(\mathcal{W})$ is a decreasing function of τ (i.e. $\delta^{(\tau)}\mathcal{F}_\phi$ is negative) during regular periods, we use the following method [29].

Fix a potential ψ. If T is the length of the experimental raster plot, one

divides it into k blocs of length n, such that $T = k \times n$. Call:

$$P_{n,k}(\omega) = \frac{1}{n} \log \left(\frac{1}{k} \sum_{j=0}^{k-1} \exp(S_j^{(n)}(\omega)) \right), \tag{5.50}$$

with $S_j^{(n)}(\omega) = \sum_{t=jn}^{jn+n-1} \boldsymbol{\psi}(\sigma^t \omega)$. Then, the following result can be proved [29]:

$$P[\psi] = \lim_{n \to +\infty} \lim_{k \to +\infty} P_{n,k}(\omega), \tag{5.51}$$

for $\nu_{\boldsymbol{\psi}}$ almost-every ω. This computation requires the knowledge of $\boldsymbol{\psi}$.

The result in [29] can be extended straightforwardly to the following case. If $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2$ are two potentials for the same grammar, and $\delta\boldsymbol{\psi} = \boldsymbol{\psi}_2 - \boldsymbol{\psi}_1$ then:

$$P[\boldsymbol{\psi}_2] - P[\boldsymbol{\psi}_1] = \lim_{n \to +\infty} \lim_{k \to +\infty} \frac{1}{n} \log \left(\frac{1}{k} \sum_{j=0}^{k-1} \exp(\sum_{t=jn}^{jn+n-1} \delta\boldsymbol{\psi}(\sigma^t \omega)) \right) \tag{5.52}$$

Since $\delta^{(\tau)} \mathcal{F}_\phi = P\left[\boldsymbol{\psi}^{(\tau)}\right] - P\left[\boldsymbol{\psi}^{(\tau)} + \delta\mathcal{W}^{(\tau)}.\boldsymbol{\phi}(\mathcal{W}^{(\tau)})\right]$,

$$\delta^{(\tau)} \mathcal{F}_\phi = - \lim_{n \to +\infty} \lim_{k \to +\infty} \frac{1}{n} \log \left(\frac{1}{k} \sum_{j=0}^{k-1} \exp\Big(\sum_{t=jn}^{jn+n-1} \delta\mathcal{W}^{(\tau)}.\boldsymbol{\phi}(\mathcal{W}^{(\tau)}, \sigma^t(\omega^{(\tau)}))\Big) \right) \tag{5.53}$$

Expanding this equation in power series of ϵ one recovers the expansion (5.40). In particular, subtracting to (5.53) the leading term (of order $\delta\mathcal{W}^{(\tau)}.\delta\mathcal{W}^{(\tau)}$) one gets an estimation of the "linear response" term $\kappa^{(\tau)}_{i_1,j_1,i_2,j_2}$ in eq. (5.41). However, on practical grounds, the computation of (5.53) requires very long time series to reduce significantly the finite fluctuations of the empirical average. Similar problems are encountered in the computation of linear response from other methods [23].

This result holds for any adaptation rule (any polynomial $\boldsymbol{\phi}$). In the case of the adaptation rule (5.44), this reduces to computing:

$$\delta^{(\tau)} \mathcal{F}_\phi = -r_d \delta\mathcal{W}^{(\tau)}.\mathcal{W}^{(\tau)} - \lim_{n \to +\infty} \lim_{k \to +\infty} \frac{1}{n} \log \left(\frac{1}{k} \sum_{j=0}^{k-1} \exp\Big(\sum_{t=jn}^{jn+n-1} \epsilon\delta\mathcal{W}^{(\tau)}.\mathbf{H}(\mathcal{W}^{(\tau)}, \sigma^t(\omega^{(\tau)}))\Big) \right).$$

Simulation results.

We have run simulations for STDP parameters values: $r_d = 0.99$, $\epsilon = 0.001$, $\tau_+ = 16$, $\tau_- = 32$, $A_+ = 1.0$, and $T_s = 64$, corresponding to standard values [91]. We fix $\eta = -0.01$ corresponding to the intermediate regime discussed in section 5.4.1. This fixes the value of A_- via eq. (5.47). The

length of spike trains is $T = 3941 = 4096 - 128 = 128 \times 32 - 2 \times T_s$, where 32 is the number of bits in long integer, in our (machine-dependent) implementation. An extended description will be published elsewhere. The main results are summarized in fig. 5.2.

In this figure (Top) we have represented the evolution of the distribution of synaptic weights $\mathcal{W}^{(\tau)}$ (see captions). For this value of η, the histogram evolves to a unimodal distribution with some nodes having strong synaptic weights. In fig. 5.2 (center) we depict the raster plots at the beginning and at the end of synaptic adaptation process. In fig. 5.2 (bottom-left) we have shown the evolution of the Frobenius norm for $\delta\mathcal{W}^{(\tau)}$ (i.e, $\sum_{i,j}^{N} |\delta\mathcal{W}_{ij}^{(\tau)}|^2$), which converges to 0. This shows the convergence of the rule in this case. We have also plotted the average weights modification $\frac{1}{N^2}\sum_{i,j}\delta W_{ij}$. Finally in fig. 5.2 (bottom -right) is represented $\delta^{(\tau)}\mathcal{F}_\phi$ computed with 512 block of size 512. We see clearly that this quantity is negative as expected from eq. (5.43) and tends to 0. We have also represented the term $-\frac{1}{\epsilon}\delta\mathcal{W}^{(\tau)}.\delta\mathcal{W}^{(\tau)}$ corresponding to the first order expansion in eq. (5.43).

Approximated Gibbs distribution

Here we use the algorithm presented in chapter 3 to see which is the Gibbs distribution given in 5.34 from the exacts statistics.

Simulation set-up The simulation parameters we used here were: $N = 10,\ \gamma = 0.995,\ C = 0.2$. The external current $\mathbf{I}^{(ext)}$ in eq. (5.22) is given by $I_i^{ext} = 0.01$ while $\sigma_B = 0.01$. We register the activity after 4000 steps of adaptation with the STPD rule proposed in (5.44).

When considering a large number of neurons, it becomes difficult to compute and check numerically this joint probability over the whole population. Here, we propose to consider a subset $\mathcal{P}_s$ of $N_s < N$ neurons. In this context we expect the potential for the whole population to be of the form (3.2):

$$\psi(\omega) = \sum_{i=0}^{N} \lambda_i^{(1)}\omega_i(0) + \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\lambda_{ij}^{(2)}\sum_{u=-T_s}^{T_s} f(u)\omega_i(0)\omega_j(u).$$

For this potential the effects of the rest of the population can be written as a bulk term modulating the individual firing rates and correlations of the observed population, leading to a marginal potential of the form:

$$\psi_{\mathcal{P}_s}(\omega) = \sum_{i\in\mathcal{P}_s} \lambda_i^{(1)}\omega_i(0) + \sum_{i,j\in\mathcal{P}_s 0}\sum_{j=0}^{N-1}\lambda_{ij}^{(2)}\sum_{u=-T_s}^{T_s} f(u)\omega_i(0)\omega_j(u). \tag{5.54}$$

Therefore, we choose randomly 2 neurons among the N and we construct from them the prefix-tree. Then, for the 2 neuron potentials forms

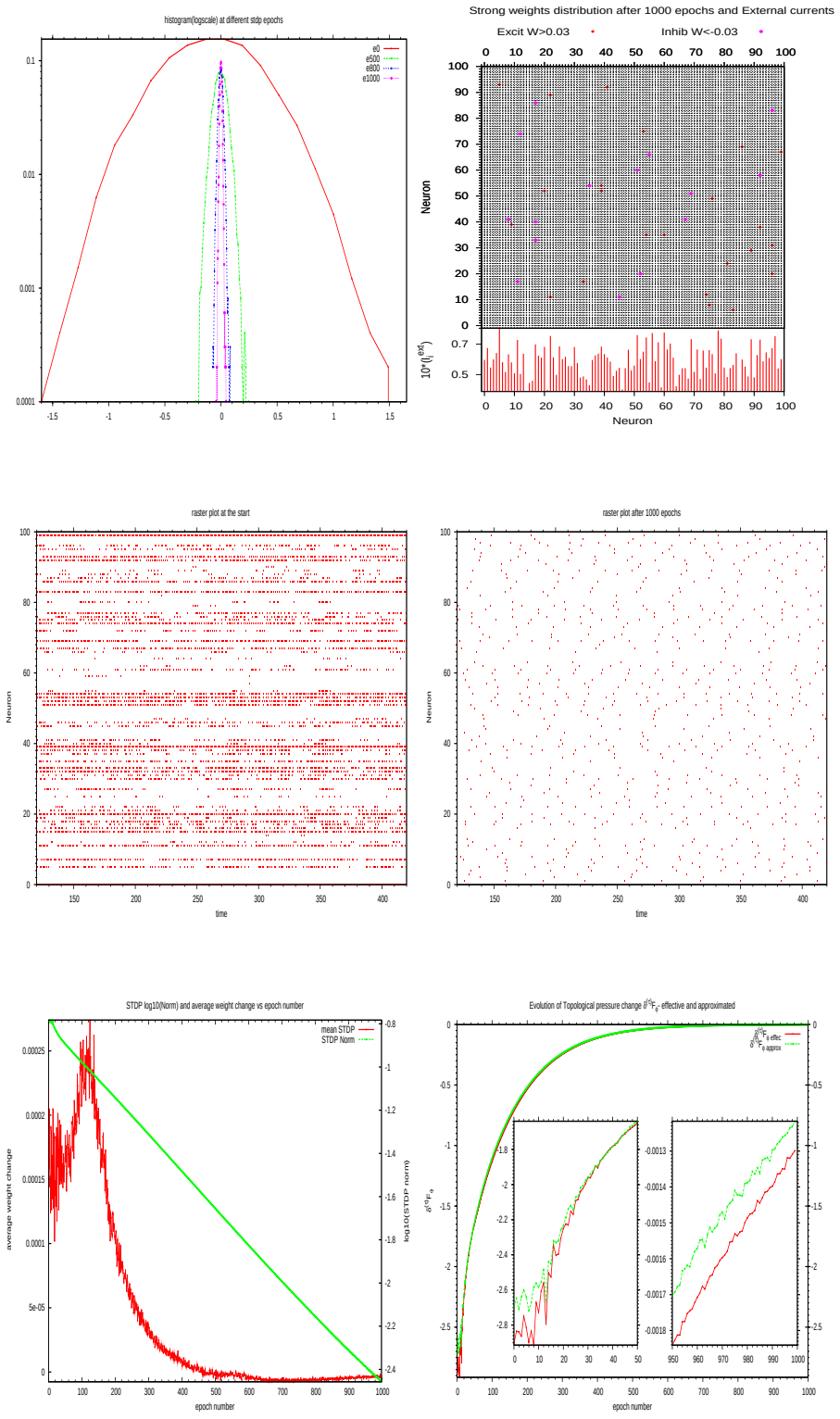

Figure 5.2: (color on line). The parameters are $N = 100, r_d = -0.4, \eta = -0.01, \epsilon = 0.01$. **Top**:(left) Histogram in log scale of synaptic connectivity in the network after 1000 epochs of STDP;(right) map of strongest connections, either inhibitory (pink dots) or excitatory (red diamonds) (i.e, $|W_{ij}| > 0.03$). We also include at the bottom of this map the external current of each neuron (multiplied by 10). **Middle**:(left) A comparison of raster plots at the beginning and (right) the end of the synapses evolution. **Bottom**:(left) Evolution of the mean synaptic change $\frac{1}{N^2}\sum_{i,j}^{N}\delta\mathcal{W}_{ij}^{(\tau)}$(red) and the Frobenius norm for STDP matrix on log scale i.e, $\log_{10}\left(\sum_{i,j}^{N}|\delta\mathcal{W}_{ij}^{(\tau)}|^2\right)$ (green); (right) Variations of Topological Pressure (with 2 zoom graphics showing sharp changes) after each epoch τ calculated explicitly in eq. (5.53) with 512 trains of size 512 (red), and first order term approximated by $\frac{1}{\epsilon}\delta\mathcal{W}^{(\tau)}.\delta\mathcal{W}^{(\tau)}$(green).

from (4.1), we estimate the coefficients that minimizes the Kullback-Leibler divergence.The probability of words of different sizes predicted by several statistical models from (4.1) versus empirical probability $\pi_\omega^{(T)}(w)$ obtained from a spike train and the corresponding $\tilde{h}$ value of the estimation process for a fixed pair of neurons are shown on figure (5.3).
Results depicted on figure (5.3) show, on one hand, that the statistics is well fitted by (5.54). Moreover, the best statistical models, are those including rate terms (the differences between their KL value is two orders of magnitude smaller that within those not disposing of rate terms). We also note that for the words with the smallest probability values, the potential do not yields a perfect matching due to finite size effects (see fig (5.3)). Especially, the small number of events due to low firing rates of neurons makes more sensitive the relation between the length of observed sequences (word size) and the spike-train length necessary to provide a good sampling and hence a reliable empirical probability.

Conclusion

We have shown, the application of principles in thermodynamic formalism, to show Gibbs distributions could constitute good prototypes for the analysis of spike trains statistics of neural networks. In particular, it exists at least one[7] exact result of this statement: BMS neural networks with noise. We also have shown how the topological pressure of the associate Gibbs potential may contain information about the evolution of the internal system parameters, such as the synaptic weights. For a proposed STDP rule, we have shown that plasticity can be seen as a sort of variational system and we have verified it numerically. Additionally we have performed parametric estimation of a finite range approximation of the asympotic Gibbs distributions after STDP evolution process. The approximated distribution seems to show that for the intermediate regime (see section 5.4.1), the activity of pairs of neurons is almost independent, however further studies over larger subsets of the network are still to be done. Also it would be possible to study the effects of other synaptic plasticity rules and network structures, including other features as the Dale's principle (a neuron is exclusivly either excitatory or inhibitory, ie. for a fixed j, all the connections W_{ij} have the same sign and the change of sign by a plasticity mechanism is forbidden).

[7] up to our knowledge, this is the only example of Neural Network where the Gibbs measure is proved to exist and explicitly computed.

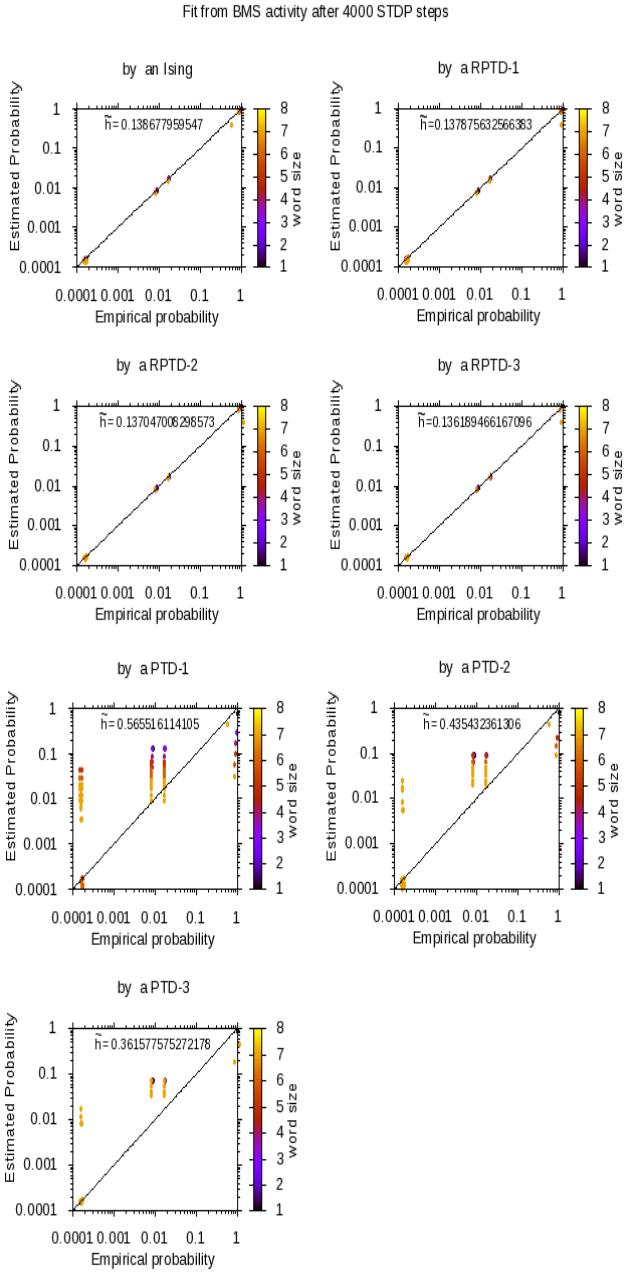

Figure 5.3: The probability of words of different sizes predicted by several statistical models from (4.1) versus empirical probability $\pi_\omega^{(T)}(w)$ obtained from a spike train generated by dynamics (5.22) after 4000 epochs of adaptation.The $\tilde{h}$ value (2.38) for each fitting model is shown inside the graphic. The potential is a pair potential of the form (5.54). Recall that **RPTD** Models include firing rates but **PTD** models do not.

Chapter 6

Conclusion

“”

–

Contents

6.1 Discussion of the thesis work

Position of the Problem How can we characterize neural network dynamics, at least its statistical properties? At the current state-of-the-art, still much of the role of neural assemblies and their internal interactions remains unknown. The simultaneously recording of the activity of groups of neurons (up to several hundreds) over a dense configuration, supplies a critical database to unravel the role of specific neural assemblies, therefore, new approaches and methods need to be developed/integrated within the theoretical and Computational Neuroscience. A group of approaches which originated in statistical physics, more precisely in Ising models and spin glass theory, have already proven to be an interesting approach [166, 175], and provide interesting and useful applications in the field of neural decoding and machine learning[165]. Nevertheless, in spite of their success, Ising models current approaches present several limitations that have been recently pointed out in different works[159**?** , 134] . In particular we remark that as small ensemble description, Ising models suffer from an intrinsic weakness that provide limited prediction power for large ensemble statistical characterization, convincingly investigated in [159] (this will be further discussed in the last section).

Strategy Since Ising models belongs to the class of models issued from a Maximum entropy principle under constraints, in this thesis, we have adopted a pragmatic approach based on the idea that maximum entropy statistical models (called Gibbs Distributions) obtained via the applications of principles in thermodynamic formalism (which are basically the same as those used in statistical physics) could constitute good prototypes for the analysis of real spike trains statistics. In this way, our goal is not to prove theorems but instead to use already proved theorems. However, theorems have hypotheses whose plausible validity is also discussed. This work extending maximum entropy modeling to characterize Spike Train Statistics in Neural networks is based in Gibbs Distributions and the thermodynamic formalism [161, 102] and provides a Numerical implementation of the method is given and numerical test cross-validating it. The implementation is intended to be open source an a ready to-use C++ header to be used as a plug-in by other users.

Results summary -discussion In chapter 2 we have presented the framework of finite range Gibbs distributions adapt for analysis of spike train statistics. We have use as starting point to relate them to Markov

chains of higher order, although a more general presentation of the formalism and its relation to neural activity as a dynamical systems is presented chapter 5.

For finite range observables, the Gibbs potential is expressed as a linear combination of "spiking events" here called monomials (e.g. neuron i fires at time t_i and neuron j fires at time t_j), where each monomial has an associated parameter (corresponding in the Ising models to the local magnetization and pairwise couplings) and where the largest "event" duration is called the potential range $R+1$. Consequently, high-order spatial and temporal correlations can be taken into account, (in the spirit of Jaynes principle this means to add arbitrary constraints). We are able to associate to this potential a Markov chain of memory R, and to compute its time-invariant (stationary) measure. This is far from straightforward when considering general parametric potentials with arbitrary coefficients (Ising potentials constitute a very special case since their range is 1 so they associate to memoryless process).

Using Gibbs distributions results and a closed expression for the Kullback-Leibler Divergence(KLD) between a Gibbs measure and another arbitrary invariant measure and we can estimate the parameter values of the Gibbs potential in order to minimize the KLD. This results enables us to propose a parameter estimation algorithm which lies on the the spectral properties of the Perron-Frobenius matrix [171, 66] and this is (up to our knowledge) entirely new in this field.

Additionally, the present formalism allows us to provide closed-form calculations of interesting parameters related to spectral properties. We, for instance, propose an indirect estimation of the entropy, via an explicit formula. We also provide numbers for the average values of the related observable, probability measure, etc. This means that as soon as we obtain the numerical values of the Gibbs distribution up to some numerical precision, all other statistical parameters come for free without additional approximations.

An important feature is that the rate of convergence or put differently the size of finite sample fluctuations of the parameter estimation are analytically related to the second order derivative of the topological pressure thanks to the theory of Large deviations. A result of a related kind it that the framework allows to check the confidence level for the estimate potential ie. the probability of be close to a wrong model, which decays as the exponentially with the finite sample size but depends also on the KLD

A step further, the non-trivial but very precious virtue of the method is that it allows us to efficiently compare models. We thus not only estimate the optimal parameters of a model, but can also determine among a set of

models which model is the most relevant. This means, for instance, that we can determine if either only rates, or rates and correlations matters, for a given piece of data. Another example is to detect if a given spike pattern is significant, with respect to a model not taking this pattern into account. The statistical significance mechanism provides numbers that are clearly different for models corresponding or not to a given empirical distribution, providing also an absolute test about the estimation significance.

We have also mentioned and provided a basic example about the estimation method's ability to retrieve slow time-varying parameters.

Although this was not fully discussed in the thesis manuscript, the formalism allows to integrate restrictions in the grammar (the allowed and forbidden transitions) observed from the empirical data or analytically known from the dynamical system. This is indeed quite important, because grammar encodes/contains might contain key features of the system such as effects of the refractory period and the thumbprint of effective network interactions. We believe that this capability of the framework, to include explicitly grammar effects opens interesting perspectives for future applications on the framework.

In chapter 5, we have shown that Gibbs distributions may naturally arise from adaptation process as statistical models for the network activity [26, 205]. More precisely, we submitted a recurrent networks of BMS-model neurons (The BMS, -for G. Beslon, O, Mazet and H. Soula neuron model, which is a time-discretised Leak-Integrate-and-Fire neuron model), where available analytical results on recurrent networks dynamics are available in [24], and in particular the existence of a Gibbs measure for describing its spike train statistics has been recently proved [25]. The thermodynamic formalism enable one to relate the internal parameters of the system such as the synaptic weights to the Topological Pressure of the Gibbs potential and furthermore, to analyze its evolution. Examples of how to include several adaptation rules within the formalism were provided. Moreover, we have shown that for one of those examples of STDP rule, the synaptic plasticity can be written as a gradient system for the variation of the Topological Pressure. Additionally, a finite range approximation expression for the asymptotic Gibbs distribution after the STPD process was proposed. Such results where verified numerically. Besides, in the light of the ensemble of results provide a new insight about how adaptation process could be perceived (minimization of a certain system function), and the type of non-trivial dynamics that this adaptation mechanisms may achieve

These all elements push forward the state of the art regarding statistical analysis of spike train a step further, since they constitute several advantages of Gibbs distribution approach with respect to previous attempts (see

the comparison and discussion in the next section).

Applications to the analysis of Biological data are preliminary, but they have greatly helped us to gain insights about the framework and how the implementation behaves facing real datasets and also they have confronted us with several other issues/difficulties arising when treating real data. Thanks to this recently acquired knowledge, we expect be able to extend our data analysis in the short future and provide more interesting results from the biological perspective.

Finally, we would like to remark that the numerical implementation of this formalism has been integrated to EnaS library in the format of a C++ header, which constitutes a free/open access source to anyone interested in use/apply the Gibbs distribution framework in Neuroscience and other research fields. However, this implementation remains a first approach, since the numerical perspectives offered by the change of numerical paradigm i.e, from distribution sampling (for instance, by means of Markov Chain Monte-Carlo methods) to a matrix computations, are still to be fully implemented (see the next sections for details)

6.2 Discussion and comparison with existing methods

Let us first summarize the advantages and drawbacks of our method compared with the existing ones. For this, we list some keywords in the approaches used by the community and discuss the links with our own work.

- **Maximum entropy.** The formalism that we use corresponds to a maximum entropy method but without limitations on the number or the type of constraints. Actually, on mathematical grounds, it allows infinitely many constraints. Moreover, we do not need to compute the entropy.

- **Markovian approaches.** Our method is based on a Markovian approach where the memory depth of the Markov chain can be arbitrary long (actually the formalism that we use allows to theoretically consider processes with infinite memory, called *chains with complete connections [116]*, see [25] for an application to spike train statistics). As we developed, the link between the potential extracted from the maximum entropy principle, by fixing *ad hoc* observables, and a Markov chain is not straightforward, since a potential of this kind is not normalized.

- **Monte-Carlo methods.** Equation (2.11) enables us to generate spike trains Gibbs-distributed with and arbitrary potential (not normalized).

The convergence is ensured by eq. (5.11). We do not need to assume detailed balance.

- **Determining an effective synaptic connectivity between neurons.** Interactions between neurons occur via synapses (or gap junction). This interaction is not instantaneous, it requires some delay. As a matter of fact, estimating the synaptic conductancies via the spike statistics requires therefore to consider time-dependent potentials. Our formalism allows this. Determining an effective synaptic connectivity between neurons from spike trains for a BMS model has been presented in [25]

- **Grammar** Indeed, the formalism allows to integrate restrictions in the grammar, (the allowed and forbidden transitions) observed from the empirical data or analytically known from the dynamical system, through the RPF operator. This capability is not a *priori* or intrinsically present in previous maximum entropy approaches.

- **Boltzmann learning.** Our approach can be viewed as "Boltzmann learning" (as presented e.g. in [160]) without restrictions on the parameters that we learn and additionally without using a Monte Carlo approach (which assumes detailed balance). Furthermore, our "learning" rule uses a criterion which is strictly convex.

- **Non Stationarity.** The formalism here presented permits the analysis of the evolution of Gibbs Distributions induced by evolution of system internal parameters. An application of this idea was developed in chapter 5 for neural networks under synaptic plasticity mechanisms and it was published in [26]. In this works, we also suggest that Gibbs measures may result from related adaptation rules whenever synaptic weights converge to a fix value. In this case, synaptic weights resume the whole history of the neuronal network contained in the structure of the generating function of cumulants (the topological pressure) and in the structure of allowed/forbidden raster plots (grammar).

  Furthermore, the estimation method proposed in this thesis (i.e, the learning scheme we implemented) remains well-behaved for estimating non-stationary potentials with slow time-varying parameters. On the other hand, up to our knowledge, Ising-like related methods do not allow to treat in a straightforward way the time-evolution of the distribution although interesting results in time-varying couplings have quite recently appeared ([158]).

- **Variable Length Markov models** Ferrari and Wyner [61] uses Variable Length Markov Chains (VLMC) with a criteria based on the Kullback - Leibler divergence, to approximate the minimal context structure (history

dependence for each pattern, which means to identify the "relevant" conditional intensities of blocks in order to describe the sequence). However, up to our knowledge, this models can not take profit of Gibbs distribution framework.

- **Hidden Markov chains.** An alternative approach to our method could be Hidden Markov chains models although we don't know references for applications in the domain of spike trains analysis to which we could compare.

- **Parallel computation suitability.** Our estimation method relies on matrix computations which have intrinsical adequacy to parallel computation. In particular, one could use parallel implementations of fast Krylov-subspace algorithms (e.g, Lanczos and Arnoldi algorithms) to calculate eigenvalue and eigenvectors of large sparse matrices. On the other hand, MCMC methods can not be parallelized in straightforward manner (for details, see [155, 210]) and this is why, fast approximated techniques like mean-field and TAP equations are currently the main way to approach parameter estimation for Ising-like Potentials.

- **Performances.** At its current implementation level, the proposed method allows us to analyze the statistics of small groups (up to 8/12) of neurons. The parametric statistical potential of Markov processes up to range 16/20 is calculable, thus considering up to 2^{20} states for the process. The implementation considers several well-established numerical methods, in order to be applicable to a large set of possible data. With respect to the state of the art, this method allows us to consider non-trivial statistics (e.g. beyond rate models and even models with correlation), thus targeting models with complex spike patterns. This method is in a sense the next step after Ising models, known as being able to represent a large but limited part of the encoded information (e.g. [166, 125]). Another very important difference with respect to other current methods is that we perform the explicit variational optimization of a well defined quantity, i.e., the KL-divergence between the observed and estimated distributions. The method proposed here does not rely on Markov Chain Monte Carlo methods but on a spectral computation based on the PF matrix, providing exact formula, while the spectral characteristics are easily obtained from standard numerical methods.

  The main drawback of our method is that it *does not allow to treat a large number of neurons and simultaneously a large range*. This is due to the evident fact that the number of monomials combinatorially increases as N, R growth. However, this is not a problem intrinsic to our approach but

to parametric estimations potentials of the form (2.18). We believe that other form of potential could be more efficient (see [25] for an example). We also want to emphasize that, when considering Ising like statistics our algorithm is *less performing* than the existing ones (although improvements in speed and memory capacity thanks to the use of parallel computation algorithms remain an open and natural development path), for the simple reason that the latter has been developed and optimized using the tremendous results existing in statistical physics, for spins systems. Their extensions to models of the general form (2.18) seems rather delicate, as suggested by the nice work in [120] where extension between the 1-step Markov case is already cumbersome.

- **Mean-field methods.** Mean-field methods aim at computing the average value of observables ("order parameters") relevant for the characterization of statistical properties of the system. Typical examples are magnetization in ferromagnetic models (corresponding to rates in spiking neurons models), but more elaborated order parameters are known e.g. in spin glasses [126] or in neural networks [181]. Those quantities obey equations (usually called mean-field equations) which are, in most cases, not explicitly solvable. Therefore, approximations are proposed from the simplest (naive mean-field equations) to more complex estimations, with significant results developed in the realm of spins systems (Ising model, Sherrington-Kirckpatrick spin glass model [173]). Examples are the replica method [126], Thouless-Anderson-Palmer equations [194], the Plefka expansion [143], or more recently e.g. the Sessak-Monasson approximation [172] (for a recent review on mean-field methods see [136]). Since the seminal paper by Schneidman and collaborators [166] those techniques have also been applied to spike trains statistics analysis assuming that neurons dynamics generates a spike statistics characterized by a Gibbs distribution with an Ising Hamiltonian. In their most common form these methods do not consider dynamics (e.g time correlations) and their extension to the time-dependent case (e.g. dynamic mean-field methods) is far from being straightforward (see e.g. [182, 181, 11, 163, 60] for examples of such developments). Moreover, exact mean-field equations and their approximations usually only provide a probability measure at positive distance to the true (stationary) probability measure of the system (this distance can be quantified in the setting of information geometry using e.g. the KL distance [4]). This is the case whenever the knowledge of the sought order parameters is not sufficient to determine the underlying probability.

  The present work can, in some sense, be interpreted in the realm of mean-field approaches. Indeed, we are seeking an hidden Gibbs measure and

we have only information about the average value of ad hoc observables. Thus, equation (2.26) is a mean-field equation since it provides the average value of an observable with respect to the Gibbs distribution. There are therefore L such equations, where L is the number of monomials in the potential ψ. Are all these equations relevant ? If not, which one are sufficient to determine unequivocally the Gibbs distribution ? Which are the order parameters ? The method consisting of providing a hierarchy of mean-field approximations which starts with the Bernoulli model (all monomials but the rate terms are replaced by a constant), then Ising (all monomials but rate and spatial correlations are replaced by a constant), while progressively diminishing the KL divergence allows to answer the question of the relevant order parameters and can be interpreted as well in the realm of information geometry. This hierarchical approach is a strategy to cope with the problem of combinatorial explosion of terms in the potential when the number of neurons or range increases. But the form of potential that we consider does not allow a straightforward application of the methods inherited from statistical mechanics of spin systems. As a consequence, we believe that instead of focusing too much on these methods it should be useful to adopt techniques based on large deviations (which actually allows the rigorous foundation of dynamic mean field methods for spin-glasses [11] and neural networks [163, 60]). This is what the present formalism offers.

6.3 Limitations and their particular perspectives

At the current state of the art, the method we propose is limited by three aspects.

First of all, the formalism is developed for a stationary spike-train, i.e. for which the statistical parameters are constant. This is indeed a strong limitation, especially in order to analyze biological data, though several related approaches consider the same restrictive framework. This drawback is overcome at two levels. At the implementation level we show here how using a sliding estimation window and assuming an adiabatic, i.e. slowly varying, distribution we still can perform some relevant estimation. In a nutshell, the method seems still usable and we are now currently investigating this on both simulated and biological data, this being another study on its own. At a more theoretical level, we are revisiting the thermodynamic formalism developed here for time varying parameters (in a similar way as the so called inhomogeneous Poisson process with time varying rates). Though this yields non-trivial developments beyond the scope of this work, it seems that we can generalize the present formalism in this direction.

Secondly, the present implementation has been optimized for dense statistical distributions, i.e., in the case where almost all possible spike combinations are observed. Several mechanisms, such as look-up tables, make this implementation very fast. However, if the data is sparse, as it may be the case for several biological neural regimes, a dual implementation has to be provided using data structure, such as associative tables, well adapted to the fact that only a small amount of possible spike combinations are observed. This complementary implementation has been made available and validated against the present one. This is going to analyze sparse Markov processes up to range much higher than 16/20. Again this is not a trivial subject and this aspect must be developed in a next study as well as the applicability of parallel computing alternatives (e.g. sparse matrix storage, parallel fast-eigenvalue algorithms, etc.).

Finally, given an assembly of neurons, every statistical tools available today provide only the analysis of the statistics a small subset of neurons, and it is known that this only partially reflects the behavior of the whole population [109]. In particular among theoretical goals, we consider revisiting the analysis presented in [159] to predict bounds for large ensemble prediction by Gibbs distributions including time dependent terms.

Unfortunately, in practice, the present method is difficult to generalize to more than 8/10 neurons because of the incompressible algorithmic complexity of the formalism although parallel computation techniques might be helpful (see perspectives). However, the barrier is not at the implementation level, but at the theoretical level, since effective statistical general models (beyond Ising models) allow for instance to analyze statistically large spiking patterns such as those observed in syn-fire chains [81] or polychronism mechanisms [140]. This may be the limit of the present class of approaches, and things are to be thought differently, probably, through the analysis of non parametric potentials such as the formal BMS-model Gibbs potential [25].

6.4 Further Theoretical and numerical Perspectives

Large deviations Extending more the results of section 2.5.7 will be very interesting. Additionally we need to include the already available ones into the current implementation, since this will permit to estimate the size of fluctuations for each coefficient and hence better identify which ones of estimated parameters are reliable, and which characteristics are intrinsic to the data and which ones are apparent constituting simply effects of the finite-size fluctuations

KL divergence scaling and model comparison In order to improve model comparison, it would be interesting to understand theoretically/numerically how the KL divergence or $\tilde{h}$ scales with the number of units considered and the range (memory depth) of models. We will learn for instance how to compare spike-trains statistical description at different population/time structure levels.

Besides, both multi-information and Kullback-Leibler based comparison techniques should need to be extended in order to take account of temporal dependencies in order to be able to compare models with different memory, but we remark that this is far from straightforward. The only work we know to exists have appeared in [7].

Integrating Information theory methods for Stimulus encoding, Binning, model interpretation etc. It is very tempting to use methods of information theory for neural data, in order to unveil relevant interpretation of spike-trains statistics modeled as Gibbs distribution. In particular in order to answer stimuli encoding questions such as how models constructed with data from spontaneous activity differ from models constructed with data where responses were conditioned on a stimulus. This would provide a better insight about the biological meaning of the different potentials.

Additionally, we could use of available theoretical results on entropy and mutual information asymptotic behavior and their dependence with respect to bin-size choices for sampling data, in order to fully understand the effects of binning when estimating the parameters of the Gibbs distribution. This will hopefully provide a better interpretation of the monomials (equivalently, the constraints in Jaynes principle) associated with the largest parameter values present in the "closest" (in the Kullback-Leibler sense) Gibbs potential form.

Another related idea: in the spirit of Ince et al.[85] it would be interesting to trying to investigate how to improve entropy and mutual-Information estimates thanks to the time-correlation structure of the system, more precisely through the knowledge of the value of the Gibbs potential parameters for the best suited (closest in the KL-divergence sense) Gibbs distribution.

Grammar Grammar encodes/contains might contain key features of the system such as effects of the refractory period and the thumbprint of effective network interactions. At the current stage, not use of the grammar induced by refractory period has been included in our biological data analysis. However, a deeper investigation of how grammar constraints may affect numerical stability

Parallelization and optimal Eigenvalue solution As stated in section 2.3.3, a key aspect of Gibbs Distribution framework is that it conduces to a the change of paradigm for the parametric estimation from Markov chain sampling to matrix computations (see section (3.1.3). This enables to use parallel computation techniques that are not yet included in the current implementation and that become mandatory since the dimension of the matrices involved, increases exponentially fast with the range of the potential and the number of neurons. An second advantage in order to future optimization of the implementation is that since we are dealing with a *sparse matrix* there exist eigenvalue sparse-specific methods such as Krylov subspace iterations methods, see [52] and the detailed survey in [9].

Goodness-of-Fit test Several ready to use GOF implementations exists (see section 3.2). However in order to use them, I would be necessary to set-up a correct interface to introduce within those implementations the EnaS output. More precisely, to adequately present the estimated Gibbs distribution or vector at least a vector of marginal probabilities obtained from it. A point of attention needs to be done here since we recall that Gibbs distributions do not provide a continuous uni-variate distribution (see section 2.5.9) and whether or not it satisfies the validity criteria for the proposed test needs to be analyzed previously for each case.

Neyman-Pearson criteria This criteria presented in section 2.5.9, constitute the optimal statistical test and offers a control of confidence intervals for model discrimination given the finite-sample length T. It could be very useful to integrate it to EnaS capabilities although its implementation is cumbersome.

Variable Length Markov models An interesting path consist in the use of the VLMC method [61] which finds the minimal context structure, as a tool to improve model selection (i.e, a priori guess) within the set of full Markov chains of fixed memory.

Estimating the memory of Sequence Up to the author knowledge none of the methods discussed in section 2.5.6 are included on a freely available implementation. It would be interesting to integrate them to the EnaS library in order to pre-analyze the training sequence (the observed data) and take account of its results to help the choice of candidate models. However their implementation is not straightforward. For the first criteria, which consists in evaluating the block entropy difference, there are unavoidable

finite sample biases in every block entropy estimators which need to be corrected by some choice of bias-corrector [88]. For the second criteria, the Morvai-Weiss technique requires a maximization within the space of sequence suffixes of different length which reflects a non trivial search with computational effort to be determined.

Non-stationary higher-order Markov chain sampling In order to validate numerically our parametric estimation method for time-varying parameters of Gibbs distributions with memory (see section 2.5.10) we need to formally resolve and numerically implement the question of how to construct a Markov chain of higher-order with time-dependent transition matrix and implement.

Other numerical Perspectives Benchmarking our method with respect to other alternative methods as MCMC or GLM, and further applications of the library to other computational fields like machine learning constitute other important perspectives by their own.

6.5 Brief Conclusion

We hope to have shown clearly that the framework of thermodynamic formalism and links to Statistical Physics is a relevant source of methods for the analysis of spike train statistics. We believe that it remains a useful source concerning the challenging perspectives that might derive from this work, in addition to the overwhelming richness of ergodic theory that might not been fully considered in neuroscience. We hope that this work will attract the attention of the neuroscience community on these points.

Appendix A

Publications of the Author Arising from this Work

Description

The following paper, when be completed by including some results about the analysis on experimental data of chapter 4, will be submitted to Journal of Physiology - Paris ISSN: 0928-4257. Imprint: ELSEVIER.

,,

Entropy-based parametric estimation of spike train statistics

J. C. Vasquez [*], T. Viéville [*], B. Cessac [*] [†]

August 27, 2010

Abstract

We propose a generalisation of the existing maximum entropy models used for spike trains statistics analysis, based on the thermodynamic formalism from ergodic theory, and allowing one to take into account memory effects in dynamics. We propose a spectral method which provides directly the "free-energy" density and the Kullback-Leibler divergence between the empirical statistics and the statistical model. This method does not assume a specific Gibbs potential form. It does not require the assumption of detailed balance and offers a control of finite-size sampling effects, inherent to empirical statistics, by using large deviations results. A numerical validation of the method is proposed and the perspectives regarding spike-train code analysis are also discussed.

Keywords: *Spike train analysis , Higher-order correlation , Statistical Physics , Gibbs Distributions , Maximum Entropy*
PACS: *05.10.-a , 87.19.lo , 87.19.lj*
MCS(2000): *37D35 , 37M25 , 37A30*

1 Introduction

Processing and encoding of information in neuronal dynamics is a very active research field [59], although still much of the role of neural assemblies and their internal interactions remains unknown [55]. The simultaneously recording of the activity of groups of neurons (up to several hundreds) over a dense configuration, supplies a critical database to unravel the role of specific neural assemblies. In complement of descriptive statistics (e.g. by means of cross-correlograms or joint peri-stimulus time histograms), somehow difficult to interpret for a large number of units (review in [8, 37]), is the specific analysis of multi-units spike-patterns, as found e.g. in [1]. This approach develops algorithms to detect common patterns in a data block, as well as performing combinatorial analysis to compute the expected probability of different kind of patterns. The

[*]INRIA, 2004 Route des Lucioles, 06902 Sophia-Antipolis, France.
email: Juan-Carlos.Vasquez@sophia.inria.fr
[†]Laboratoire J. A. Dieudonné,U.M.R. C.N.R.S. N°6621, Université de Nice Sophia-Antipolis, France.

,,
main difficulty with such type of approaches is that they rely on a largely controversial assumption, Poissonian statistics (see [57, 58, 66]), which moreover, is a minimal statistical model largely depending on the belief that firing rates are essentially the main characteristic of spike trains.

A different approach has been proposed in [66]. They have shown that a model taking into account pairwise synchronizations between neurons in a small assembly (10-40 retinal ganglion cells) describes most (90%) of the correlation structure and of the mutual information of the block activity, and performs much better than a non-homogeneous Poissonian model. Analogous results were presented the same year in [73]. The model used by both teams is based on a probability distribution known as the Gibbs distribution of the Ising model which comes from statistical physics. The parameters of this distribution relating, in neural data analysis, to the firing rate of neurons and to their probability of pairwise synchronisation have to be determined from empirical data. Note that this approach has been previously presented in neuroscience, but in a slightly different and more general fashion, by [45, 39, 46] (it was referred as "log-linear models"). The use of Ising model in neural decoding (especially of visual stimuli) has been largely exploited by several other authors [19, 53, 72, 78]. In particular, it is believed by some of them [19] that the pairwise coupling terms inferred from simultaneous spikes corresponds, in the model, to effective couplings between ganglion cells. In this spirit, computing the parameters of the Ising model would provide an indirect access to ganglion cells connections. In addition, an increasing number of different theoretical and numerical developments of this idea have recently appeared. In particular, in [80], the authors propose a modified learning scheme and thanks to concepts taken from physics, such as heat capacity, explore new insights like the distribution of the underlying density of states; additionally in [61, 60] authors study and compare several approximate, but faster, estimation methods for learning the couplings and apply them on experimental and synthetic data drawing several results for this type of modeling.

However, it might be questionable whether more general form of Gibbs distributions (e.g. involving n-uplets of neurons) could improve the estimation and account for deviations to Ising-model ([72, 80]) and provide a better understanding of the neural code from the point of view of the maximal entropy principle [34]. As a matter of fact, back to 1995, [46] already considered multi-unit synchronizations and proposed several tests to understand the statistical significance of those synchronizations and the real meaning of their corresponding value in the energy expansion. A few years later, [45] generalized this approach to arbitrary spatio-temporal spike patterns and compared this method to other existing estimators of high-order correlations or Bayesian approaches. They also introduced a method comparison based on the Neyman-Pearson hypothesis test paradigm. Though the numerical implementation they have used for their approach presented strong limitations, they have applied this methods successfully to experimental data from multi-units recordings in the pre-frontal cortex, the visual cortex of behaving monkeys, and the somato-sensory cortex of anesthetized rats. Several papers have pointed out the importance of temporal patterns of activity at the network level [41, 83, 69], and recently [78] have shown the insufficiency of Ising model to predict the temporal statistics of the neural activity. As a consequence, a few authors (we know only one reference, [44]) have attempted to define time-dependent Gibbs

,,
distributions on the base of a Markovian approach (1-step time pairwise correlations) and convincingly showed a clear increase in the accuracy of the spike train statistics characterization. Namely, this model produces a lower Jensen-Shannon Divergence, when analyzing raster data generated by a Glauber spin-glass model, but also *in vivo* multineuron data from cat parietal cortex in different sleep states.

To summarize, the main advantages of all these 'Ising-like' approaches are:

- (i) to be based on a widely used principle, the maximal entropy principle [34] to determine statistics from the empirical knowledge of *(ad hoc)* observables;
- (ii) to propose statistical models having a close analogy with Gibbs distributions of magnetic systems, hence disposing of several deep theoretical results and numerical methods (Monte-Carlo methods, Mean-Field approximations, series expansions), resulting in a fast analysis of experimental data from large number of neurons.

However, as we argue in this paper, this approaches presents also, in its current state, fundamental weaknesses:

- (i) The maximal entropy principle leads, in its classical formulation, to a parametric form, corresponding to chosing a finite set of *ad hoc* constraints, which only provides an approximation of the real statistics, while the distance (say measured by the Kullback-Leibler divergence) between the model and the hidden distribution can be quite large [21]. Moreover, when considering time dependent correlations, this procedure leads to Gibbs potential which requires a proper renormalisation in order to be related to a Markov chain (see section 2.3.6).
- (ii) The Gibbs distributions considered by these approaches, with the naive form "$\frac{1}{Z}e^{-\beta H}$", where Z is a constant (while it depends on boundary terms in the general case) have a limited degree of application; in particular they do not extend easily to time dependent sequences with long memory, as spikes train emitted from neural networks might well be. Especially, considering already one time step Markov processes leads to substantial complications a shown in [44]. The "partition function" is not a constant (see eq. (1) of paper [44]) and needs to be approximated (eq. (4) of the same paper) using the (unproved) assumption of detailed balance, which is moeover a sufficient but non necessary condition for the existence of an equilibrium state, and may hardly generalize to more elaborated models.
- (iii) It does not allow to treat in a straightforward way the time-evolution of the Gibbs distribution (e.g. induced by mechanisms such as synaptic plasticity).

However, more general forms of Gibbs distributions have been introduced since long [74, 64, 5], in a theory called "thermodynamic formalism" introduced in the realm of dynamical systems and ergodic theory, allowing to treat infinite time sequences of processes with long (and even infinite [43]) memory. In this paper, we use the thermodynamic formalism to propose a generalisation of the existing models used for spike

,,
trains statistics analysis which results in a more powerful framework that overcomes some of the weaknesses mentioned above. Our results are grounded on well established theoretical basements (see e.g. [38]) completed by recent results of the authors dealing with collective spike trains statistics in neural *networks* models [12, 11]. The theoretical framework of our approach is presented in the section 2. We propose a global approach to spike train analysis, going beyond the usual approaches essentially because it allows us to take into account (long term) *memory effects* in dynamics (sections 2.1,2.2). As a matter of fact we deal with models considering *spatio-temporal* and time-*causal* structure of spike trains emitted by neural networks together with the fact that some spike sequences (or "words") might be forbidden by dynamics, introducing the notion of *grammar*. We propose a spectral method which provides directly the "free energy density" and the Kullback-Leibler divergence between the empirical statistics and the statistical model (section 2.3). This method does not assume a specific potential form and allows us to handle correctly non-normalized potentials. It does not require the assumption of detailed balance (necessary to apply Markov Chain Monte-Carlo (MCMC) methods) and offers a control of finite-size sampling effects, inherent to empirical statistics, by using large deviations results (Section 2.4). The price to pay is to introduce a somewhat heavy, but necessary, mathematical formalism. In several places we make connections with existing methods to clarify these concepts.

These theoretical basements allows us to propose, in section 3, a numerical method to parametrically estimate, and possibly compare, models for the statistics of simulated multi-cell-spike trains. Our method is not limited to firing rates models, pairwise synchronizations as [66, 73, 72] or 1-step time pairwise correlations models as [44], but deals with general form of Gibbs distributions, with parametric potentials corresponding to a spike n-uplets expansion, with multi-units and multi-times terms. The method is exact (in the sense that is does not involve heuristic minimization techniques). Moreover, we perform fast and reliable estimate of quantities such as the Kullback- Leibler divergence allowing a comparison between different models, as well as the computation of standard statistical indicators, and a further analysis about convergence rate of the empirical estimation and large deviations.

In section 4 we perform a large battery of tests allowing us to experimentally validate the method. First, we analyse the numerical precision of parameter estimation. Second, we generate synthetic data with a given statistics, and compare the estimation obtained using these data for several models. Moreover, we simulate a neural network and propose the estimation of the underlying Gibbs distribution parameters whose analytic form is known [11]. We also perform the estimation for several models using data obtained from a simulated neural network with stationary dynamics after Spike-Time dependent synaptic plasticity. Finally, we show results on the parameters estimation from synthetic data generated by a non-stationary statistical model.

,,

2 Spike trains statistics from a theoretical perspective.

2.1 General context

2.1.1 Neural network dynamics.

We consider the evolution of a network of N neurons, described by a dynamical model, that is, either a deterministic dynamical system or a stochastic dynamical system (usually governed by both a deterministic evolution map and additive noise). We assume that there is a minimal time scale δ, set to $\delta = 1$ without loss of generality, at which dynamics can be time-discretized. Typically, this can be the minimal resolution of the spike time, constrained by biophysics and by measurements methods (see [9] for a discussion on time discretisation in spiking neural networks). The typical neuron models we think of are punctual conductance based generalized Integrate-and-Fire (IF) models with exponential synapses (gIF) [13]. Actually, the formalism developed here has been rigorously funded in [11] for Leaky-Integrate-and-Fire (LIF) models with noise. We further assume the network parameters (synaptic weights, currents, etc..) to be fixed in this context (see [13] for a discussion). This means that we assume observing a period of time where the system parameters are essentially constants. In other words, we focus here on *stationary* dynamics. This restriction is further discussed in section 4.3.5.

We are interested in situations where neurons dynamics, and especially spikes occurrences, do not show any regularity or exact reproducibility and require a statistical treatment. This is obviously the case for stochastic evolutions but this also happens in the deterministic case, whenever dynamics exhibits initial conditions sensitivity. This leads us to the choice of the statistical formalism proposed here, called the "thermodynamic formalism[1] " (see [12] for an extended discussion).

2.1.2 Dynamics and raster plots.

Each neuron of index $i = 0 \ldots N-1$ is characterized by its state, X_i, which belongs to some (bounded) set $\mathscr{I} \in \mathbb{R}^M$. M is the number of variables characterizing the state of one neuron (we assume that all neurons are described by the same number of variables). A typical example is $M = 1$ where $X_i = V_i$ is the membrane potential of neuron i and $\mathscr{I} = [V_{min}, V_{max}]$ but the present formalism affords extensions to such additional characteristics as activation variables (e.g. for the Hodgkin-Huxley model [31] $M = 4$). The variable $X = [X_i]_{i=0}^{N-1}$ represents the state of a network of N neurons. Without loss of generality, we assume that all neurons have the same properties so that $X \in \mathscr{M} = \mathscr{I}^N$, where $\mathscr{M}$ is the phase space where dynamics occurs. The evolution of the network over an infinite time is characterized by a *trajectory* $\tilde{X} \stackrel{\text{def}}{=} \{X(t)\}_{t=-\infty}^{+\infty}$.

One can associate to each neuron i a variable $\omega_i(t) = 1$ if neuron i fires between $[t, t+1[$ and $\omega_i(t) = 0$ otherwise. A "spiking pattern" is a vector $\omega(t) \stackrel{\text{def}}{=} [\omega_i(t)]_{i=0}^{N} - 1$

[1] This terminology has been introduced by Sinai [74], Ruelle [64] and Bowen [5] because of its analogy with statistical physics. But it does not relies on the principles of thermodynamics. Especially, the maximization of statistical entropy, discussed below, does not requires the invocation of the second principle of thermodynamics.

,,

which tells us which neurons are firing at time t. In this setting, a "raster plot" is a sequence $\omega \stackrel{\text{def}}{=} \{\omega(t)\}_{t=-\infty}^{+\infty}$, of spiking patterns. Finally a *spike block* is a finite set of spiking pattern, written:

$$[\omega]_{t_1}^{t_2} = \{\omega(t)\}_{\{t_1 \leq t \leq t_2\}},$$

where spike times have been prescribed between the times t_1 to t_2.

To each trajectory $\tilde{X} = \{X(t)\}_{t=-\infty}^{+\infty}$ is associated a raster plot $\omega = \{\omega(t)\}_{t=-\infty}^{+\infty}$. This is the sequence of spiking patterns displayed by the neural network when it follows the trajectory $\tilde{X}$. We write $\tilde{X} \to \omega$. On the other way round, we say that an infinite sequence $\omega = \{\omega(t)\}_{t=-\infty}^{+\infty}$ is *an admissible raster plot* if dynamics allows a trajectory $\tilde{X}$ such that $\tilde{X} \to \omega$. We call Σ the set of admissible raster plots. The dynamics of the neurons state induces therefore a dynamics on the set of admissible raster plots, represented by the *left shift*, σ, such that $\sigma\omega = \omega' \Leftrightarrow \omega'(t) = \omega(t+1), \forall t \geq 0$. Thus, in some sense, raster plots provide a code for the trajectories $\tilde{X}$. Note that the correspondence may not be one-to-one [10].

Though dynamics produces many possible raster plots, it is important to remark that it is not able to produce *any* possible sequence of spiking patterns. This depends on the system properties (e.g., refractoriness forbids raster plots with spike interval below $1ms$) and parameters (e.g., after synaptic weight adaptation the dynamics often appears more constrained). For example, inhibition may prevent a neuron to fire whenever a group of pre-synaptic neurons has fired before. There are therefore *allowed* and *forbidden* sequences, constrained by dynamics. This corresponds to the following *crucial* property, often neglected in entropy estimations of spike trains [59]. The set of admissible raster plot Σ is *not the set of all possible raster plots*. Indeed, considering spike blocks of size n there are 2^{Nn} possible spike blocks but quite a bit less *admissible* raster plots (the exponential rate of growths in the number of admissible raster plots is given by the topological entropy which is an upper bound for the Kolmogorov-Sinai entropy defined in eq. (3), footnote 6).

2.1.3 Transition probabilities.

Typically, the network dynamics and the related spikes fluctuate in an unpredictable manner. The spike response itself is not sought as a deterministic response in this context, but as a conditional probability [59]. "Reading out the code" consists of inferring such probability. Especially, the probability that a neuron emits a spike at some time t depends on the history of the neural network. However, it is impossible to know explicitely its form in the general case since it depends on the past evolution of all variables determining the neural network state $\mathbf{X}$. A possible simplification is to consider that this probability depends *only* on the spikes emitted in the past by the network. In this way, we are seeking a family of transition probabilities of the form $P\left[\omega(t) \,|\, [\omega]_{-\infty}^{t-1}\right]$ from which all spike trains statistical properties can be deduced. These transition probabilities are called *conditional intensity* in [35, 7, 18, 36, 82, 51, 81, 56] and they are essential to determine completely the spike trains statistics. The price to pay is that we have to consider processes with memory (which is not so shocking when dealing with neural networks).

,,

These transition probabilities are unknown for most models but an explicit computation can be rigorously achieved in the case of a discrete time Leaky Integrate-and-Fire (LIF) neural networks with noise, in the stationary case (e.g. time independent stimulus) (see eq. (40) and [11]). Stationarity means here that the transition probability does not depend explicitely on t so that one can focus on transition probabilities of the form $P\left[\omega(0) \,|\, [\omega]_{-\infty}^{-1}\right]$ and infer the probability of any spike block by the classical Chapman-Kolmogorov equation [27]. To our knowledge this is the only example where the complete spike trains statistics can be rigorously and analytically computed. Note that the transition probability depends on a *unbounded* past in the LIF model. Indeed, the state of a neuron is reset whenever it fires, so the probability of a given spiking pattern at time 0 depends on the past up to a time when each neuron has fired at least once. However, this time cannot be bounded (though the probability that it is larger than some τ decays exponentially fast with τ) [11].

2.1.4 Gibbs distribution.

As far as the present paper is concerned, the main result in [11] states that some neural networks models *do have Gibbs distributions*, though of a quite more complex form than currently used in the literature. More precisely it is rigorously proved in [11] that in discrete-time LIF models[2] with noise the statistics of spike trains is characterized by a *Gibbs distribution* which is also an *equilibrium state*, where the potential can be *explicitely computed*, but *has infinite range*.

Let us be more explicit. Since we are using the terms "Gibbs distribution" and "equilibrium state" in a more general sense than the definition used in the neuroscience community for spike train statistics analysis, we give here the definition of these two terms. In several places in the paper we show the link between this formalism and the usual form, and explain why we need to use the present formalism for spike train analysis. The main difference is that we consider probability distributions on a set of spatio-temporal sequences where the "space" is the network, and where *time is infinite* so that the spike train probability distributions is defined on infinite time sequences [3]. This is the natural context when considering transition probabilities as introduced in the previous section. The price to pay is a more complex formulation than the classical $\frac{1}{Z}\exp(-\beta H)$, but the reward is having a formalism allowing us to handle spike trains statistics including memory terms, and an explicit way to compute the free energy density and the Kullback-Leibler divergence between the empirical statistics and a statistical model, as developped in the rest of the paper.

A probability distribution μ_ϕ on the set of *infinite* spike sequences Σ (raster plots) is a Gibbs distribution if there exists a function[4] $\phi : \Sigma \to \mathbb{R}$, called a *potential*, such

[2] Without restriction on the synaptic weights except that they are finite.

[3] This corresponds to the "thermodynamic limit" in statistical physics but in our case thermodynamic limit means "time tends to infinity" instead of "dimension of the system tends to infinity". As a matter of fact the number of neurons, N, is fixed and finite in the whole paper.

[4] Some regularity conditions, associated with a sufficiently fast decay of the potential at infinity, are also required, that we do not state here [38].

,,

that the probability of a spike block $[\omega]_t^{t+n}$, for any $-\infty < t < +\infty$, and $n > 0$, obeys:

$$c_1 \leq \frac{\mu_\phi\left([\omega]_t^{t+n}\right)}{\exp\left[-(n+1)P(\phi) + \sum_{k=t}^{t+n}\phi(\sigma^k\omega)\right]} \leq c_2, \tag{1}$$

where $P(\phi), c_1, c_2$ are some constants with $0 < c_1 \leq 1 \leq c_2$. Recall that σ is the shift on rasters defined in section 2.1.2. Basically, this expresses that, as n becomes large, $\mu_\phi\left([\omega]_t^{t+n}\right)$ behaves[5] like $\frac{\exp\left[\sum_{k=t}^{t+n}\phi(\sigma^k\omega)\right]}{\exp[(n+1)P(\phi)]}$.

An *equilibrium state* is a probability distribution μ_ϕ which satisfies the following variational principle:

$$P(\phi) \stackrel{\text{def}}{=} h(\mu_\phi) + \mu_\phi(\phi) = \sup_{\mu \in m^{(inv)}(\Sigma)} h(\mu) + \mu(\phi), \tag{2}$$

where $m^{(inv)}(\Sigma)$ is the set of invariant probability measures on Σ, $h(\mu)$ is the entropy[6] of the probability μ, and $\mu(\phi) \stackrel{\text{def}}{=} \int \phi d\mu$ is the average of ϕ with respect to the probability μ. Note that the notion of Gibbs distribution and equilibrium state are not equivalent in general [38], but in the present context, they are[11].

The term $P(\phi)$, called the *topological pressure* in this context, is the formal analog of a thermodynamic potential (free energy density). It is a generating function for the cumulants of ϕ (see section 2.2.3 for explicit examples).

2.1.5 Gibbs potential.

In the case of discrete-time LIF models the potential ϕ is the log of the probability transition $P\left[\omega(t) \,|\, [\omega]_{-\infty}^{t-1}\right]$ [11]. We believe that this statement extends to more general situations: if a spike train is characterized by a Gibbs distribution then a natural candidate for the Gibbs potential is the log of the conditional intensity. Let us insist on this result. Beyond the mathematical intrincacies grounding this statement, this choice is natural because it provides a (time) *causal* potential with *memory*. As a consequence, the statistics of spikes at a given time are causaly related to the past spikes. This corresponds to potential having a *range* that can be large. A potential has range R if $\phi([\omega]_{-\infty}^0) = \phi([\omega]_{-(R-1)}^0)$. In terms of the transition probability, this corresponds to a system with a memory depth $R-1$ (the probability that a neuron spike at time t

[5] In the sense of (1).Thus,"behaves like" does not mean "is equal to".

[6] The Kolmogorov-Sinai entropy or entropy rate of a probability μ is:

$$h[\mu] = \lim_{n \to +\infty} \frac{h^{(n)}[\mu]}{n}, \tag{3}$$

where

$$h^{(n)}[\mu] = -\sum_{\omega \in \Sigma^{(n)}} \mu\left([\omega]_0^{n-1}\right) \log \mu\left([\omega]_0^{n-1}\right), \tag{4}$$

$\Sigma^{(n)}$ being the set of admissible sequences of length n. This quantity provides the exponential rate of growth of admissible blocks having a positive probability under μ, as n growths. It is positive for chaotic system and it is zero for periodic systems.

,,

depends on the spikes emitted by the network, up to time $t-(R-1)$ back in the past[7]). Unfortunately even if the simplest known example of neural network model, the LIF, the range is (mathematically) infinite[8]. Is the situation simpler for more complex neural networks models, for real neural networks ? Fortunately, finite range approximations can be proposed, with a good control on the degree of approximation, as we now develop.

2.2 Markov approximations.

In the sequel, we make the assumption that the spike trains statistics of the system that we are observing is described by a Gibbs distribution whose potential has to be determined from empirical data.

2.2.1 Range-R potential.

It is always possible to propose Markov approximations of ϕ, even in the case where the Gibbs potential depends on spike sequences with unbounded length. This is the strategy that we now develop. We approximate the exact transition probability by a transition probability with finite memory of depth $R-1$, $P\left[\omega(0)\,|\,[\omega]_{-(R-1)}^{-1}\right]$. In this context, as shown in [11], the exact Gibbs potential can be approximated[9] by a *range-R potential* with a parametric form:

$$\psi(\omega)=\sum_{l=1}^{R}\sum_{(i_1,t_1),\ldots,(i_l,t_l)\in\mathscr{P}(N,R),}\lambda^{(l)}_{i_1,n_{i_1},\ldots,i_l,n_{i_l}}\omega_{i_1}(n_{i_1})\ldots\omega_{i_l}(n_{i_l}). \tag{5}$$

This form is nothing but a Taylor expansion of $\log(P\left[\omega(0)\,|\,[\omega]_{-(R-1)}^{-1}\right])$, where one collects all terms of form $\omega_{i_1}^{k_1}(n_{i_1})\ldots\omega_{i_l}^{k_l}(n_{i_l})$, for integer $k_1,\ldots k_l$'s, using that $\omega_i^k(n)=\omega_i(n)$, for any $k>0$ and any i,n. Here $\mathscr{P}(N,R)$ is the set of non repeated pairs of integers (i,n) with $i\in\{0,\ldots,N-1\}$ and $n\in\{0\ldots,R-1\}$.

Such form of potential is a linear combination of *monomials*. An *order-n monomial* is a product $\omega_{i_1}(t_1)\ldots\omega_{i_n}(t_n)$, where $0\leq i_1\leq i_2\leq\cdots\leq i_n\leq N-1$, $0\leq t_1\leq t_2\leq\cdots\leq t_n<\infty$ and such that there is no repeated pair (i_k,t_k), $k=1\ldots n$. The monomial $\omega_{i_1}(t_1)\ldots\omega_{i_n}(t_n)$ takes values in $\{0,1\}$ and is 1 if and only if each neuron i_l fires at time t_l, $l=1\ldots n$. On phenomenological grounds the monomial $\omega_{i_1}(t_1)\ldots\omega_{i_n}(t_n)$ corresponds to a spike n-uplet $(i_1,t_1),\ldots,(i_n,t_n)$ (neuron i_1 fires at time t_1, neuron i_2 at time t_2, etc ...).

2.2.2 Further approximations.

The potential (5) remains quite cumbersome since the number of terms in (6) explodes combinatorially as N,R growth. Equivalently, in terms of the classical Jaynes approach

[7] Hence range 1 or equivalently memory depth 0 means time independent events.

[8] Though the variation of ϕ decays exponentially fast ensuring the existence of a thermodynamic limit.

[9] In the case of LIF models the Kullback-Leibler divergence between the exact Gibbs distribution and its approximation by the potential (5) decays exponentially fast with R.

,,

where the Gibbs distribution is obtained via the maximisation of statistical entropy under constraints (see section 2.4.3), one has to fix a number of constraints that growths exponentially fast with N, R. As a consequence, one is typically lead to consider parametric forms where monomials have been removed (or, sometimes, added) in the expansion. This constitutes a coarser approximation to the exact potential, but more tractable from the numerical or empirical point of view. To alleviate notations we write, in the rest of paper, the parametric potential in the form:

$$\psi = \sum_{l=1}^{L} \lambda_l \phi_l, \tag{6}$$

where ϕ_l's are monomials. The choice of the parametric form defines what we call a "statistical model", namely a Gibbs distribution, denoted μ_ψ in the sequel, for the potential (6). The question is "how far is this distribution from the true statistics" ?

2.2.3 Examples of range-R potentials

Bernoulli potentials The easiest example of potential are range-1 potentials (memoryless) where $\psi(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0)$. The corresponding Gibbs distribution provides a statistical model where neurons are independent.

"Ising" like potentials. This type of potential has been introduced by Schneidman and collaborators in [67]. It reads, in our notations,

$$\psi(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \lambda_{ij} \omega_i(0)\omega_j(0). \tag{7}$$

The corresponding Gibbs distribution provides a statistical model where synchronous pairwise correlations between neurons are taken into account, but neither higher order spatial correlations nor other time correlations are taken into account. As a consequence, the corresponding "Markov chain" is memoryless.

Pairwise Time-Dependent-k potentials with rates (RPTD-k).

An easy generalization of (7) is:

$$\psi(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \sum_{\tau=-k}^{k} \lambda_{ij\tau} \omega_i(0)\omega_j(\tau), \tag{8}$$

called *Pairwise Time-Dependent k (RPTD-k) with Rates* potentials in the sequel.

Pairwise Time-Dependent k (PTD-k) potentials.

A variation of (8) is to avoid the explicit constraints associated to firing rates :

$$\sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \sum_{\tau=-k}^{k} \lambda_{ij\tau} \omega_i(0)\omega_j(\tau), \tag{9}$$

called **Pairwise Time-Dependent k (PTD-k)** potentials in the sequel.

,,

2.2.4 Encoding spike blocks

To each spike block of length R, $[\omega]_k^{k+R-1}$, $k \in \mathbb{Z}$, one can associate an integer:

$$w_k = \sum_{t=k}^{k+R-1} \sum_{i=0}^{N-1} 2^{i+Nt} \omega_i(t). \tag{10}$$

One has 2^{NR} such possible blocks (though some of them can be forbidden by dynamics).

We use the following notation:

$$\sigma w_k = \sum_{t=k+1}^{k+R} \sum_{i=0}^{N-1} 2^{i+Nt} \omega_i(t), \tag{11}$$

so that, w_k represents the block $[\omega]_k^{k+R-1}$ and $\sigma w_k = w_{k+1}$ represents the block $[\omega]_{k+1}^{k+R}$. In this setting a range-R potential is therefore a vector in the space $\mathscr{H} \stackrel{\text{def}}{=} \mathbb{R}^{2^{NR}}$ with components $\psi_w \stackrel{\text{def}}{=} \psi(\omega)$. This amounts to recoding spiking sequences as sequences of spike blocks of length R, associated with words w_k, taking into account the memory depth of the Markov chain.

2.3 Determining the statistical properties of a Gibbs distribution.

We now introduce the thermodynamic formalism allowing us to compute numerically the main statistical properties of a Gibbs distribution. This approach is different from a classical approach in statistical physics where one tries to compute the partition function. The present approach gives directly the topological pressure (corresponding to the free energy density in the thermodynamic limit) from which the statistical properties can be inferred.

2.3.1 The Ruelle-Perron-Frobenius operator.

Once the parametric form of the potential is given, the statistical properties of the Gibbs distribution are obtained by the Ruelle-Perron-Frobenius (RPF) operator introduced by Ruelle in [62]. In the present case this is a positive $2^{NR} \times 2^{NR}$ matrix, $L(\psi)$, with entries

$$L_{w',w}(\psi) = e^{\psi_{w'}} G_{w',w} \quad , \tag{12}$$

(while it acts on functional spaces in the infinite range case).

The matrix G is called the *grammar*. It encodes in its definition the *essential fact* that the underlying dynamics is not able to produce all possible raster plots:

$$G_{w',w} = \begin{cases} 1, & \text{if the transition } w' \to w \text{ is admissible;} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Since we are considering blocks of the form[10] $w' \sim [\omega]_k^{k+R-1} = \omega(k) \dots \omega(k+R-$

[10] Since dynamics is assumed stationnary the result actually does not depend on k.

,,

1) and $w \sim [\omega]_{k+1}^{k-R} = \omega(k+1)\dots\omega(k-R)$, the transition $w' \to w$ is legal if[11] w' and w have the spiking patterns $\omega(k+1)\dots\omega(k+R-1)$ in common. Thus, while there are 2^{NR} blocks for a network of N neurons, the matrix G has at most 2^N non zero entries on each line. As a consequence $L(\psi)$ is *sparse*.

Note also that all non zeroes entries $L_{w',w}(\psi)$ on a given line are equal. This degeneracy comes from our choice to represent ψ as a vector in $\mathscr{H}$ which is the easiest for numerical purposes. This has consequences discussed in the section 2.3.6.

2.3.2 The Ruelle-Perron-Frobenius theorem

In the present paper we make the assumption that the underlying (and hidden) dynamics is such that the $L(\psi)$ matrix is primitive, i.e. $\exists n > 0$, s.t. $\forall w, w'$ $L^n_{w',w}(\psi) > 0$. This assumption holds for Integrate-and-Fire models with noise and is likely to hold for more general neural networks models where noise renders dynamics ergodic and mixing [11]. Note, on the opposite, that if this assumption is not fulfilled there are little chances to characterize spike trains statistics with a (unique) Gibbs distribution.

Then, $L(\psi)$ obeys the Perron-Frobenius theorem[12]:

Theorem 1 *$L(\psi)$ has a unique maximal and strictly positive eigenvalue $s(\psi) = e^{P(\psi)}$ associated with a right eigenvector $\boldsymbol{b}(\psi)\rangle$ and a left eigenvector $\langle \boldsymbol{b}(\psi)$, with positive and bounded entries, such that $L(\psi)\boldsymbol{b}(\psi)\rangle = s(\psi)\boldsymbol{b}(\psi)\rangle$, $\langle \boldsymbol{b}(\psi)L(\psi) = s(\psi)\langle \boldsymbol{b}(\psi)$. Those vectors can be chosen such that $\langle \boldsymbol{b}(\psi).\boldsymbol{b}(\psi)\rangle = 1$ where . is the scalar product in $\mathscr{H}$. The remaining part of the spectrum is located in a disk in the complex plane, of radius strictly lower than $s(\psi)$. As a consequence, for all $\mathbf{v}$ in $\mathscr{H}$,*

$$\frac{1}{s(\psi)^n} L^n(\psi)\mathbf{v} \to \boldsymbol{b}(\psi)\rangle\langle \boldsymbol{b}(\psi).\mathbf{v}, \tag{14}$$

as $n \to \infty$.

The Gibbs-probability of a spike block w of length R is

$$\mu_\psi(w) = b_w(\psi)\rangle\langle b_w(\psi), \tag{15}$$

where $b_w(\psi)\rangle$ is the w-th component of $\boldsymbol{b}(\psi)\rangle$.

As a consequence, the assumption of primitivity guarantees the existence and uniqueness of a Gibbs distribution. Note that it is more general than the detailed balance assumption.

2.3.3 Computing averages of monomials

Since $\mu_\psi[\phi_l] = \sum_w \mu_\psi[w]\phi_l(w)$ one obtains using (15):

$$\mu_\psi[\phi_l] = \sum_{w \in \mathscr{H}} b_w(\psi)\rangle\phi_l(w)\langle b_w(\psi). \tag{16}$$

This provides a fast way to compute $\mu_\psi[\phi_l]$.

[11] Additional transitions are usually forbidden by dynamics. As a consequence, those transitions have a zero probability of occurence and they can be detected on empirical sequences (see section 3.4.1).

[12] This theorem has been generalized by Ruelle to infinite range potentials under some regularity conditions [63, 64].

,,

2.3.4 The topological pressure.

The RPF theorem gives a direct access to the topological pressure $P(\psi)$ which is the logarithm of the leading eigenvalue $s(\psi)$, easily obtained by a power method (see eq. (14)). In the case of range-R potentials (6) where the topological pressure $P(\psi)$ becomes a function of the parameters $\lambda = (\lambda_l)_{l=1}^{L}$, we write $P(\lambda)$. One can show that the topological pressure is the generating function for the cumulants of the monomials ϕ_l:

$$\frac{\partial P(\lambda)}{\partial \lambda_l} = \mu_\psi[\phi_l]. \tag{17}$$

Higher order cumulants are obtained likewise by successive derivations. Especially, second order moments related to the central limit theorem obeyed by Gibbs distributions [5, 38] are obtained by second order derivatives. As a consequence of this last property, the topological pressure's Hessian is positive and the topological pressure is *convex* with respect to λ.

2.3.5 Entropy

Since μ_ψ is a Gibbs distribution, for the potential ψ, therefore, an exact expression for the Kolmogorov-Sinai entropy (3) can be readily obtained:

$$h\left[\mu_\psi\right] = P(\lambda) - \sum_l \lambda_l \mu_\psi\left[\phi_l\right]. \tag{18}$$

2.3.6 Normalisation

When switching from the potential (5), which is the polynomial expansion of the log of the conditional intensity, to a simplified parametric form (6), one introduces several biases. First, one may add terms which are not in the original potential. Second, (5) must satisfy a constraint corresponding to the fact that $\psi(\omega)$ is the log of a probability. Such a potential is called *normalized*. Its main characteristics are (i) the topological pressure is zero; (ii) the right eigenvector $\mathbf{b}(\psi)\rangle$ has all its components equal. The reason is simple: when the potential is the log of a transition probability the RPF operator satisfies $\sum_{w\in\mathscr{H}} L_{w'w} = \sum_{\omega(0)\in\{0,1\}^N} P\left[\omega(0)\,|\,[\omega]_{-(R-1)}^{-1}\right] = 1, \forall w' \in \mathscr{H}$, where w' corresponds e.g. to $[\omega]_{-R}^{-1}$, and w to $[\omega]_{-(R-1)}^{0}$. Thus, the largest eigenvalue $s(\psi)$ is 1, and the corresponding right eigenvector has all its components equal.

On the opposite, the parametric form (6) where λ_l are free parameters is in general *not normalized*, with deep consequences discussed in the next sections. However, there exists a transformation allowing to convert an arbitrary range-R potential to a normalized potential Ψ by the transformation:

$$\Psi_{w'w} = \psi_{w'} - \log(b_{w'}(\psi)\rangle) + \log(b_w(\psi)\rangle) - P(\psi). \tag{19}$$

Let us give two examples. First, if ψ is normalized then $b_{w'}(\psi)\rangle = b_w(\psi)\rangle$ and $P(\psi) = 0$ so that (fortunately) $\Psi = \psi$. Second, if ψ has range-1 then, according to the computations done in section 2.3.8, $\Psi_{w'w} = -\log(Z) + \psi_{w'} = -\log(Z) + \psi(\omega(0))$.

,,

Here, the normalisation only consists of removing the log of a (constant) partition function.

In the general case, the potential (19) has range-$R+1$. The corresponding RPF operator $L(\Psi)$, is therefore the transition matrix for a R-step Markov chain. Thus, switching to a parametric form (6) without constraint on the λ_l's we end up with a redundant transition probability of form $P'(\omega(0)\,|\,[\omega]_{-R}^{-1})$ while the right transition probability is $P(\omega(0)\,|\,[\omega]_{-(R-1)}^{-1})$. Since, obviously $P'(\omega(0)\,|\,[\omega]_{-R}^{-1}) = P(\omega(0)\,|\,[\omega]_{-(R-1)}^{-1})$ the final form of the normalized potential can be easily simplified.

2.3.7 Probability of arbitrary spike blocks

Using the normalized potential (19) the probability of a *admissible* spike block of size strictly larger than R, $[\omega]_t^{t+n+R-1}$, $t \in \mathbb{Z}$, $n > 0$ is given by:

$$\mu_\psi\left[[\omega]_t^{t+n+R-1}\right] = \mu_\psi\left[\omega(t), \omega(t+1)\ldots\omega(t+n+R-1)\right] = \mu_\psi\left[w_t, \ldots, w_{t+n}\right],$$

where the word w_k encodes the block $[\omega]_k^{k+R-1}$. As a consequence,

$$\mu_\psi\left[w_t, \ldots, w_{t+n}\right] = \mu_\psi[w_t] L_{w_t, w_{t+1}}(\Psi) L_{w_{t+1}, w_{t+2}}(\Psi)\ldots L_{w_{t+n-1}, w_{t+n}}(\Psi). \tag{20}$$

This is the classical Chapman-Kolmogorov equation. Returning to the initial (non-normalised) potential (6) this relation reads, using (19):

$$\mu_\psi\left[w_t, \ldots, w_{t+n}\right] = \mu_\psi[w_t] \frac{b_{w_{t+n}}(\psi)\rangle}{b_{w_t}(\psi)\rangle} \frac{1}{e^{(n+1)P(\psi)}} e^{\sum_{k=t}^{t+n}\psi_{w_k}}. \tag{21}$$

One checks[13] that $\mu_\psi\left[w_t, \ldots, w_{t+n}\right]$ satisfies the definition of a Gibbs distribution (1) with $P(\psi) = \log s(\psi)$ and $w_k = \sigma^k w_0$.

On the opposite, for blocks of size $0 < n < R+1$ then

$$\mu_\psi\left[[\omega]_t^{t+n}\right] = \sum_{w \subset [\omega]_t^{t+n}} \mu_\psi(w),$$

where the sum holds on each word w containing the block $[\omega]_t^{t+n}$.

2.3.8 Links with the simple Gibbs form.

In this section we make the link between our formalism and previous approaches using the simple Gibbs formulation.

As a preliminary remark note that the Gibbs-probability of a spike block w of length R, given by (15), *hasn't* the form $\frac{1}{Z}e^{\psi(w)}$, with Z constant, *except when* $R = 1$. The case $R = 1$ corresponds to a Markov chain without memory, where therefore the spiking pattern $w_t = \omega(t)$ is independent on $w_{t-1} = \omega(t-1)$. Examples are the Bernoulli model (where moreover spikes are spatially independent) or the Ising model (where spikes are spatially correlated but not time correlated). In this case, all transitions

[13] Taking into account the fact that symbols w_k encode spike blocks of length R.

,,

are allowed, thus the RPF matrix reads $L_{w',w}(\psi) = e^{\psi_{w'}}$ and does not depend on w. As a consequence, all lines are linearly dependent which implies that there are $N-1$ 0-eigenvalues while the largest eigenvalue is $Z \stackrel{\text{def}}{=} s(\psi) = Tr(L(\psi)) = \sum_{w' \in \mathscr{H}} e^{\psi_{w'}}$. The corresponding left eigenvector is $\langle \mathbf{b}(\psi) = (1,\ldots,1)$ and the right eigenvector is $b_{w'}(\psi)\rangle = \frac{e^{\psi_{w'}}}{Z}$, so that $\langle \mathbf{b}(\psi).\mathbf{b}(\psi)\rangle = 1$. Thus, the Gibbs distribution is, according to (15), $\mu_{w'} = \frac{e^{\psi_{w'}}}{Z}$.

Consider now larger ranges. Recall first that a potential of form (6) is in general not normalized. To associate it to a Markov chain one has to use the transformation (19) and the probability of a spiking pattern sequence is given by (21). In particular, the joint probability of two admissible successive blocks w',w reads $\mu_\psi(w,w') = \mu_\psi[w']\frac{b_w(\psi)\rangle}{b_{w'}(\psi)\rangle}\frac{1}{e^{P(\psi)}}e^{\psi_{w'}}$. One can introduce a formal Hamiltonian $H_{ww'} = \psi_{w'} + \log(b_w(\psi)\rangle)$ and a "conditional" partition function $Z(w') = e^{P(\psi)}b_{w'}(\psi)\rangle$ such that $\mu_\psi(w|w') = \frac{1}{Z(w')}e^{H_{ww'}}$ with $Z(w') = \sum_{w\in\mathscr{H}} e^{H_{ww'}}$ but here the partition function depends on w' (compare with eq. (1) in ref [44]). This corresponds, in statistical mechanics, to have interactions with a boundary. In this setting, the free energy density (topological pressure) is obtained (away from phase transitions[14]), via $\frac{1}{n}\log Z_n(w') \to P(\psi)$ as $n\to\infty$, $\forall w'$, requiring to consider a thermodynamic limit, as we do in the present setting.

As a conclusion, starting from an a priori form of a potential (6), obtained e.g. by Jaynes argument (see section 2.4.3) one obtains a non normalized potential which cannot be directly associated with a Markov chain, and the corresponding Gibbs measure hasn't the simple Gibbs form used for Ising model, as soon as one introduces memory terms in the potential. However, the thermodynamic formalism allows one to treat this case without approximations, or assumptions such as detailed balance, and gives direct access to the topological pressure.

2.3.9 Comparing several Gibbs statistical models.

The choice of a potential (6), i.e. the choice of a set of observables, fixes a statistical model for the statistics of spike trains. Clearly, there are many choices of potentials and one needs to propose a criterion to compare them. The Kullback-Leibler divergence,

$$d(\nu,\mu) = \limsup_{n\to\infty} \frac{1}{n} \sum_{[\omega]_0^{n-1}} \nu\left([\omega]_0^{n-1}\right) \log\left[\frac{\nu\left([\omega]_0^{n-1}\right)}{\mu\left([\omega]_0^{n-1}\right)}\right], \tag{22}$$

where ν and μ are two invariant probability measures, provides some notion of asymmetric "distance" between μ and ν.

The computation of $d(\nu,\mu)$ is delicate but, in the present context, the following holds. For ν an invariant measure and μ_ψ a Gibbs measure with a potential ψ, both defined on the same set of sequences Σ, one has [5, 63, 38, 16]:

$$d\left(\nu,\mu_\psi\right) = P(\psi) - \nu(\psi) - h(\nu). \tag{23}$$

[14] This requires a sufficiently fast decay of the potential, as mentioned in the footnote 4

,,

This is the key of the algorithm that we have developed.

2.4 Computing the Gibbs distribution from empirical data.

2.4.1 Empirical Averaging

Assume now that we observe the spike trains generated by the neural network. We want to extract from these observations informations about the set of monomials ϕ_l constituting the potential and the corresponding coefficients λ_l.

Typically, one observes, from $\mathcal{N}$ repetitions of the same experiment, i.e. submitting the system to the same conditions, $\mathcal{N}$ raster plots $\omega^{(m)}, m = 1 \dots \mathcal{N}$ on a finite time horizon of length T. These are the basic data from which we want to extrapolate the Gibbs distribution. The key object for this is the *empirical* measure. For a fixed $\mathcal{N}$ (number of observations) and a fixed T (time length of the observed spike train), the *empirical average* of a function $f : \Sigma \to \mathbb{R}$ is:

$$\overline{f}^{(\mathcal{N},T)} = \frac{1}{\mathcal{N}T} \sum_{m=1}^{\mathcal{N}} \sum_{t=1}^{T} f(\sigma^t \omega^{(m)}). \tag{24}$$

Typical examples are $f(\omega) = \omega_i(0)$ in which case the empirical average of f is the firing rate[15] of neuron i; $f(\omega) = \omega_i(0)\omega_j(0)$ then the empirical average of f measures the estimated probability of spike coincidence for neuron j and i; $f(\omega) = \omega_i(\tau)\omega_j(0)$ then the empirical average of f measures the estimated probability of the event "neuron j fires and neuron i fires τ time step later" (or sooner according to the sign of τ).

Note that in (24) we have used the shift σ^t for the time evolution of the raster plot. This notation is compact and well adapted to the next developments than the classical formula, reading, e.g., for firing rates $\frac{1}{\mathcal{N}T} \sum_{m=1}^{\mathcal{N}} \sum_{t=1}^{T} f(\omega^{(m)}(t))$.

The empirical measure is the probability distribution $\pi^{(T)}$ such that, for any function[16] $f : \Sigma \to \mathbb{R}$,

$$\pi^{(T)}(f) = \overline{f}^{(\mathcal{N},T)}. \tag{25}$$

Equivalently, the empirical probability of a spike block $[\omega]_{t_1}^{t_2}$ is given by:

$$\pi^{(T)}\left([\omega]_{t_1}^{t_2}\right) = \frac{1}{\mathcal{N}T} \sum_{m=1}^{\mathcal{N}} \sum_{t=1}^{T} \chi_{[\omega]_{t_1}^{t_2}}(\sigma^t \omega^{(m)}), \tag{26}$$

where $\chi_{[\omega]_{t_1}^{t_2}}$ is the indicatrix function of the block $[\omega]_{t_1}^{t_2}$ so that $\sum_{t=1}^{T} \chi_{[\omega]_{t_1}^{t_2}}(\sigma^t \omega^{(m)})$ simply counts the number of occurences of the block $[\omega]_{t_1}^{t_2}$ in the empirical raster $\omega^{(m)}$.

2.4.2 Estimating the potential from empirical average

The empirical measure is what we get from experiments while it is assumed that spike statistics is governed by an hidden Gibbs distribution μ_ψ that we want to determine or approximate. Clearly there are infinitely many *a priori* choices for this distribution,

[15] Recall that we assume dynamics is stationary so rates do not depend on time.

[16] In fact, it is sufficient here to consider monomials.

,,

corresponding to infinitely many a priori choices for the potential ψ. However, the ergodic theorem (the law of large number) states that $\pi^{(T)} \to \mu_\psi$ as $T \to \infty$ where μ_ψ is the sought Gibbs distribution. Equivalently, the Kullback-Leibler divergence $d\left(\pi^{(T)}, \mu_\psi\right)$ between the empirical measure and the sought Gibbs distribution *tends to* 0 as $T \to \infty$.

Since we are dealing with finite samples the best that we can expect is to find a Gibbs distribution which *minimizes* this divergence. This is the core of our approach. Indeed, using eq. (23) we use the approximation[17]:

$$d(\pi^{(T)}, \mu_\psi) = P(\psi) - \pi^{(T)}(\psi) - h(\pi^{(T)}). \tag{27}$$

The advantage is that this quantity can be numerically estimated, since for a given choice of ψ the topological pressure is known from the Ruelle-Perron-Frobenius theorem, while $\pi^{(T)}(\psi)$ is directly computable. Since $\pi^{(T)}$ is fixed by the experimental raster plot, $h(\pi^{(T)})$ is independent of the Gibbs potential, so we can equivalently minimize:

$$\tilde{h}[\psi] = P[\psi] - \pi^{(T)}(\psi), \tag{28}$$

without computing the entropy $h(\pi^{(T)})$.

This relation holds for any potential. In the case of a parametric potential of the form (6) we have to minimize

$$\tilde{h}[\lambda] = P[\lambda] - \sum_{l=1}^{L} \lambda_l \pi^{(T)}(\phi_l). \tag{29}$$

Thus, from (17) and (24), given the parametric form, the set of λ_l's minimizing the KL divergence are given by:

$$\mu_\psi[\phi_l] = \pi^{(T)}(\phi_l), \qquad l = 1 \ldots L. \tag{30}$$

Before showing why this necessary condition is also sufficient, we want to comment this result in connection with standard approaches ("Jaynes argument").

2.4.3 Inferring statistics from empirical averages of observables: The Jaynes argument.

The conditions (30) impose constraints on the sought Gibbs distribution. In view of the variational principle (2) the minimization of KL divergence *for a prescribed parametric form of the Gibbs potential* is equivalent to *maximizing the statistical entropy under the constraints (30)*, where the λ_l's appear as adjustable Lagrange multipliers. This is the Jaynes argument [34] commonly used to introduce Gibbs distributions in statistical physics textbooks, and also used in the fundating paper of Schneidman et al. [67]. There is however an important subblety that we want to outline. The Jaynes argument provides the Gibbs distribution which minimizes the KL divergence with respect to the empirical distribution *in a specific class of Gibbs potentials*. Given a parametric form for the potential it gives the set of λ_l's which minimizes the KL divergence for the set

[17] This is an approximation because $\pi^{(T)}$ is not invariant [38]. It becomes exact as $T \to +\infty$.

,,

of Gibbs measures having *this form of potential* [21]. Nevertheless, the divergence can still be quite large and the corresponding parametric form can provide a poor approximation of the sought measure. So, in principle one has to minimize the KL divergence with respect to several parametric forms. This is a way to compare the statistical models. The best one is the one which minimizes (29), i.e. knowing if the " model"ψ_2 is significantly "better" than ψ_1, reduces to verifying:

$$\tilde{h}[\psi_2] \ll \tilde{h}[\psi_1], \tag{31}$$

easily computable at the implementation level, as developed below. Note that $\tilde{h}$ has the dimension of entropy. Since we compare entropies, which units are bits of information, defined in base 2, the previous comparison units is well-defined.

2.4.4 Convexity.

The topological pressure is convex with respect to λ. As being the positive sum of two (non strictly) convex criteria $P[\psi]$ and $-\pi^{(T)}(\psi)$ in (29), the minimized criterion is convex. This means that the previous minimization method intrinsically converges towards a global minimum.

Let us now consider the estimation of an hidden potential $\psi = \sum_{l=1}^{L} \lambda_l \phi_l$ by a test potential $\psi^{(test)} = \sum_{l=1}^{L^{(test)}} \lambda_l^{(test)} \phi_l^{(test)}$. As a consequence, we estimate ψ with a set of parameters $\lambda_l^{(test)}$, and the criterion (29) is minimized with respect to *those parameters* $\lambda_l^{(test)}$, $l = 1 \ldots L^{(test)}$.

Several situations are possible. First, ψ and $\psi^{(test)}$ have the same set of monomials, only the λ_l's must be determined. Then, the unique minimum is reached for $\lambda_l^{(test)} = \lambda_l, l = 1 \ldots L$. Second, $\psi^{(test)}$ contains all the monomials of ψ plus additional ones (*overestimation*). Then, the $\lambda_l^{(test)}$'s corresponding to monomials in ψ converge to λ_l while the coefficients corresponding to additional monomials converge to 0. The third case corresponds to *underestimation*. $\psi^{(test)}$ contains less monomials than ψ or distinct monomials. In this case, there is still a minimum for the criterion (29), but it provides a statistical model (a Gibbs distribution) at *positive KL distance* from the correct potential [21]. In this case adding monomials to $\psi^{(test)}$ will improve the estimation. More precisely, if for a first test potential the coefficients obtained after minimisation of $\tilde{h}$ are $\lambda_l^{(test)}, l = 1 \ldots L^{(test)}$ and for a second test potential they are $\lambda_l^{'(test)}, l = 1 \ldots L^{'(test)}, L^{'(test)} > L^{(test)}$ then $\tilde{h}(\lambda_1^{(test)}, \ldots, \lambda_{L^{(test)}}^{(test)}) \geq \tilde{h}(\lambda_1^{'(test)}, \ldots, \lambda_{L^{'(test)}}^{(test)})$. For the same l the coefficients $\lambda_l^{(test)}$ and $\lambda_l^{'(test)}$ can be quite different.

Note that these different situations are not inherent to our procedure, but to the principle of finding an hidden probability by maximizing the statistical entropy under constraints, when the full set of constraints is not known[18]. Examples of these cases are provided in section 4. As a matter of fact, we have therefore two strategies to

[18]The problem of estimating the memory order of the underlying markov chain to a given sequence, which means, in our framework, to find the the potential range, has been a well known difficult question in coding and information theory [49]. Some of the current available tests might offer additional algorithmic tools that would be explored in a forthcoming paper

,,

estimate an hidden potential. Either starting from a minimal form of test potential (e.g. Bernoulli) and adding successive monomials (e.g. based on heuristical arguments such as "pairwise correlations do matter") to reduce the value of $\tilde{h}$. The advantage is to start from potentials with a few number coefficients, but where the knowledge of the coefficients at a given step cannot be used at the next step, and where one has no idea on "how far" we are from the right measure. The other strategy consists of starting from the largest possible potential with range R [19] . In this case it is guarantee that the test potential is at the minimal distance from the sought one, in the set of range-R potential, while the minimization will remove irrelevant monomials (their coefficient vanishes in the estimation). The drawback is that one has to start from a very huge number of monomials (2^{NR}) which reduces the number of situations one can numerically handle. These two approaches are used in section 4.

2.4.5 Finite sample effects and large deviations.

Note that the estimations crucially depend on T. This is a central problem, not inherent to our approach but to all statistical methods where one tries to extract statistical properties from finite empirical sample. Since T can be small in practical experiments, this problem can be circumvented by using an average over several samples (see eq. (24) and related comments). Nevertheless it is important to have an estimation of finite sampling effects, which can be addressed by the large deviations properties of Gibbs distributions.

For each observable ϕ_l, $l = 1 \dots L$, the following holds, as $T \to +\infty$ [22]:

$$\mu_\psi \left\{ \omega, |\pi^{(T)}(\phi_l) - \mu_\psi(\phi_l)| \geq \varepsilon \right\} \sim \exp(-T I_l(\varepsilon)), \tag{32}$$

where $I_l(x) = \sup_{\lambda_l \in \mathbb{R}} (\lambda_l x - P[\lambda])$, is the Legendre transform of the pressure $P[\lambda]$.

This result provides the convergence rate with respect to T. This is very important, since, once the Gibbs distribution is known, one can infer the length T of the time windows over which averages must be performed in order to obtain reliable statistics. This is of particular importance when applying statistical tests such as Neymann-Pearson for which large deviations results are available in the case of Markov chains and Gibbs distributions with finite range potentials [50].

Another important large deviation property also results from thermodynamic formalism [38, 14, 22]. Assume that the experimental raster plot ω is distributed according to the Gibbs distribution μ_ψ, with potential ψ, and assume that we propose, as a statistical model, a Gibbs distribution with potential $\psi' \neq \psi$. The Gibbs measure of spike blocks of range (15) is a vector in $\mathscr{H}$ and $\pi^{(T)} = \left(\pi^{(T)}(w)\right]\Big)_{w \in \mathscr{H}}$ is a random vector. Now, the probability $\mu_\psi \left\{ \|\pi^{(T)} - \mu_{\psi'}\| < \varepsilon \right\}$ that $\pi^{(T)}$ is ε-close to the "wrong" probability $\mu_{\psi'}$ decays exponentially fast as,

$$\mu_\psi \left\{ \|\pi^{(T)} - \mu_{\psi'}\| < \varepsilon \right\} \sim \exp(-T \inf_{\mu, \|\mu - \mu_{\psi'}\| < \varepsilon} d(\mu, \mu_\psi))). \tag{33}$$

[19] ibid.

,,

Thus, this probability decreases exponentially fast with T, with a rate given (for small ε) by $T\,d(\mu_\psi, \mu_{\psi'})$. Therefore, a difference of η in the Kullback-Leibler divergences $d(\pi^{(T)}, \mu_\psi)$, $d(\pi^{(T)}, \mu_{\psi'})$ leads to a ratio $\frac{\mu_\psi\left\{\|\pi^{(T)}-\mu_\psi\|<\varepsilon\right\}}{\mu_\psi\left\{\|\pi^{(T)}-\mu_{\psi'}\|<\varepsilon\right\}}$ of order $\exp -T\eta$. Consequently, for $T \sim 10^8$ a divergence of order $\eta = 10^{-7}$ leads to a ratio of order $\exp(-10)$. Illustrations of this are given in section 4.

2.4.6 Other statistics related to Gibbs distributions.

The K-L divergence minimization can be completed with other standard criteria for which some analytical results are available in the realm of Gibbs distributions and thermodynamic formalism. Fluctuations of monomial averages about their mean are Gaussian, since Gibbs distribution obey a central limit theorem with a variance controlled by the second derivative of $P(\lambda)$. Then, using a χ^2 test seems natural. Examples are given in section 4. In order to compare the goodness-of-fit (GOF) for probability distributions of spike blocks, we propose at the descriptive level the box plots tests. On the other hand, quantitative methods to establish GOF are numerous and can be classified in families of 'test Statistics', the most important being the Power-Divergence methods (eg. Pearson-χ^2 test), the Generalized Kolmogorov-Smirnov (KS) tests (eg. the KS and the Watson-Darling test) and the Phi-Divergence methods (eg. Cramer-von Mises test)[20, 17]. Finally, to discriminate 2 Gibbs measures one can use the Neyman-Pearson criteria since large deviations results for the Neyman-Pearson risk are available in this case [50]. In the present paper we have limited our analysis to the most standard tests (diagonal representations, box plots, χ^2).

3 Application: parametric statistic estimation.

Let us now discuss how the previous piece of theory allows us to estimate, at a very general level, parametric statistics of spike trains.

We observe N neurons during a stationary period of observation T, assuming that statistics is characterized by an unknown Gibbs potential of range R. The algorithmic[20] procedure proposed here decomposes in three steps:

1. *Choosing a statistical model*, i.e. fixing the potential (6) (equivalently, the relevant monomials or "observables").
2. *Computing the empirical average of observables*, i.e. determine them from the raster, using eq. (24).
3. *Performing the parametric estimation*, i.e. use a variational approach to determine the Gibbs potential.

Let us describe and discuss these three steps, and then discuss the design choices.

[20] The code is available at `http://enas.gforge.inria.fr/classGibbsPotential.html`

,,

3.1 Choosing a model: rate, coincidence, spiking pattern and more.

3.1.1 The meaning of monomials.

In order to understand the power of representation of the proposed formalism, let us start reviewing a few elements discussed at a more theoretical level in the previous section.

We start with a potential limited to a unique monomial.

- If $\psi = \omega_i(0)$, its related average value measures the firing probability or *firing rate* of neuron i;
- If $\psi(\omega) = \omega_i(0)\,\omega_j(0)$, we now measure the probability of spikes coincidence for neuron j and i, as pointed out at the biological level by, e.g, [28] and developed by [66];
- If $\psi(\omega) = \omega_i(\tau)\,\omega_j(0)$, we measure the probability of the event "neuron j fires and neuron i fires τ time step later" (or sooner according to the sign of τ); in this case the average value provides[21] the *cross-correlation* for a delay τ and the auto-correlation for $i = j$;
- A step further, if, say, $\psi(\omega) = \omega_i(0)\,\omega_j(0)\,\omega_j(1)$, we now take into account triplets of spikes in a specific pattern (i.e. one spike from neuron i coinciding with two successive spikes from neuron j);

These examples illustrate the notion of "design choice": the first step of the method being to choose the "question to ask", i.e. what is to be observed over the data. In this framework, this translates in: "choosing the form of the potential". Let us enumerate a few important examples.

3.1.2 Taking only rate or synchronization into account: Bernoulli and Ising potentials.

Rate potential are range-1 potentials, as defined before. Such models are not very interesting as such, but have two applications: they are used to calibrate and study some numerical properties of the present methods, and they are also used to compare the obtained conditional entropy with more sophisticated models.

Ising potentials have been introduced by Schneidman and collaborators in [67], taking rate and synchronization of neurons pairs, as studied in, e.g. [28]. This form is justified by the authors using the Jaynes argument.

Let us now consider potentials not yet studied, up to our best knowledge, in the present literature.

3.1.3 Taking rate and correlations into account: RPTD-k potentials.

This is a key example for the present study. On one hand, the present algorithmic was developed to take not only Bernoulli or Ising-like potential into account, but a

[21] Substracting the firing rates of i and j.

,,

large class of statistical model, including a *general second order model* (redundant monomial being eliminated), i.e. taking rate, *auto-correlation* (parametrized by $\lambda_{i\tau}$) and *cross-correlation* (parametrized by $\lambda_{ij\tau}$) into account.

Being able to consider such type of model is an important challenge, because it provides a tool to analyze not only synchronization between neurons, but more general temporal relations (see e.g. [23, 28, 6] for important applications).

Let us now turn to a specific example related to the neuronal network dynamic analysis.

3.1.4 Taking plasticity into account: "STDP" potentials

In [12] we considered Integrate-and-Fire neural networks with Spike-Time Dependent Plasticity of type:

$$W'_{ij} = \varepsilon \left[r_d W_{ij} + \frac{1}{T} \sum_{t=T_s}^{T+T_s} \omega_j(t) \sum_{u=-T_s}^{T_s} f(u)\, \omega_i(t+u) \right], \tag{34}$$

where W_{ij} is the synaptic weight from neuron j to neuron i, $-1 < r_d < 0$ a term corresponding to passive LTD, T a large time, corresponding to averaging spike activity for the synaptic weights update, and,

$$f(x) = \begin{cases} A_- e^{\frac{x}{\tau_-}}, & x < 0, \quad A_- < 0; \\ A_+ e^{-\frac{x}{\tau_+}}, & x > 0, \quad A_+ > 0; \\ 0, & x = 0; \end{cases}$$

with $A_- < 0$ and $A_+ > 0$, is the STDP function as derived by Bi and Poo [4]. The shape of f has been obtained from statistical extrapolations of experimental data. $T_s \stackrel{\text{def}}{=} 2\max(\tau_+, \tau_-)$ is a characteristic time scale. We argued that this synaptic weight adaptation rule produces, when it has converged, spike trains distributed according to a Gibbs distribution with potential:

$$\psi(\omega) = \sum_{i=0}^{N} \lambda_i^{(1)} \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \lambda_{ij}^{(2)} \sum_{u=-T_s}^{T_s} f(u) \omega_i(0) \omega_j(u). \tag{35}$$

When considering a large number of neurons, it becomes difficult to compute and check numerically this joint probability over the whole population. Here, we propose to consider a subset $\mathscr{P}_s$ of $N_s < N$ neurons. In this case, the effects of the rest of the population can be written as a bulk term modulating the individual firing rates and correlations of the observed population, leading to a marginal potential of the form:

$$\psi_{\mathscr{P}_s}(\omega) = \sum_{i \in \mathscr{P}_s} \lambda_i^{(1)'} \omega_i(0) + \sum_{i,j \in \mathscr{P}_s 0} \sum_{j=0}^{N-1} \lambda_{ij}^{(2)} \sum_{u=-T_s}^{T_s} f(u) \omega_i(0) \omega_j(u). \tag{36}$$

Here, the potential is a function of both past and future. A simple way to embed this potential in our framework, is to shift the time by an amount of T_s, using the stationarity assumption.

,,

3.1.5 The general case: Typical number of observed neurons and statistics range.

The previous piece of theory allows us to take any statistics of range R, among any set of N neurons into account. At the numerical level, the situation is not that simple, since it appears, as detailed in the two next sections, that both the memory storage and computation load are in $O(2^{NR})$, except if the grammar is very restrictive, and the possible spike pattern blocks very sparse. Hopefully, we are going to see that estimation algorithms are rather efficient, thus do not lead to a complexity larger than $O(2^{NR})$.

It is clear that the present limitation is *intrinsic* to the problem, since we have *at least*, for a statistics of range R, to count the number of occurrences of blocks of N neurons of size R, and there are (at most) 2^{NR} of them. Fastest implementations must be based on the *partial* observation of only a subset of, e.g., the most preeminent occurrences.

Quantitatively, we consider "small" values of N and R, typically a number of neurons equal to $N \in \{1, \simeq 8\}$, and Markov chain of range $R = \{1, \simeq 16\}$, in order to manipulate quantities of dimension $N \leq 8$, and $R \leq 16$, and such that $N(R+1) \leq 18$.

3.2 Computing the empirical measure: prefix-tree construction.

For one sample ($\mathcal{N} = 1$),the empirical probability (25) of the block $[\omega]_{-D}^{t}$, $-D < t \leq 0$ is given by

$$\pi^{(T)}([\omega]_{-D}^{t}) = \frac{\#[\omega]_{-D}^{t}}{T}.$$

thus obtained counting the number of occurrence $\#[\omega]_{-D}^{t}, -D < t \leq 0$ of the block $[\omega]_{-D}^{t}$ in the sequence $[\omega]_{-T}^{0}$. Since we assume that dynamics is stationary we have, $\pi^{(T)}([\omega]_{-D}^{t}) = \pi^{(T)}([\omega]_{0}^{t+D})$.

We observe that the data structure size has to be of order $O(2^{NR})$ (lower if the distribution is sparse), but does not depends on T. Since many distributions are sparse (not all blocks occur, because the distribution is constrained by a grammar), it is important to use a sparse data structure, without storing explicitly blocks of occurence zero.

Furthermore, we have to study the distribution at several ranges R and it is important to be able to factorize these operations. This means counting in one pass, and in a unique data structure, block occurrences of different ranges.

The chosen data structure is a tree of depth $R+1$ and degree 2^N. The nodes at depth D count the number of occurrences of each block $[\omega]_{-D+t}^{t}$, of length up to $D \leq R+1$[22]. It is known (see, e.g., [29] for a formal introduction) that this is a suitable data structure (faster to construct and to scan than hash-tables, for instance) in this context. It allows to maintain a computation time of order $O(TR)$, which does not depends on the structure size.

3.2.1 The prefix-tree algorithm.

Since we use such structure in a rather non-standard way compared to other authors, e.g. [29, 26], we detail the method here.

[22] The code is available at `http://enas.gforge.inria.fr/classSuffixTree.html`.

,,

We consider a spike train ω_{-T}^{0}, where time is negative. The prefix-tree data structure for the present estimation procedure is constructed iteratively.

1. Each spiking pattern at time t, $\omega(t)$, is encoded by an integer $w(t)$.
2. This given, before any symbol has been received, we start with the empty tree consisting only of the root.
3. Then suppose for $-D < t \leq 0$ that the tree $\mathscr{T}([\omega]_{-T}^{t-1})$ represents $[\omega]_{-T}^{t-1}$. One obtains the tree $\mathscr{T}([\omega]_{-T}^{t})$ as follows:
   (a) One starts from the root and takes branches corresponding to the observed symbols $\omega(t-D+1), \cdots, \omega(t)$.
   (b) If ones reaches a leaf before termination, one replaces this leaf by an internal node and extends on the tree.
   (c) Each node or leaf has a counter incremented at each access, thus counting the number of occurrence $\#[\omega]_{-D}^{t}, -D < t \leq 0$ of the block $[\omega]_{-D}^{t}$ in the sequence $[\omega]_{-T}^{0}$.

The present data structure not only allows us to perform the empirical measure estimation over a period of time T, but can also obviously be used to aggregate several experimental periods of observation. It is sufficient to add all observations to the same data structure.

3.2.2 Generalization to a sliding window.

Though we restrict ourselves to stationary statistics in the present work, it is clear that the present mechanism can be easily generalized to the analysis of non-stationary data set, using a sliding window considering the empirical measure in $[t, t+T[$, then $[t+1, t+1+T[$, etc.. This is implemented in the present data structure by simply counting the block occurrences observed at time t and adding the block occurrences observed at time T, yielding a minimal computation load. The available implementation has already this functionality (see section 4.3.5 for an example).

3.3 Performing the parametric estimation

In a nutshell, the parametric estimation reduces to minimizing (27), by calculating the topological pressure $P(\psi) \equiv P(\lambda)$ using (14) and the related theorem. The process decomposes into the following steps.

3.3.1 Potential eigen-elements calculation.

It has been shown in the theoretical section that the Ruelle-Perron-Frobenius operator eigen-elements allows one to derive all characteristics of the probability distribution. Let us now describe at the algorithmic level how to perform these derivations.

,,

1. The first step is to calculate the right-eigenvector $\mathbf{b}(\psi)\rangle$ of the $L(\psi)$ operator, associated to the highest eigenvalue, using a standard power-method series:

$$\begin{aligned} s^{(n)} &= \|L(\psi)\mathbf{v}^{(n-1)}\| \\ \mathbf{v}^{(n)} &= \frac{1}{s^{(n)}} L(\psi)\mathbf{v}^{(n-1)} \end{aligned}$$

   where $\mathbf{v}^{(n)}$ is the n-th iterate of an initial vector $\mathbf{v}^{(0)}$ and $s^{(n)}$ is the n-th iterate of an initial real value $s^{(0)}$. With this method the pair $(s^{(n)}, \mathbf{v}^{(n)})$ converges to $(s(\psi), \mathbf{b}(\psi)\rangle)$ as soon as $\mathbf{v}^{(0)}$ is not orthogonal to $\mathbf{b}(\psi)\rangle$. In our case, after some numerical tests, it appeared a good choice to either set $\mathbf{v}^{(0)}$ to an uniform value, or to use the previous estimated value of $\mathbf{b}(\psi)\rangle$, if available. This last choice is going to speed up the subsequent steps of the estimation algorithm.

   The key point, in this iterative calculation, is that $L(\psi)$ is (hopefully) a sparse $2^{NR} \times 2^{NR}$ matrix, as outlined in the section 2.3.1. As a consequence calculating $L(\psi)\mathbf{v}$ is a $O(2^{N+NR}) \ll O(2^{2NR})$ operation, making explicit the grammar in the implementation.

   The required precision on $(s(\psi), \mathbf{b}(\psi)\rangle)$ must be very high, for the subsequent steps to be valid, even if the eigenvector dimension is huge (it is equal to 2^{NR}), therefore the iteration must be run down to the smallest reasonable precision level (10^{-24} in the present implementation).

   We have experimented that between 10 to 200 iterations are required for an initial uniform step in order to attain the required precision (for $NR \in 2..20$), while less than 10 iterations are sufficient when starting with a previously estimated value.

   From this 1st step we immediately calculate:

   (a) The topological pressure $P(\psi) = \log(s(\psi))$.

   (b) The normalized potential Ψ_w (this normalized potential is also stored in a look-up table). This gives us the transition matrix, which can be used to generate spike trains distributed according the Gibbs distribution μ_ψ and used as benchmarks in the section 4.

2. The second step is to calculate the left eigenvector $\langle\mathbf{b}(\psi)$, this calculation having exactly the same characteristics as for $\mathbf{b}(\psi)\rangle$.

   From this 2nd step one immediately calculates:

   (a) The Gibbs probability of a block w given by (15), from which probabilities of any block can be computed (section 2.3.7).

   (b) The theoretical value of the observables average $\mu_\psi(\phi_l)$, as given in (16).

   (c) The theoretical value of the distribution entropy $h\left[\mu_\psi\right]$, as given in (18).

   After both steps, we obtain all useful quantities regarding the related Gibbs distribution: probability measure, observable value prediction, entropy. These algorithmic loops are direct applications of the previous piece of theory and show the profound interest of the proposed framework: given a Gibbs potential, all other elements can be derived directly.

,,

3.3.2 Estimating the potential parameters.

The final step of the estimation procedure is to find the parameters λ such that the Gibbs measure fits at best with the empirical measure. We have discussed why minimizing (27) is the best choice in this context. Since $h(\pi^{(T)})$ is a constant with respect to λ, it is equivalent to minimize $\tilde{h}[\psi_\lambda]$ eq. (29), where $\mu_\psi(\phi_l)$ is given by (16. Equivalently, we are looking for a Gibbs distribution μ_ψ such that $\frac{\partial P[\psi_\lambda]}{\partial \lambda_l} = \pi^{(T)}(\phi_l)$ which expresses that $\pi^{(T)}$ is tangent to P at ψ_λ [38].

3.3.3 Matching theoretical and empirical observable values.

As pointed out in the theoretical part, the goal of the estimation is indeed to find the parameters λ for which theoretical and empirical observable values match. The important point is that this is exactly what is performed by the proposed method: minimizing the criterion until a minimum is reached, i.e. until the gradient vanishes corresponding to a point where $\mu_\psi(\phi_l) = \pi^{(T)}(\phi_l)$, thus where theoretical and empirical observable values are equal. Furthermore, this variational approach provides an effective method to numerically obtain the expected result.

At the implementation level, the quantities $\pi^{(T)}(\phi_l)$ are the empirical averages of the observables, i.e. the observable averages computed on the prefix tree. They are computed once from the prefix tree. For a given λ, $P(\lambda)$ is given by step 1.a of the previous calculation, while $\mu_\psi(\phi_l)$ is given by the step 2.b. It is thus now straightforward[23] to delegate the minimization of this criterion to any standard powerful non-linear minimization routine.

We have implemented such a mechanism using the GSL[24] implementation of non-linear minimization methods. We have also made available the GSL implementation of the simplex algorithm of Nelder and Mead which does not require the explicit computation of a gradient like in eq. (29). This alternative is usually less efficient than the previous methods, except in situations, discussed in the next section, where we are at the limit of the numerical stability. In such a case the simplex method is still working, whereas other methods fail.

[23] Considering a simple gradient scheme, there is always a $\varepsilon^k > 0$, small enough for the series λ_l^k and $\tilde{h}^k$, defined by:

$$\lambda_l^{k+1} = \lambda_l^k + \varepsilon^k \frac{\partial \tilde{h}}{\partial \lambda_l}(\lambda_l^k)$$

$$0 \leq \tilde{h}^{k+1} < \tilde{h}^k,$$

to converge, as a bounded decreasing series, since:

$$\tilde{h}^{k+1} = \tilde{h}^k - \varepsilon^k \left| \frac{\partial \tilde{h}}{\partial \lambda_l} \right|^2 + O\left((\varepsilon^k)^2\right).$$

[24] The GSL `http://www.gnu.org/software/gsl` multi-dimensional minimization algorithms taking the criteria derivatives into account used here is the Fletcher-Reeves conjugate gradient algorithm, while other methods such as the Polak-Ribiere conjugate gradient algorithm, and the Broyden-Fletcher-Goldfarb-Shannon quasi-Newton method appeared to be less efficient (in precision and computation times) on the benchmarks proposed in the result section. Anyway, the available code `http://enas.gforge.inria.fr/classIterativeSolver.html` allows us to consider these three alternatives, thus allowing to tune the algorithm to different data sets.

,,

3.3.4 Measuring the precision of the estimation.

Once the quantity $\tilde{h}[\psi] = P[\psi] - \pi^{(T)}(\psi)$ (eq. (29)) has been minimized the Kullback-Leibler divergence $d(\pi^{(T)}, \mu_\psi) = \tilde{h}[\psi] - h(\pi^{(T)})$ determines a notion of "distance" between the empirical measure $\pi^{(T)}$ and the statistical model μ_ψ. Though it is not necessary to compute $d(\pi^{(T)}, \mu_\psi)$ for the comparison of two statistical models $\mu_\psi, \mu_{\psi'}$, the knowledge of $d(\pi^{(T)}, \mu_\psi)$, even approximate, is a precious indication of the method precision. This however requires the computation of $h(\pi^{(T)})$.

Though the numerical estimation of $h(\pi^{(T)})$ is a far from obvious subject, we have implemented the entropy estimation using definitions (3) and (4). In order to interpolate the limit (4), we have adapted an interpolation method from [29] and used the following interpolation formula. Denote by $h(\pi^{(T)})^{(n)}$ the entropy estimated from a raster plot of length T, considering cylinders of size n. We use the interpolation formula $h(\pi^{(T)})^{(n)} \simeq h^\infty + \frac{k}{n^c}$, where $h^\infty, k, c > 0$ are free parameters, with $h(\pi^{(T)})^{(n)} \to h^\infty$, as $n \to +\infty$. The interpolation formula has been estimated in the least square sense, calculating $h(\pi^{(T)})^{(n)}$ on the prefix-tree. The formula is linear with respect to h^∞ and k, thus has a closed-form solution with respect to these two variables. Since the formula is non-linear with respect to c, an iterative estimation mechanism is implemented.

3.4 Design choices: genesis of the algorithm.

Let us now discuss in details the design choices behind the proposed algorithm.

The fact that we have an implementation able to efficiently deal with higher-order dynamics is the result of computational choices and validations, important to report here, in order for subsequent contributor to have the benefit of this part of the work.

3.4.1 Main properties of the algorithm.

Convexity. As indicated in the section 2.4.4 there is a unique minimum of the criterion. However, if $\psi^{(test)}$ contains monomials which are not in ψ, the procedure converges but there is an indeterminacy in the λ_l's corresponding to exogenous monomials. The solution is not unique, there is a subspace of equivalent solutions. The rank of the topological pressure Hessian is an indicator of such a degenerate case. Note that these different situations are not inherent to our procedure, but to the principle of finding an hidden probability by maximizing the statistical entropy under constraints, when the full set of constraints is not known [21].

Finite sample effects. As indicated in the section 2.4.5 the estimations crucially depend on T. This is a central problem, not inherent to our approach but to all statistical methods where one tries to extract statistical properties from finite empirical sample. Since T can be small in practical experiments, this problem can be circumvented by using an average over several samples. In the present thermodynamic formalism it is possible to have an estimation of the size of fluctuations as a function of the potential, using the central limit theorem and the fact that the variance of fluctuations is given by the second derivative of the topological pressure. This is a further statistical test

,,

where the empirical variance can be easily measured and compared to the theoretical predictions.

Numerical stability of the method. Two factors limitate the stability of the method, from a numerical point of view.

The first factor is that the RPF operator is a function of the *exponential* of the potential $\psi = \sum_l \lambda_l \phi_l$. As a consequence, positive or negative values of ψ yield huge or vanishing value of L_ψ, and numerical instabilities easily occurs.

However, though numerical instabilities are unavoidable, the good news is that they are easily detected, because we have introduced a rather large set of numerical tests in the code:

1. Negligible values (typically lower than 10^{-4}) are set to zero, implicitly assuming that they correspond to hidden transition in the grammar.

2. Huge value (typically higher than 10^4) generate a warning in the code.

3. Several coherent tests regarding the calculation of the RPF eigen-elements are implemented: we test that the highest eigenvalue is positive (as expected from the RPF theorem), and that the left and right RPF related eigenvectors yield equal eigenvalues, as expected; we also detect that the power-method iterations converge in less than a maximal number of iteration (typically 2^{10}). We never found this spurious condition during our numerical tests. When computing the normalized potential (19), we verify that the right eigenvalue is 1 up to some precision, and check that the normal potential is numerically normalized (i.e. that the sum of probabilities is indeed 1, up to some "epsilon").

In other words, we have been able to use all what the piece of theory developed in the previous section makes available, to verify that the numerical estimation is valid.

The second factor of numerical imprecision is the fact that some terms $\lambda_l \phi_l$ may be negligible with respect to others, so that the numerical estimation of the smaller terms becomes unstable with respect to the imprecision of the higher ones. This has been extensively experimented, as reported in the next section.

Relation with entropy estimation. The construction of a prefix-tree is also the basis of efficient entropy estimation methods [29, 68]. See [26] for a comparative about entropy estimation of one neuron spike train (binary time series). Authors numerically observed that the context-tree weighting methods [42] is seen to provide the most accurate results. This, because it partially avoids the fact that using small word-lengths fails to detect longer-range structure in the data, while with longer word-lengths the empirical distribution is severely under-sampled, leading to large biases. This statement is weaken by the fact that the method from [68] is not directly tested in [26], although a similar prefix-tree method has been investigated.

However the previous results are restrained to relative entropy estimation of "one neuron" whereas the analysis of entropy of a *group of neurons* is targeted if we want to better investigate the neural code. In this case [68] is directly generalizable to non-binary (thus multi-neurons) spike trains, whereas the context-tree methods seems in-

,,
trinsically limited to binary spike-trains [42], and the numerical efficiency of these methods is still to be studied at this level.

Here, we can propose an estimation for the KS-entropy from eq. (18). Clearly, we compute here the entropy of a Gibbs statistical model μ_ψ while methods above try to compute this entropy from the raster plot. Thus, we do not solve this delicate problem, but instead, propose a method to benchmark these methods from raster plots obeying a Gibbs statistics, where the Gibbs distribution approaches at best the empirical measures obtained from experiments.

3.4.2 Key aspects of the numerical implementation.

Estimating the grammar from the empirical measure.

The grammar defined in (13) is implemented as a Boolean vector indexed by w and estimated by observing, in a prefix-tree of depth at least $R+1$, whose blocks $[\omega]_{-R-1}^{0}$ occur at least once (allowed transition). We make therefore here the (unavoidable) approximation that unobserved blocks correspond to forbidden words (actually, our implementation allows to consider that a block is forbidden if it does not appear more than a certain threshold value). There is however, unless a priori information about the distribution is available, no better choice. The present implementation allows us to take into account such a priori information, for instance related to global time constraints on the network dynamics, such as the refractory period. See [12] for an extended discussion.

Potential values tabulation.

Since the implementation is anyway costly in terms of memory size, we have choosen to pay this cost but obtaining the maximal benefit of it and we used as much as possible tabulation mechanisms (look-up tables) in order to minimize the calculation load. All tabulations are based on the following binary matrix:

$$\mathbf{Q} \in \{0,1\}^{L\times 2^{NR}},$$

with $\mathbf{Q}_{l,w} = \phi_l([\omega]_{-R}^{0})$, where w is given by (10). $\mathbf{Q}$ is the matrix of all monomial values, entirely defined by the choice of the parameter dimensions N, R and D. It corresponds to a "look-up table" of each monomial values where w encodes $[\omega]_{-R}^{0}$. Thus the potential (6) writes $\psi_w = (\mathbf{Q}\lambda)_w$. We thus store the potential exponential values as a vector and get values using a look-up table mechanism, speeding-up all subsequent computations.

This allows to minimize the number of operations in the potential eigen-elements calculation.

3.4.3 Appendix: About other estimation alternatives.

Though what is proposed here corresponds, up to our best knowledge, to the best we can do to estimate a Gibbs parametric distribution in the present context, this is obviously not the only way to do it, and we have rejected a few other alternatives, which

,,

appeared less suitable. For the completeness of the presentation, it is important to briefly discuss these issues.

Avoiding RPF right eigen-element's calculation. In the previous estimation, at each step, we have to calculate step 1 of the RPF eigen-element's derivation for the criterion value calculation and step 2 of the RPF eigen-element's derivation for the criterion gradient calculation. These are a costly $O(2^{N+NR})$ operations.

One idea is to avoid step 2 and compute the criterion gradient numerically. We have explored this track: we have calculated $\frac{\partial \tilde{h}}{\partial \lambda_l} \simeq \frac{\tilde{h}(\lambda_l+\varepsilon)-\tilde{h}(\lambda_l-\varepsilon)}{2\varepsilon}$ for several order of magnitude, but always found a poorer convergence (more iteration and a biased result) compared to using the closed-form formula. In fact, each iteration is not faster, since we have to calculate $\tilde{h}$ at two points thus, to apply step 1, at least two times. This variant is thus to be rejected.

Another idea is to use a minimization method which does not require the calculation of the gradient: we have experimented this alternative using the simplex minimization method, instead of the conjugate gradient method, and have observed that both methods correctly converge towards a precise solution in most cases, while the conjugate gradient method is faster. However, there are some cases with large range potential, or at the limit of the numerical stability where the simplex method may still converge, while the other does not.

Using a simple Gibbs form. Using the Gibbs form

$$\mu_\psi[w_t,\ldots,w_{t+n}] = \frac{e^{\sum_{k=t}^{t+n}\psi_{w_k}}}{Z_n}, \text{ with } Z_n = \sum_{w_t,\ldots,w_{t+n}} e^{\sum_{k=t}^{t+n}\psi_{w_k}},$$

where Z_n is a constant, could provide an approximation of the right Gibbs distribution and of the topological pressure, avoiding the power-method internal loop. Furthermore, instead of a costly $O(2^{N+NR})$ operation, calculating Z_n (and derivatives) would require a simple scan of the prefix-tree (since values are calculated at each step weighted by the empirical measure values) thus $O(2^{NR})$ operations. This apparent gain is unfortunately impaired since the amount of calculation is in fact rather heavy. Moreover, as widely commented on section 2, the result is biased with a *non negligible* additional bias increasing with the range R of the potential. Finally, it has been observed as being slower than for the basic method.

About analytical estimation of the RPF eigen-element's. The costly part of the RPF eigen-element's computation is the estimation of the highest eigenvalue. It is well-known that if the size of the potential is lower than five, there are closed-form solutions, because this problem corresponds to finding the root of the operator characteristic polynomial. In fact, we are going to use this nice fact to cross-validate our method in the next section. However, except for toy's potentials (with $2^{NR} < 5 \Leftrightarrow NR \leq 2$!), there is no chance that we can not do better than *numerically* calculating the highest eigenvalue.

,,

And the power method is known as the most powerful way to do it, in the general case. We thus have likely optimal choices at this stage.

Using other approximations of the KL-divergence criterion. Let us now discuss another class of variants: the proposed KL-divergence criterion in (22) and its empirical instantiation in (27) are not the only one numerical criterion that can be proposed in order to estimate the Gibbs distribution parameters. For instance, we have numerically explored approximation of the KL-divergence of the form:

$$d(\nu,\mu) \simeq \sum_{n=R}^{R'} \frac{\alpha_n}{n} \sum_{[\omega]_0^{n-1}} \nu\left([\omega]_0^{n-1}\right) \log\left[\frac{\nu\left([\omega]_0^{n-1}\right)}{\mu\left([\omega]_0^{n-1}\right)}\right],$$

and have obtained coherent results (for $\alpha_n = 1$), but not quantitatively better than what is observed by the basic estimation method, at least for the set of performed numerical tests.

All these variants correspond to taking into account the same kind of criterion, but some other weighted evaluations of the empirical average of the observable. There is no reason to use it unless some specific a priori information on the empirical distribution is available.

Another interesting track is to use (19) which allows us to write a KL-divergence criterion, not on the probability block, but on the conditional probability block, as proposed in [14, 15] in a different context. We have considered this option. However a straightforward derivation allows one to verify, that this in fact corresponds the same class of criterion but with a different empirical observable average estimation. At the numerical level, we did not observe any noticeable improvement.

Using score matching based estimation. We are here in a situation where we have to estimate a parametric statistical distribution, whose closed-form is given up to a scale factor Z_n. Such model contains a normalization constant whose computation may be considered as too difficult for practical purposes, as it is the case for some maximum likelihood estimations. Score-matching methods [32] are based on the gradient of the log-density with respect to the data vector, in which the normalization constant is eliminated. However, the estimation criterion is no more the KL-divergence, and there is no guaranty that the obtained solution is not biased with respect to a well-defined statistical quantity. As such it is another candidate to estimate Gibbs distribution. However, thanks to the eigen decomposition of the RPF operator, we do not need to use this trick, since we obtain a tractable calculation of the normalization constant at each step of the estimation and can minimize a well-defined criterion, as proposed in this paper.

We have numerically checked such modification of the criterion in which we do not consider the KL-divergence criterion, but the *ratio* between two conditional probabilities, as defined in (19). Considering this ratio allows to eliminate the scale factor Z_n. This is the same spirit as score matching based estimation, more precisely, it corresponds to a discrete form of it, where the gradient of the log-density is replaced by finite

,,
difference. We have obtain correct results for simple forms of potential, but have experimented that the method is numerically less robust than using the unbiased method developed in this paper. This confirms that using the eigen-decomposition of the RPF operator, is the key for numerically stable estimations of such parametric statistics.

Estimation in the case of a normalized potential. In the case where the potential is normalized, the criterion (29) is a simple linear criterion, thus unbounded and its minimization is meaningless. In this singular case, its is obvious to propose another criterion for the estimation of the parameters. A simple choice is to simply propose that the theoretical likelihood of the measure matches the estimated one, in the *least square sense*. This has been integrated in the available code.

4 Results

4.1 Basic tests: validating the method

4.1.1 Method

Knowing the potential ψ, it is easy to generate a spike train of length T, distributed according to μ_ψ, using the Chapman-Kolmogorov equations (20). Thus, we have considered several examples of Gibbs potentials, where, starting from a sample raster plot $[\omega]_{-T}^{0}$ distributed according to μ_ψ, we use our algorithm to recover the right form of ψ.

Given a potential of range-R of the parametric form (6) and a number of neurons N we apply the following method:

1. Randomly choosing the parameter's values $\lambda_l, l = 1 \dots L$ of the Gibbs potential;
2. Generating a spike train realization of length T;
3. From these values re-estimating a Gibbs potential:
   (a) Counting the block occurrences, thus the probabilities $\pi^{(T)}$ from the prefix-tree,
   (b) Minimizing (29), given $\pi^{(T)}$, as implemented by the proposed algorithm.
   (c) Evaluating the precision of the estimation as discussed in the previous section.

In the previous method, there is a way to simulate "infinite" ($T = +\infty$) sequences, by skipping step 2., and filling the prefix-tree in step 3.a directly by the exact probability $\mu_\psi(w)$ of the blocks w.

At a first glance, this loop seems to be a "tautology" since we re-estimate the Gibbs potential parameters from a one-to-one numerical process. However, this is not the case for three reasons:

,,

1. For $T = +\infty$ using the same potential for the prefix-tree generation and for the parameters estimation, must yield the same result, but *up to the computer numerical precision*. This has to be controlled due to the non-linear minimization loop in huge dimension. This is obviously also a way to check that the code has no mistake.

2. For $T < +\infty$ using the same potential allows us to study the numerical precision of the estimation in the realistic situation of finite size data set, providing quantitative estimations about the truncation effects to be expected.

3. Using different potentials between simulated data generation and the parameters value estimation allows us to study numerically to which extends we can only correctly estimate the parameter's values, even if huge state vectors are involved. Quantitative errors are obtained. We can also perform comparison between different statistical models, as detailed in the sequel.

4.1.2 An illustrative example to understand what the algorithm calculates

Let us start with very simple example, for which we can make explicit what the algorithm calculates, thus helping the reader to understand in details what the output is.

We consider a situation where the number L of parameters λ_l is known (only the values of the λ_l's are unknown). We start from rather basic examples and then increase their complexity. In the first examples analytical expression for the topological pressure, entropy, RPF eigen-vectors and invariant measure are available. Thus we can check that we re-obtain, from the estimation method, the related values up to the numerical imprecision.

One neuron and range-2. Here $\psi(\omega) = \lambda_1\,\omega_0(0) + \lambda_2\,\omega_0(0)\,\omega_0(1)$. We obtain analytically:

$$\begin{array}{lcl}
s(\psi) & = & \frac{1+B+\sqrt{(1-B)^2+4A}}{2}, \\
P(\psi) & = & \log s(\psi), \\
\langle \mathbf{b}(\psi) & = & (1, s(\psi)-1, A, B(s(\psi)-1),) \\
\mathbf{b}(\psi)\rangle & = & (s(\psi)-B, s(\psi)-B, 1, 1)^T, \\
\mu_\psi & = & \frac{1}{s(\psi)^2+A-B}\,(s(\psi)-B, A, A, B(s(\psi)-1)), \ , \\
h\left[\mu_\psi\right] & = & \log(s(\psi)) - \lambda_1 \frac{\partial s(\psi)}{\partial \lambda_1} - \lambda_2 \frac{\partial s(\psi)}{\partial \lambda_2} \\
r & = & \frac{A+B(s(\psi)-1)}{s(\psi)^2+A-B}, \\
C & = & \frac{B(s(\psi)-1)}{s(\psi)^2+A-B},
\end{array}$$

with $A = e^{\lambda_1} = e^{\psi_{10}}, B = e^{\lambda_1+\lambda_2} = e^{\psi_{11}}$ and where T denotes the transpose. We remind that the index vector encodes spike blocs by eq. (10). Thus, the first index (0) corresponds to the bloc 00, 1 to 01, 2 to 10 and 3 to 11. r is the firing rate, C the probability that the neuron fires two successive time steps. This is one among the few models for which a closed-form solution is available.

,,

The following numerical verifications have been conducted. A simulated prefix-tree whose nodes and values has been generated using (6) with $\lambda_1 = \log(2), \lambda_2 = \log(2)/2$. We have run the estimation program of λ_i's and have obtained the right values with a precision better than 10^{-6}. We also obtain a precision better than 10^{-6} for $s(\psi), r, C, h\left[\mu_\psi\right]$. This first test simply states that the code has no mistake.

A step further, we have used this simple potential to investigate to which extends we can detect if the model is of range-1 (i.e. with $\lambda_2 = 0$) or range-2 (i.e. with a non-negligible value of λ_2). To this purpose, we have generated a range-2 potential and have performed its estimation using a range-1 and a range-2 potential, comparing the entropy difference (Fig. 4.1.2).

As expected the difference is zero for a range-2 model when $\lambda_2 = 0$, and this difference increases with λ_2. Less obvious is the fact that curves saturate for high values of λ_2. Indeed, because of the exponential function, high values of λ yield huge or vanishing values of the RPF operator, thus numerical instabilities. This instability is detected by our algorithm. Note that values of λ larger than 10 in absolute value have little sense from a statistical analysis of spike trains perspective.

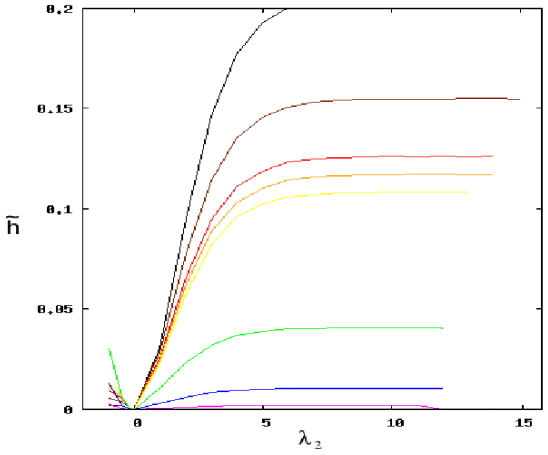

Figure 1: Entropy difference, using $\tilde{h}$, defined in (29), between the estimations of a range-1 and a range-2 model. The range-2 model writes $\phi = -\lambda_1\,\omega_0(0) - \lambda_2\,\omega_0(0)\,\omega_0(1)$ for $\lambda_1 = \{-1\,(black), -0.5\,(brown), -0.2\,(red), -0.1\,(orange), 0\,(green), 1\,(blue), 2\,(Magenta)\}$. λ_2 is a free parameter, in abscissa of this curve. The range-1 corresponds to $\lambda_2 = 0$.

We also have generated a range-1 potential and have performed its estimation, using a range-1 versus a range-2 model, and found always that using range-2 model is as good as using a model of range-1 (not shown).

Two neurons and range-2 (Ising). Here $\psi(\omega) = \lambda_1\,\omega_1(0) + \lambda_2\,\omega_2(0) + \lambda_3\,\omega_1(0)\,\omega_2(0)$. The largest eigenvalue of the RPF operator is $Z = s(\psi) = A + B + C + D$, with $A = 1, B = e^{\lambda_1}, C = e^{\lambda_2}, D = e^{\lambda_1+\lambda_2+\lambda_3}$ and the topological pressure is $\log s(\psi)$. Here the Gibbs distribution has the classical form. We still obtain numerical precision better than 10^{-4}, for standard values of λ, e.g., $\lambda_1 = 1, \lambda_2 = \log(2), \lambda_3 = \log(2)/2$.

,,

Two neurons and pattern of spikes. A step further, we have considered $\psi(\omega) = \lambda_1 \omega_1(0) + \lambda_2 \omega_2(0) + \lambda_3 \omega_1(0)\omega_2(1)\omega_1(2)$, and $\psi(\omega) = \lambda_1 \omega_1(0) + \lambda_2 \omega_2(0) + \lambda_3 \omega_1(0)\omega_2(1)\omega_2(2)\omega_3(3)$, for random values drawn in $]-1,0[$, i.e., considering the statistical identification of *spike patterns*. We still obtain numerical precision better than 10^{-3}, for these standard values of λ, though the precision decreases with the number of degrees of freedom, as expected, while it increases with the observation time. This is investigated in details in the remainder of this section.

When considering larger neuron N and range-R the main obstacle toward analytical results is the Galois theorem which prevent a general method for the determination of the largest eigenvalue of the RPF operator. Therefore, we only provide numerical results obtained for more general potentials.

4.1.3 Gibbs potential precision paradigm: several neurons and various ranges.

In order to evaluate the numerical precision of the method, we have run the previous benchmark considering potentials with all monomial of degree less or equal to 1, and less or equal to 2, at a various ranges, with various numbers of neurons. Here we have chosen $T = +\infty$ and used the same potential for the prefix-tree generation and for the parameters value estimation. The computation time is reported in Table 1 and the numerical precision in Table 2, for $NR \leq 16$. This benchmark allows us to verify that there is no "surprise" at the implementation level: computation time increases in a supra-linear way with the potential size, but, thanks to the chosen estimation method, remains tractable in the size range compatible with available memory size. This is the best we can expect, considering the intrinsic numerical complexity of the method. Similarly, we observe that while the numerical precision decreases when considering large size potential, the method remains stable. Here tests has been conducted using the standard 64-bits arithmetic, while the present implementation can easily be recompiled using higher numerical resolution (e.g. "long double") if required.

A step further, this benchmark has been used to explore the different variants of the estimation method discussed in the previous section (avoiding some RPF eigen-element's calculation, using other approximations of the KL-divergence criterion, ..) and fix the details of the proposed method.

,,

Table 1: Cpu-time order of magnitude in second (using Pentium M 750 1.86 GHz, 512Mo of memory), for the estimation of a potential with all monomial of degree less or equal to 1, for ψ_1 and less or equal to 2, for ψ_2, (i.e., $\psi_1(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0)$, $\psi_2(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \sum_{\tau=-T_s}^{T_s} \lambda_{ij\tau} \omega_i(0)\omega_j(\tau)$) at a range-$R = 2T_s + 1$ with N neurons. We clearly observe the exponential increase of the computation time. Note that the present implementation is not bounded by the computation time, but simply by the exponential increase of the memory size.

ψ_1	R=1	R=2	R=4	R=8	R=16	ψ_2	R=1	R=2	R=4	R=8	R=16
N=1	2.0e-06	3.0e-06	8.0e-06	7.8e-05	2.9e-01	N=1	4.5e-16	4.0e-06	4.0e-06	7.2e-04	3.7e-02
N=2	4.0e-06	1.0e-06	3.0e-05	6.7e-02		N=2	3.0e-06	5.0e-06	4.0e-04	1.1e+00	
N=4	1.3e-05	3.8e-05	8.3e-02			N=4	1.9e-05	1.2e-03	3.6e+00		
N=8	2.4e-03	3.2e-01				N=8	6.6e-03	6.2e-01			

Table 2: Numerical precision of the method using synthetic data, for the estimation of ψ_1 and ψ_2, at a range-R with N neurons. The Euclidean distance $|\bar{\lambda} - \tilde{\lambda}|$ between the estimated parameter's value $\tilde{\lambda}$ and the true parameter's value $\bar{\lambda}$ is reported here, when the $\bar{\lambda}_l$'s are randomly drawn in $[-1,1]$. We clearly observe the error increase, but the method remaining numerically stable.

ψ_1	R=1	R=2	R=4	R=8	R=16	ψ_2	R=1	R=2	R=4	R=8	R=16
N=1	5.0e-09	2.2e-02	6.3e-03	1.3e-02	6.9e-03	N=1	1.1e-10	1.9e-02	7.2e-03	4.8e-03	9.2e-02
N=2	1.1e-08	1.3e-02	9.2e-03	5.2e-03		N=2	1.1e-09	4.8e-03	3.7e-03	2.3e-03	
N=4	8.0e-09	8.5e-03	6.8e-03			N=4	3.7e-08	2.6e-03	5.8e-02		
N=8	3.8e-08	5.1e-03				N=8	6.0e-06	2.4e-02			

4.2 More general tests: applying the method

4.2.1 Test framework.

In order to test more general potentials for $N = 2$ neurons we explicit here the forms (7), (8), (9), where $k \in \mathbb{N}$:

$$\begin{aligned} \text{Ising} &: \psi(\omega) = \lambda_1\,\omega_1(0) + \lambda_2\,\omega_2(0) + \lambda_3\,\omega_1(0)\,\omega_2(0). \\ \text{RPTD}-k &: \psi(\omega) = \lambda_1\,\omega_1(0) + \lambda_2\,\omega_2(0) + \sum_{i=-k}^{i=k} \hat{\lambda}_i\,\omega_1(0)\omega_2(i). \\ \text{PTD}-k &: \psi(\omega) = \sum_{i=-k}^{i=k} \hat{\lambda}_i\,\omega_1(0)\omega_2(i). \end{aligned} \tag{37}$$

test 1 (estimation precision). Given a selected potential of form (37) we choose randomly its coefficients $\bar{\lambda}_l$ from an uniform distribution on $[-2,0]$ and we generate a spike-train of length $T = 4 \times 10^8$. Then we construct a prefix-tree from a sample of length $T_0 \ll T$ (typically $T_0 = 10^7$) taken from the generated spike-train. For each sample of length T_0 we propose a randomly chosen set of "initial guess" coefficients, used to start the estimation method, distributed according to $\tilde{\lambda}_l^{(0)} = \bar{\lambda}_l(1 + (U[0,1] - 0.5)x/100)$, where x is the initial percentage of bias from the original set of generating coefficients and $U[0,1]$ is a uniform random variable on $[0,1]$. Call $\tilde{\lambda}_l$ the values

,,

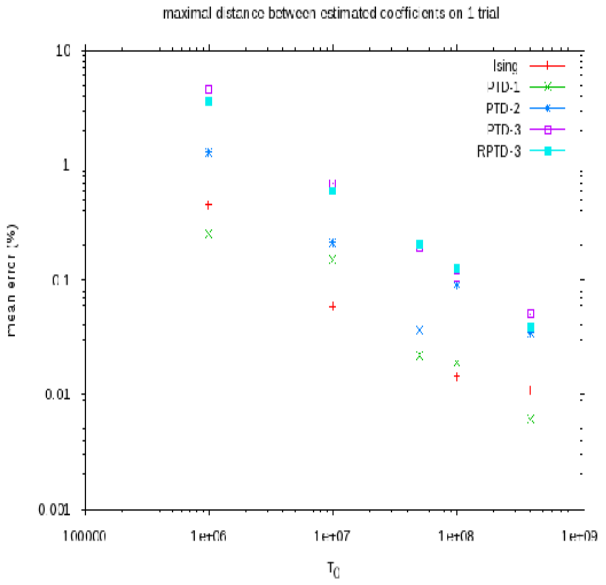

Figure 2: Mean error (in percentage) vs T_0 size.

obtained after convergence of the algorithm. Results show that:

(i) the error $E(|\tilde{\lambda}_l - \bar{\lambda}_l|)$ increases with the range of the potential and it decreases with T_0;

(ii) the error is independent of the initial bias percentage (see figs 4.2.1);

(iii) $\tilde{h}[\psi] = P[\psi] - \pi^{(T)}(\psi)$ is fairly constant with respect to the length T_0 (not shown).

Test 2 (Models comparison). We select a potential form ψ from those proposed in (37); we choose randomly its coefficients λ_l from an uniform distribution in $[-2,0]$; we generate a spike-train of length $T = 1 \cdot 10^8$ and we construct the prefix-tree with the spike-train obtained. Using this prefix-tree we estimate the coefficients $\lambda_i^{\psi_m}$ that minimizes the KL divergence for several statistical models ψ_m proposed in (37). The coefficients $\lambda_i^{\psi_m}$ and $\tilde{h} = P[\psi_m] - \pi^{(T)}(\psi_m)$ are averaged over 20 samples and error bars are computed. Results show that :

(i) The 'best' statistical models (i.e the ones with lowest mean value KL divergence) have the same monomials as the statistical model that generated the spike-train, plus, possibly additional monomials. For example, in (37), **RPTD**-1 contains **Ising**, and also the **PTD**-1 but not **PTD**-2. We choose the model with the minimal number of coefficients in agreement with section 2.4.4.

,,

(ii) The value of the additional coefficients of an over-esimated model (corresponding to monomials absent in the corresponding potential) are almost null up to the numerical error.

(iii) For all the 'best' suited statistical models (in the sense of (i)), the criterion $\tilde{h}[\psi]$ (29) averaged over trials, is fairly equal for these models up to a difference of order $\delta \approx 10^{-6}$, and the difference with respect to other types of statistical models is at least of 4 orders of magnitude lower. We recall that, according to section 2.4.5, the deviation probability is of order to $\exp(-\delta T)$. After estimation from a raster generated with an Ising model, the ratio of the deviation probabilities (33) between an **Ising** and a **RPTD-**1 model is $\sim \eta = \exp(0.0000115 \times 10^8)$, while between the **Ising** and the **PTD-**3 $\sim \eta = \exp(0.00072 \times 10^8)$ meaning that the **PTD-**3 provide a worst estimation.

(iv) The predicted probability of words corresponds very well with the empirical value.

In order to extend the model comparison we introduce the following notations: let w be a word (encoding a spiking pattern) of length R, $P_{est}(w)$ its mean probability over trials calculated with the estimated potential, $P_{emp}(w)$ its mean empirical average over trials (i.e average of form (24) including a time average $\pi^{(T)}$ and a sample average, where the samples are contiguous pieces of the raster of length $T_0 \ll T$), and $\sigma_{emp}(w)$ the standard deviation of $P_{emp}(w)$. We now describe the comparison methods.

We first use the box-plot method [25] which is intended to graphically depict groups of numerical data through their 'five-number summaries' namely: the smallest observation (sample minimum), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation (sample maximum)[25]. Figure 4.2.1 shows, in log-scale, the box-plot for the distribution of the quantity defined as:

$$\varepsilon(w) = \left| \left(P_{est}(w) - P_{emp}(w)\right) / \sigma_{emp}(w) \right| \tag{38}$$

that is taken as a weighted measure of the deviations. We have considered this distribution when it takes into account, either all the words up to a given size R_{max}, or only the words of that given size. There is no visual difference for $R_{max} = 7$. The results shows that only models containing the generating potential have the lower deviations value with very similar box. On the other hand a "bad" statistical model shows a much more extended error distribution .

Finally a χ^2 estimation is computed as $\chi^2 = \frac{1}{N_{\text{words}} - L} \sum_w \varepsilon(w)^2$ where $\varepsilon(w)$ is given by (38). Values are reported in tables 3, using all words or only those of size R_{max}. Since the number of words is high, it is clear that the lower the error, the lower the

[25] The largest (smallest) observation is obtained using parameter dependent bounds, or "fences", to filter aberrant uninteresting deviations. Call $\beta = Q3 - Q1$ and let k denote the parameter value, usually between 1.0 and 2.0. Then the bound correspond to $Q3 + k\beta$ for the largest observation (and for the smallest one to $Q1 - k\beta$). A point x found above (below) is called "mild-outlier" if $Q3 + k < x < Q3 + 2k\beta$ (respectively, $Q1 - 2k\beta < x < Q3 - k\beta$) or extreme outlier if $x > Q3 + 2k\beta$ (respectively, $x < Q1 - 2k\beta$). We have used a fence coefficient $k = 2.0$ to look for outliers.

,,

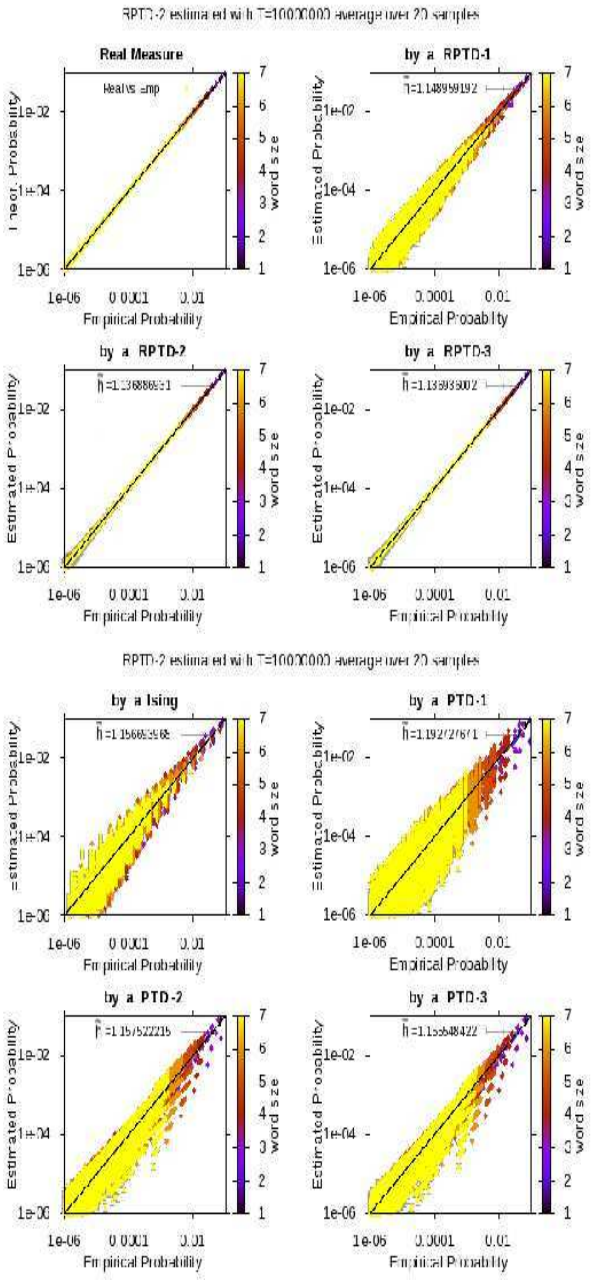

Figure 3: Figure 1 (top left) Expected probability μ_ψ versus empirical probability $\pi_\omega^{(T)}(w)$; Figure 2 (top right) to 8 (bottom right) Predicted probability versus empirical probability $\pi_\omega^{(T)}(w)$ for several models. The generating potential is a **RPTD-2**.

,,

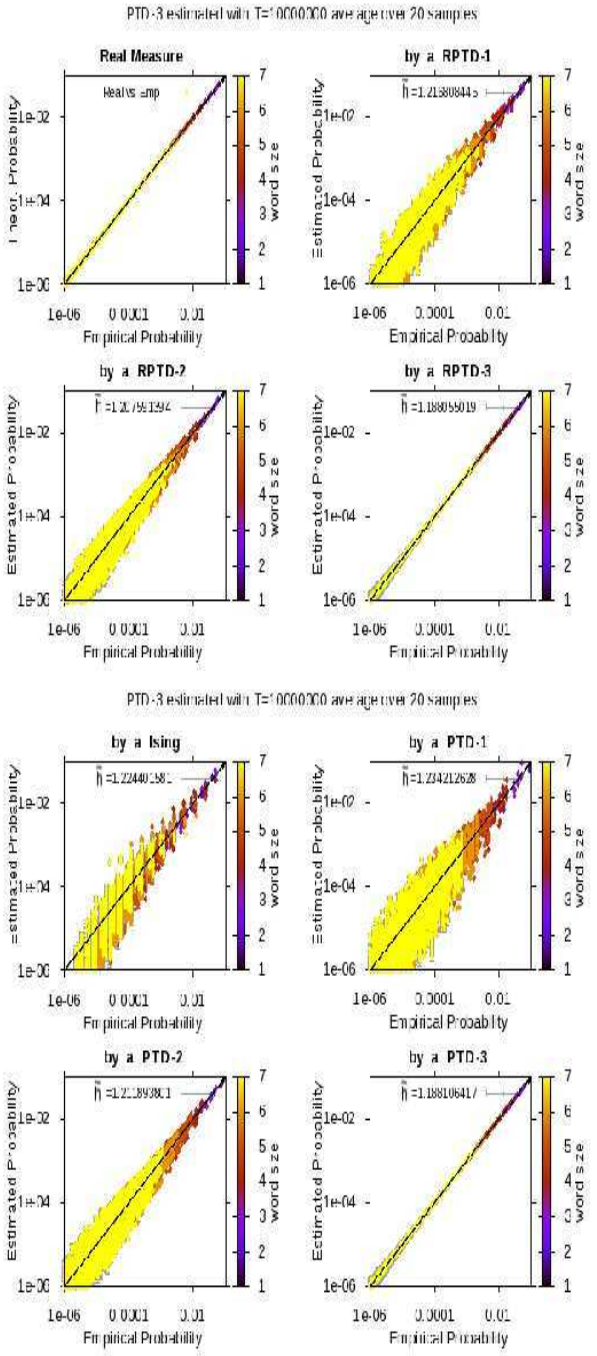

Figure 4: Same as previous figure where generating potential is a **PTD-3**.

”

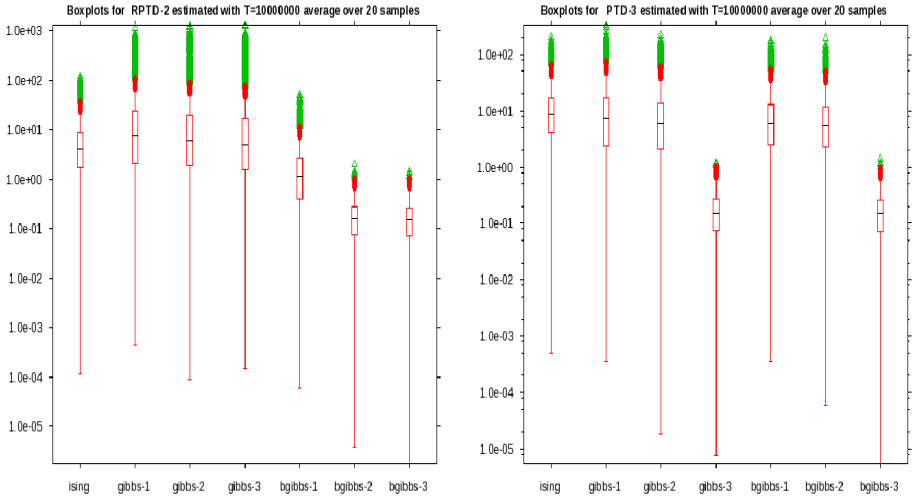

Figure 5: The box-plot (in log-scale) of the distributions of weighted deviations of word's probability versus their empirical probability, for several statistical models, using a generating potential of the form **(left) RPTD**-2 and **(right) PTD**-3. Midliers Outliers (see footnote 25) are shown by red dots and extreme outliers by green dots.

χ^2 estimated value. Note that χ^2 test assumes Gaussian fluctuations about the mean value, which are satisfied for finite-range Gibbs distributions, as can be easily seen by expanding the large deviations function I_l in (32) up to the second order in ε. However, when comparing two different Gibbs distributions it might be that the deviations from the expected value of one Gibbs distribution compared to the expected value of the other Gibbs distribution is well beyond the mean-square deviation of the Gaussian fluctuations distribution, giving rise to huge χ^2 coefficients, as we see in the tables 3.

4.3 Spike train statistics in a simulated Neural Network

Here we simulate an Integrate-and-Fire neural network whose spike train statistics is explicitely and rigorously known [11] while effects of synaptic plasticity on statistics

Table 3: χ^2 **coefficient calculated**: (left) with all words of size < 7; (right) with words of size 7 only. See text for details.

Estimating \Generating	RPTD-2	PTD-3	Estimating \Generating	RPTD-2	PTD-3
Ising	135.427	415.965	Ising	121.825	347.502
PTD-1	3146.17	564.396	PTD-1	2839.36	468.763
PTD-2	3319.75	290.93	PTD-2	2537.39	229.255
PTD-3	2533.35	0.0571905	PTD-3	2053.72	0.057065
RPTD-1	13.9287	274.773	RPTD-1	11.6167	218.458
RPTD-2	0.0607027	223.516	RPTD-2	0.0605959	176.598
RPTD-3	0.0556114	0.0539691	RPTD-3	0.0553242	0.0541206

,,

have been studied in [12].

4.3.1 Network dynamics.

The model is defined as follows. Denote by V_i the membrane potential of neuron i and W_{ij} the synaptic weight of neuron j over neuron i, I_i^{ext} an external input on neuron i. Each neuron is submitted to noise, modeled by an additional input, $\sigma_B B_i(t)$, with $\sigma_B > 0$ and where the $B_i(t)$'s are Gaussian, independent, centered random wariable with variance 1. The network dynamics is given by:

$$V_i(t+1) = \gamma V_i(1 - Z[V_i(t)]) + \sum_{j=1}^{N} W_{ij} Z[V_j(t)] + I_i^{ext} + \sigma_B B_i(t); \qquad i = 1 \dots N, \quad (39)$$

where $\gamma \in [0,1[$ is the leak in this discrete time model ($\gamma = 1 - \frac{dt}{\tau}$). Finally, the function $Z(x)$ mimics a spike: $Z(x) = 1$ if $x \geq \theta = 1$ and 0 otherwise, where θ is the firing threshold. As a consequence, equation (39) implements both the integrate and firing regime. It turns out that this time-discretisation of the standard integrate-and-Fire neuron model, which as discussed in e.g. [33], provides a rough but realistic approximation of biological neurons behaviors. Its dynamics has been fully characterized for $\sigma_B = 0$ in [10] while the dynamics with noise is investigated in [11]. Its links to more elaborated models closer to biology is discussed in [13].

4.3.2 Exact spike trains statistics.

For $\sigma_B > 0$ there is a unique Gibbs distribution in this model, whose potential is explicitely known. It is given by:

$$\phi(\omega_{-\infty}^{0}) = \sum_{i=1}^{N} \left[\omega_i(0) \log\left(\pi\left(\frac{\theta - C_i(\underline{\omega})}{\sigma_i(\underline{\omega})} \right) \right) + (1 - \omega_i(0)) \log\left(1 - \pi\left(\frac{\theta - C_i(\underline{\omega})}{\sigma_i(\underline{\omega})} \right) \right) \right], \quad (40)$$

where $\pi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-\frac{u^2}{2}} du$, $\underline{\omega} = \omega_{-\infty}^{-1}$, $C_i(\underline{\omega}) = \sum_{j=1}^{N} W_{ij} x_{ij}(\underline{\omega}) + I_i^{ext} \frac{1-\gamma^{t+1-\tau_i(\underline{\omega})}}{1-\gamma}$, $x_{ij}(\underline{\omega}) = \sum_{l=\tau_i(\underline{\omega})}^{t} \gamma^{t-l} \omega_j(l)$, $\sigma_i^2(\underline{\omega}) = \sigma_B^2 \frac{1-\gamma^{2(t+1-\tau_i(\underline{\omega}))}}{1-\gamma^2}$. Finally, $\tau_i(\underline{\omega})$ is the last time, before $t = -1$, where neuron i has fired, in the sequence $\underline{\omega}$ (with the convention that $\tau_i(\underline{\omega}) = -\infty$ for the sequences such that $\omega_i(n) = 0, \forall n < 0$). This potential has infinite range but range $R \geq 1$ approximations exist, that consist of replacing $\underline{\omega} = \omega_{-\infty}^{-1}$ by ω_{-R}^{-1} in (40). The KL divergence between the Gibbs measure of the approximated potential and the exact measure decays like γ^R. Finite range potentials admit a polynomial expansion of form (5).

4.3.3 Numerical estimation of spike train statistics

Here we have considered only one example of model (39) (more extended simulations and results will be provided elsewhere). It consists of 4 neurons, with a *sparse* connectivity matrix so that there are neurons without synaptic interactions. The synaptic weigths matrix is:

,,

$$\mathscr{W} = \begin{pmatrix} 0 & -0.568 & 1.77 & 0 \\ 1.6 & 0 & -0.174 & 0 \\ 0 & 0.332 & 0 & -0.351 \\ 0 & 1.41 & -0.0602 & 0 \end{pmatrix},$$

while $\gamma = 0.1, \sigma_B = 0.25, I_i^{ext} = 0.5$.

First, one can compute directly the theoretical entropy of the model using the results exposed in the previous section: the entropy of the range-R approximation, that can be computed with our formalism, converges exponentially fast with R to the entropy of the infinite range potential. For these parameters, the asymptotic value is $h = 0.57$.

Then, we generate a raster of length $T = 10^7$ for the 4 neurons and we compute the KL divergence between the empirical measure and several potentials including:

- (i) The range-R approximation of (40), denoted $\phi^{(R)}$. Note that $\phi^{(R)}$ does not contain all monomials. In particular, *it does not have the Ising term (the corresponding coefficient is zero)*.
- (ii) A Bernoulli model ϕ^{Ber};
- (iii) An Ising model ϕ^{Is};
- (iv) A one-time step Ising Markov model (as proposed in [44]) ϕ^{MEDF} [26] ;
- (v) A range-R model containing all monomials ϕ^{all}.

Here we can compute the KL divergence since we known the theoretical entropy. The results are presented in the table (4). Note that the estimated KL divergence of range-1 potentials slightly depend on R since the RPF operator, and thus the pressure, depend on R.

Table 4: Kullback-Leibler divergence between the empirical measure of a raster generated by (39) (See text for the parameters value) and the Gibbs distribution, for several statistical models.

	$\phi^{(R)}$	ϕ^{Ber}	ϕ^{Is}	ϕ^{MEDF}	ϕ^{all}
R=1	0.379	0.379	0.312	1.211	0.309
R=2	0.00883	0.299871	0.256671	0.257068	0.0075
R=3	-0.001	0.250736	0.215422	0.200534	0.0001

We observe that our procedure recovers the fact that the range-R potential $\phi^{(R)}$ is the best to approximate the empirical measure, in the sense that it minimizes the KL divergence and that it has the minimal number of terms (ϕ^{all} does as good as $\phi^{(R)}$ for the KL divergence but it contains more monomials whose coefficient (almost) vanish in the estimation).

[26] or equivalently, a **RPTD**-1 from (37)

,,

4.3.4 Synaptic plasticity.

Here the neural network with dynamics given by (39) has been submitted to the STDP rule (34). The goal is to check the validity of the statistical model given by (35), predicted in [12]. We use spike-trains of length $T = 10^7$ from a simulated network with $N = 10$ neurons.

Previous numerical explorations of the noiseless case, $\sigma_B = 0$, have shown [10, 13] that a network of N such neurons, with fully connected graph, where synapses are taken randomly from a distribution $\mathscr{N}(0, \frac{C^2}{N})$, where C is a control parameter, exhibits generically a dynamics with very large periods in determined regions of the parameters-space (γ, C). On this basis, we choose; $N = 10$, $\gamma = 0.995$, $C = 0.2$. The external current $\mathbf{I}^{(ext)}$ in eq. (39) is given by $I_i^{ext} = 0.01$ while $\sigma_B = 0.01$. Note that fixing a sufficiently large average value for this current avoids a situation where neurons stops firing after a certain time ("neural death").

We register the activity after 4000 steps of adaptation with the STPD rule proposed in (34). In this context we expect the potential for the whole population to be of the form (35) and for a subset of the population of the form (36). Therefore, we choose randomly 2 neurons among the N and we construct from them the prefix-tree. Then, for the 2 neuron potentials forms from (37), we estimate the coefficients that minimizes the Kullback-Leibler divergence.The probability of words of different sizes predicted by several statistical models from (37) versus empirical probability $\pi_\omega^{(T)}(w)$ obtained from a spike train and the corresponding $\tilde{h}$ value of the estimation process for a fixed pair of neurons are shown on figure (4.3.4).
Results depicted on figure (4.3.4) show, on one hand, that the statistics is well fitted by (36). Moreover, the best statistical models, are those including rate terms (the differences between their KL value is two orders of magnitude smaller that within those not disposing of rate terms). We also note that for the words with the smallest probability values, the potential do not yields a perfect matching due to finite size effects (see fig (4.3.4)). Especially, the small number of events due to low firing rates of neurons makes more sensitive the relation between the length of observed sequences (word size) and the spike-train length necessary to provide a good sampling and hence a reliable empirical probability.

4.3.5 Additional tests: the non-stationary case

Here we present results of the parameter estimation method applied to a spike train with statistics governed by a non-stationary statistical model of range 1, i.e. with time varying coefficients for rate or synchronization terms. Since the generation of spike-trains corresponding to more general higher time-order non-stationary process is not trivial, these potentials with higher range values will be analyzed in a forthcoming paper.

In the following we use an Ising potential form (37) with time-varying coefficients $\psi = (\omega) = \lambda_1(t)\,\omega_1(0) + \lambda_2(t)\,\omega_2(0) + \lambda_3(t)\,\omega_1(0)\,\omega_2(0)$.. The procedure to generate a non stationary spike-train of length T is the following. We fix a time dependent form for the 3 coefficients $\lambda_i(t)$. From the initial value of the $\lambda_i's$ (say at time t) we compute the invariant measure of the RPF operator. From this, we draw a Chapman-

,,

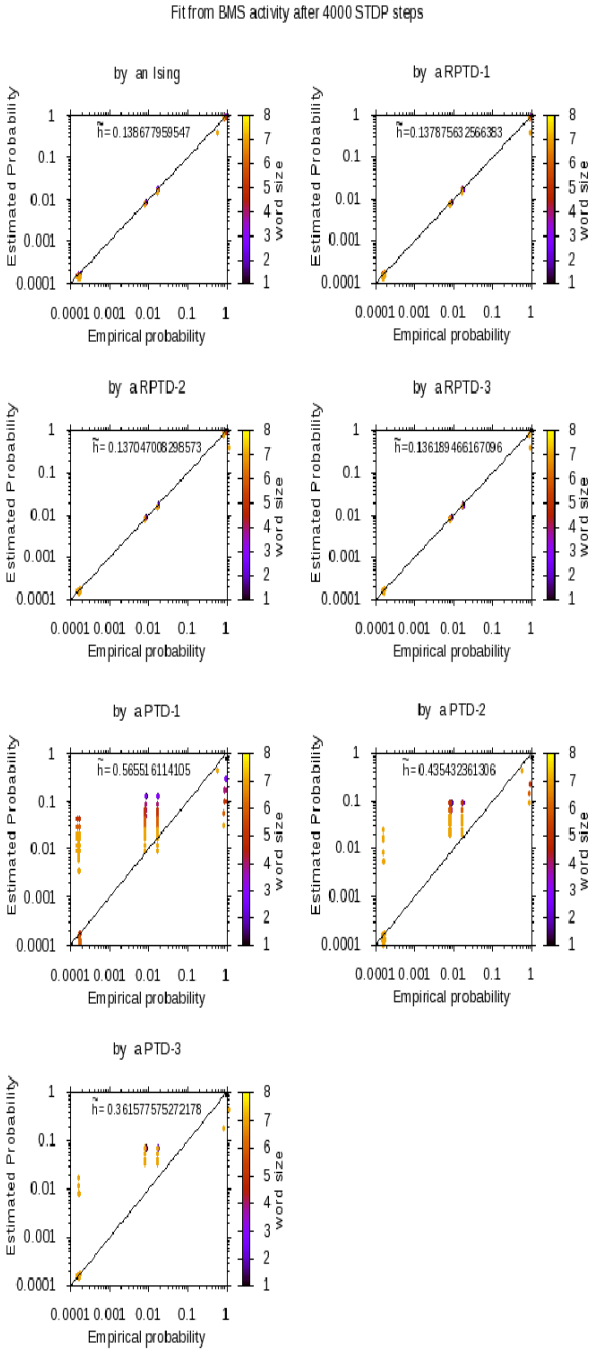

Figure 6: The probability of words of different sizes predicted by several statistical models from (37) versus empirical probability $\pi_{\omega}^{(T)}(w)$ obtained from a spike train generated by dynamics (39) after 4000 epochs of adaptation.The $\tilde{h}$ value (29) for each fitting model is shown inside the graphic. The potential is a pair potential of the form (36). Recall that **RPTD** Models include firing rates but **PTD** models do not.

,,

Kolmogorov equation (20) with a time dependent RPF operator computed using the next coefficient values $\lambda_i(t+1)$.

With the generated spike-train, we perform the parameter estimation, but computing the empirical average over an small fraction of it which means a time window of size $T_0 = \frac{T}{M} << T$. Then, we slide the observation window and estimate again the coefficients value. We have verified that estimation procedure can recover correctly the coefficient values, for several types of time dependence, provided their variations be not too fast, and that the sliding window size be not too large with respect to T. We present the reconstruction of the parameters with a sinusoidal time-dependence given by $\lambda_0(t) = 0.4 + 0.3\sin\left(\frac{4\pi t}{T-T_0}\right)$.

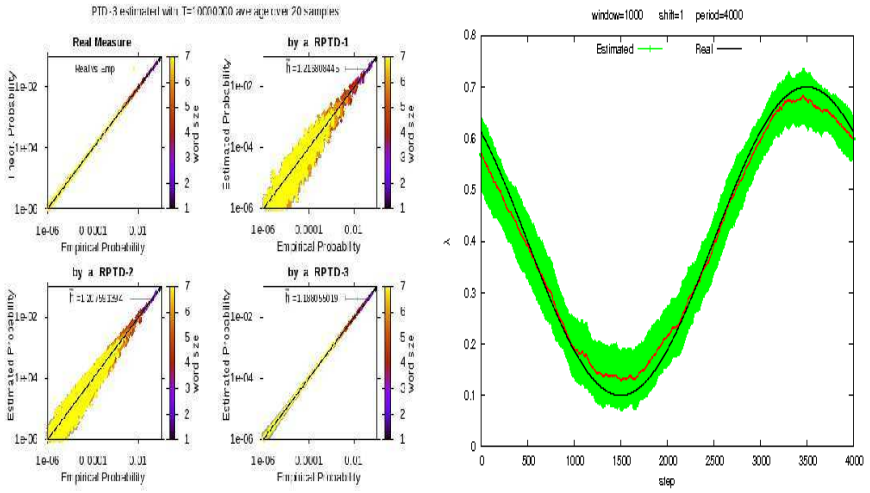

Figure 7: **Estimation of coefficients on a Non-Stationary process genated by an Ising model and sinousoidal time dependence**. Real value(black) and estimated parameter with its error bars (green) computed over 20 trials. The time shift is $\tau = 1$, Window size is fixed 1000, but oscillation period corresponds to 2000 (left) and 4000 (right).

5 Discussion and conclusion

5.1 Comparison with existing methods

Let us first summarize the advantages and drawbacks of our method compared with the existing ones. For this, we list some keywords in the approaches used by the community and discuss the links with our own work.

- **Maximum entropy.** The formalism that we use corresponds to a maximum entropy method but without limitations on the number or type of contraints. Actually, on mathematical grounds, it allows infinitely many constraints. Moreover, we do not need to compute the entropy.

,,

- **Markovian approaches.** Our method is based on a Markovian approach where the memory depth of the Markov chain can be arbitrary long (actually the formalism that we use allows to theoretically consider processes with infinite memory, called *chains with complete connections [43]*, see [11] for an application to spike train statistics). As we developed, the link between the potential extracted from the maximum entropy principle, by fixing *ad hoc* observables, and a Markov chain is not straightforward, since a potential of this kind is not normalized.

- **Monte-Carlo methods.** Equation (21) allows us to generate spike trains Gibbs-distributed with and arbitrary potential (non normalized). The convergence is ensured by eq. (14). We emphasize that we do not need to assume detailed balance. Instead, we impose a technical assumption (primitivity of the Ruelle-Perron-Frobenius matrix) which is more general than detailed balance. On the opposite, if this assumption does not hold then the unicity of the Gibbs distribution is not guarantee and, in this case, the determination of spike train statistics from empirical data becomes even more cumbersome.

- **Determining an effective synaptic connectivity between neurons.** Interactions between neurons occur via synapses (or gap junction). This interaction is not instantaneous, it requires some delay. As a matter of fact, estimating the synaptic conductances via the spike statistics requires therefore to consider time-dependent potentials. Our formalism allows this. Determining an effective synaptic connectivity between neurons from spike trains will be the subject of a forthcoming paper.

- **Boltzmann learning.** Our approach can be viewed as "Boltzmann learning" (as presented e.g. in [61]) without restrictions on the parameters that we learn, without using a Monte Carlo approach (which assumes detailed balance), and uses a criterion which is strictly convex.

- **Performances.** At its current implementation level, the proposed method allows us to analyze the statistics of small groups (up to 8/12) of neurons. The parametric statistical potential of Markov processes up to range 16/20 is calculable, thus considering up to 2^{20} states for the process. The implementation considers several well-established numerical methods, in order to be applicable to a large set of possible data. With respect to the state of the art, this method allows us to consider non-trivial statistics (e.g. beyond rate models and even models with correlation), thus targeting models with complex spike patterns. This method is in a sense the next step after Ising models, known as being able to represent a large but limited part of the encoded information (e.g. [66, 48]). Another very important difference with respect to other current methods is that we perform the explicit variational optimization of a well defined quantity, i.e., the KL-divergence between the observed and estimated distributions. The method proposed here does not rely on Monte Carlo Markov Chain methods but on a spectral computation based on the RPF operator, providing exact formula, while the spectral characteristics are easily obtained from standard numerical methods.

,,

The main drawback of our method is that it *does not allow to treat a large number of neurons and simultaneously a large range*. This is due to the evident fact that the number of monomials combinatorically increases as N, R growth. However, this is not a problem intrinsic to our approach but to parametric estimations potentials of the form (6). We believe that other form of potential could be more efficient (see [11] for an example). We also want to emphasize that, when considering Ising like statistics our algorithm is *less performant* than the existing ones (although improvements in speed and memory capacity thanks to the use of parallel computation algorithms remain an open and natural developpement path), for the simple reason that the latter has been developed and optimized using the tremendous results existing in statistical physics, for spins systems. Their extensions to models of the general form (6) seems rather delicate, as suggested by the nice work in [44] where extension between the 1-step Markov case is already cumbersome.

- **Mean-field methods.** Mean-field methods aim at computing the average value of observables ("order parameters") relevant for the characterisation of statistical properties of the system. Typical examples are magnetisation in ferromagnetic models (corresponding to rates in spiking neurons models), but more elaborated order parameters are known e.g. in spin glasses [47] or in neural networks [75]. Those quantities obey equations (usually called mean-field equations) which are, in most cases, not explicitely solvable. Therefore, approximations are proposed from the simplest (naive mean-field equations) to more complex estimations, with significant results developed in the realm of spins systems (Ising model, Sherrington-Kirckpatrick spin glass model [71]). Examples are the replica method [47], Thouless-Anderson-Palmer equations [79], the Plefka expansion [77], or more recently e.g. the Sessak-Monasson approximation [70] (for a recent review on mean-field methods see [52]). Since the seminal paper by Schneidman and collaborators [67] they have also been applied to spike trains statistics analysis assuming that neurons dynamics generates a spike statistics characterized by a Gibbs distribution with an Ising Hamiltonian. In their most common form these methods do not consider dynamics (e.g time correlations) and their extension to the time-dependent case (e.g. dynamic mean-field methods) is far from being straightforward (see e.g. [76, 75, 3, 65, 24] for examples of such developments). Moreover, exact mean-field equations and their approximations usually only provide a probability measure at positive distance to the true (stationary) probability measure of the system (this distance can be quantified in the setting of information geometry using e.g. the KL distance [2]). This is the case whenever the knowledge of the sought order parameters is not sufficient to determine the underlying probability.

  The present work can, in some sense, be interpreted in the realm of mean-field approaches. Indeed, we are seeking an hidden Gibbs measure and we have only information about the average value of ad hoc observables. Thus, equation (17) is a mean-field equation since it provides the average value of an observable with respect to the Gibbs distribution. There are therefore L such equations, where L is the number of monomials in the potential ψ. Are all these equations relevant

”

? If not, which one are sufficient to determine univoquely the Gibbs distribution ? Which are the order parameters ? The method consisting of providing a hierarchy of mean-field approximations which starts with the Bernoulli model (all monomials but the rate terms are replaced by a constant), then Ising (all monomials but rate and spatial correlations are replaced by a constant), while progressively diminishing the KL divergence allows to answer the question of the relevant order parameters and can be interpreted as well in the realm of information geometry. This hierarchical approach is a strategy to cope with the problem of combinatorial explosion of terms in the potential when the number of neurons or range increases. But the form of potential that we consider does not allow a straightforward application of the methods inherited from statistical mechanics of spin systems. As a consequence, we believe that instead of focusing too much on these methods it should be useful to adopt technics based on large deviations (which actually allows the rigorous fundation of dynamic mean field methods for spin-glasses [3] and neural networks [65, 24]). This is what the present formalism offers.

5.2 Conclusion and perspectives

The thermodynamic formalism allows us to provide closed-form calculations of interesting parameters related to spectral properties of the RPF operator. We, for instance, propose an indirect estimation of the entropy, via an explicit formula. We also provide numbers for the average values of the related observable, probability measure, etc.. This means that as soon as we obtain the numerical values of the Gibbs distribution up to some numerical precision, all other statistical parameters come for free without additional approximations.

A step further, the non-trivial but very precious virtue of the method is that it allows us to efficiently compare models. We thus not only estimate the optimal parameters of a model, but can also determine among a set of models which model is the most relevant. This means, for instance, that we can determine if either only rates, or rates and correlations matters, for a given piece of data. Another example is to detect if a given spike pattern is significant, with respect to a model not taking this pattern into account. The statistical significance mechanism provides numbers that are clearly different for models corresponding or not to a given empirical distribution, providing also an absolute test about the estimation significance. These elements push the state of the art regarding statistical analysis of spike train a step further.

At the present state of the art, the present method is limited by three bounds.

First of all, the formalism is developed for a stationary spike-train, i.e. for which the statistical parameters are constant. This is indeed a strong limitation, especially in order to analyze biological data, though several related approaches consider the same restrictive framework. This drawback is overcome at two levels. At the implementation level we show here how using a sliding estimation window and assuming an adiabatic, i.e. slowly varying, distribution we still can perform some relevant estimation. In a nutshell, the method seems still usable and we are now currently investigating this on both simulated and biological data, this being another study on its own. At a more

,,

theoretical level, we are revisiting the thermodynamic formalism developed here for time varying parameters (in a similar way as the so called inhomogeneous Poisson process with time varying rates). Though this yields non-trivial developments beyond the scope of this work, it seems that we can generalize the present formalism in this direction.

Secondly, the present implementation has been optimized for dense statistical distributions, i.e., in the case where almost all possible spike combinations are observed. Several mechanisms, such as look-up tables, make this implementation very fast. However, if the data is sparse, as it may be the case for biological, a dual implementation has to be provided using data structure, such as associative tables, well adapted to the fact that only a small amount of possible spike combinations are observed. This complementary implementation has been made available and validated against the present one. This is going to analyze sparse Markov processes up to range much higher than 16/20. Again this is not a trivial subject and this aspect must be developed in a next study as well as the applicability of parallel computing alternatives (e.g. sparse matrix storage, parallel fast-eigenvalue algorithms, etc.).

Finally, given an assembly of neurons, every statistical tools available today provide only the analysis of the statistics a small subset of neurons, and it is known that this only partially reflects the behavior of the whole population [40]. The present method for instance, is difficult to generalize to more than 8/10 neurons because of the incompressible algorithmic complexity of the formalism although parallel computation techniques might be helpful. However, the barrier is not at the implementation level, but at the theoretical level, since effective statistical general models (beyond Ising models) allow for instance to analyze statistically large spiking patterns such as those observed in synfire chains [30] or polychronism mechanisms [54]. This may be the limit of the present class of approaches, and things are to be thinked differently. We believe that the framework of thermodynamic formalism and links to Statistical Physics is still a relevant source of methods for such challenging perspectives.

acknowledgements

We are grateful to F. Grammont, C. Malot, F. Delarue and P. Reynaud-Bourret for helpful discussions and Adrien Palacios for his precious remarks and profound scientific questions at the origin of main aspects of the present work. Partially supported by the ANR MAPS & the MACACC ARC projects and PhD.D-fellowship from Research Ministry to J.C Vasquez.

References

[1] Moshe Abeles and George L. Gerstein. Detecting spatiotemporal firing patterns among simultaneously recorded single neurons. *Journal of Neurophysiology*, 60(3):909–924, September 1988.

[2] Shun-Ichi Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford Univ. Press., 2000.

[3] G. BenArous and A. Guionnet. Large deviations for langevin spin glass dynamics. *Probability Theory and Related Fields*, 102:455–509, 1995.

[4] G. Bi and M. Poo. Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, 24:139–166, March 2001.

”

[5] R. Bowen. *Equilibrium states and the ergodic theory of Anosov diffeomorphisms. Second revised version.*, volume 470 of *Lect. Notes.in Math.* Springer-Verlag, 2008.

[6] P. C. Bressloff and S. Coombes. *Synchronization of synaptically-coupled neural oscillators in: Epilepsy as a dynamic disease*, chapter 7. J.Milton and P. Jung, Springer-Verlag, 2003.

[7] D. R. Brillinger. Maximum likelihood analysis of spike trains of interacting nerve cells. *Biol Cybern*, 59(3):189âĂŞ200, 1988.

[8] Emery N. Brown, Robert E. Kass, and Partha P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7(5):456–461, May 2004.

[9] B. Cessac. Does the complex susceptibility of the hénon map have a pole in the upper-half plane ? a numerical investigation. *Nonlinearity*, 20:2883–2895, dec 2007.

[10] B. Cessac. A discrete time neural network model with spiking neurons. i. rigorous results on the spontaneous dynamics. *J. Math. Biol.*, 56(3):311–345, 2008.

[11] B. Cessac. A discrete time neural network model with spiking neurons ii. dynamics with noise. *J.Math. Biol., accepted*, 2010.

[12] B. Cessac, H. Rostro-Gonzalez, J.C. Vasquez, and T. Viéville. How gibbs distribution may naturally arise from synaptic adaptation mechanisms: a model based argumentation. *J. Stat. Phys*, 136(3):565–602, August 2009.

[13] B. Cessac and T. Viéville. On dynamics of integrate-and-fire neural networks with adaptive conductances. *Frontiers in neuroscience*, 2(2), jul 2008.

[14] J.R. Chazottes. *Entropie Relative, Dynamique Symbolique et Turbulence*. PhD thesis, Université de Provence - Aix Marseille I, 1999.

[15] J.R. Chazottes, E. Floriani, and R. Lima. Relative entropy and identification of gibbs measures in dynamical systems. *J. Statist. Phys.*, 90(3-4):697–725, 1998.

[16] J.R. Chazottes and G. Keller. *Pressure and Equilibrium States in Ergodic Theory*, chapter Ergodic Theory. Encyclopedia of Complexity and System Science, Springer, 2009. to appear.

[17] Sung Nok Chiu and Kwong Ip Liu. Generalized cramér-von mises goodness-of-fit tests for multivariate distributions. *Computational Statistics and Data Analysis*, 53(11):3817–3834, 2009.

[18] E. S. Chornoboy, L. P. Schramm, and A. F. Karr. Maximum likelihood identification of neural point process systems. *Biol Cybern*, 59(4-5):265âĂŞ275, 1988.

[19] Simona Cocco, Stanislas Leibler, and Rémi Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences of the United States*, 106(33):14058–14062, August 2009.

[20] Noel Cressie and Timothy R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):440–464, 1984.

[21] Imre Csiszar. Sanov property, generalized *i*-projection and a conditional limit theorem. *Ann. Prob*, 12(3):768–793, 1984.

[22] Amir Dembo and Ofer Zeitouni. *Large Deviation Techniques and Applications*. Springer, 1993.

[23] M. Diesmann, M-O. M-O Gewaltig, and A. Aertsen. Stable propagation of synchronous spiking in cortical neural networks. *Nature*, 402:529–533, 1999.

[24] O. Faugeras, J. Touboul, and B. Cessac. A constructive mean field analysis of multi population neural networks with random synaptic weights and stochastic inputs. *Frontiers in Neuroscience*, 2008. submitted.

[25] Michael Frigge, David C Hoaglin, and Boris Iglewicz. Implementation of the boxplot. *The American Statistician*, 43(1):50–54, February 1989.

[26] Yun Gao, Ioannis Kontoyiannis, and Elie Bienenstock. Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, 10(2):71–99, 2008.

[27] I.I. Gikhman and A.V. Skorokhod. *The Theory of Stochastic Processes*. Springer, 1979.

”

[28] F. Grammont and A. Riehle. Precise spike synchronization in monkey motor cortex involved in preparation for movement. *Exp. Brain Res.*, 128:118–122, 1999.

[29] Peter Grassberger. Estimating the information content of symbol sequences and efficient codes. *IEEE Transactions on Information Theory*, 35, 1989.

[30] J. Hertz. *Theoretical Aspects of Neural Computation.*, chapter Modelling synfire processing., pages 135–144. Wong K-Y M. King I. and Yeung D-Y (eds), Springer-Verlag, 1997.

[31] A.L. Hodgkin and A.F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117:500–544, 1952.

[32] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.

[33] E. Izhikevich. Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, 14(6):1569–1572, 2003.

[34] E.T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620, 1957.

[35] Don H. Johnson and Ananthram Swami. The transmission of signals by auditory-nerve fiber discharge patterns. *J. Acoust. Soc. Am*, 74(2):493âĂŞ501, August 1983.

[36] R. E. Kass and V. Ventura. A spike-train probability model. *Neural Comput.*, 13(8):1713âĂŞ1720, 2001.

[37] Robert E. Kass, Valrie Ventura, and Emery N. Brown. Statistical issues in the analysis of neuronal data. *Journal of Neurophysiology*, 94(1):8–25, January 2005.

[38] G. Keller. *Equilibrium States in Ergodic Theory*. Cambridge University Press, 1998.

[39] Kathryn Laskey and Laura Martignon. Bayesian learning of loglinear models for neural connectivity. In *Proceedings of the Proceedings of the Twelfth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-96)*, pages 373–380, San Francisco, CA, 1996. Morgan Kaufmann.

[40] P.E. Latham, A. Roth, M. Hausser, and M. London. Requiem for the spike? *Soc. Neurosc. Abstr.*, 32, 2006.

[41] BG Lindsey, KF Morris, R Shannon, and GL Gerstein. Repeated patterns of distributed synchrony in neuronal assemblies. *Journal of Neurophysiology*, 78:1714–1719, 1997.

[42] Michael London, Adi Shcreibman, and Idan Segev. Estimating information theoretic quantities of spike-trains using the context tree weighting algorithm. *Nature neuroscience*, 5, 2002. Appendix to: The information efficacy of a synapse.

[43] G. Maillard. *Introduction to chains with complete connections*. Ecole Federale Polytechnique de Lausanne, winter 2007.

[44] Olivier Marre, Sami El Boustani, Yves Frégnac, and Alain Destexhe. Prediction of spatiotemporal patterns of neural activity from pairwise correlations. *Physical Review Letters*, 102(13):4, April 2009.

[45] Laura Martignon, Gustavo Deco, Kathryn Laskey, Mathiew Diamond, Winrich Freiwald, and Eilon Vaadia. Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Computation*, 12(11):2621–2653, November 2000.

[46] Laura Martignon, H. von Hasseln, S. Grün, A. Aertsen, and G. Palm. Detecting higher-order interactions among the spiking events in a group of neurons. *Biological Cybernetics*, 73(1):69–81, July 1995.

[47] M. Mézard, G. Parisi, and M.A. Virasoro. *Spin-glass theory and beyond*. World scientific Singapore, 1987.

[48] Marc Mezard and Thierry Mora. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris*, 103(1–2):107–113, January–March 2009.

[49] Gusztáv Morvai and Benjamin Weiss. Estimating the lengths of memory words. *IEEE TRANSACTIONS ON INFORMATION THEORY*, 54(8):3804–3807, august 2008.

[50] A.V. Negaev. An asymptotic formula for the neyman-pearson risk in discriminating between two markov chains. *Journal of Mathematical Sciences*, 111(3):3582–3591, 2002.

[51] Murat Okatan, Matthew A.Wilson, and Emery N. Brown. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural Computation*, 17(9):1927âĂŞ1961, September 2005.

”

[52] Manfred Opper and David Saad. *Advanced Mean Field Methods: Theory and Practice*. MIT. Press., 2001.

[53] Leslie C. Osborne, Stephanie E. Palmer, Stephen G. Lisberger, and William Bialek. The neural basis for combinatorial coding in a cortical population response. *The Journal of Neuroscience*, 28(50):13522–13531, December 2008.

[54] Hélène Paugam-Moisy, Régis Martinez, and Samy Bengio. Delay learning and polychronization for reservoir computing. *Neurocomputing*, 71:1143–1158, 2008.

[55] Alexandre Pouget and Gregory C DeAngelis. Paying attention to correlated neural activity. *Nature Neuroscience*, 11(12):1371–1372, December 2008.

[56] C. Pouzat and A. Chaffiol. On goodness of fit tests for models of neuronal spike trains considered as counting processes. *http://arxiv.org/abs/0909.2785v1*, 2009.

[57] Christophe Pouzat and Antoine Chaffiol. Automatic spike train analysis and report generation. an implementation with r, r2html and star. *J Neurosci Methods*, 181:119–144, 2009.

[58] Christophe Pouzat and Antoine Chaffiol. On goodness of fit tests for models of neuronal spike trains considered as counting processes, 2009.

[59] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. Bradford Books, 1997.

[60] Yasser Roudi, Erik Aurell, and John A Hertz. Statistical physics of pairwise probability models. *Frontiers in Computational Neuroscience*, page 15, 2009.

[61] Yasser Roudi, Joanna Tyrcha, and John A Hertz. Ising model for neural data:model quality and approximate methods for extracting functional connectivity. *Physical Review E*, page 051915, 2009.

[62] D. Ruelle. Statistical mechanics of a one-dimensional lattice gas. *Commun. Math. Phys.*, 9:267–278, 1968.

[63] D. Ruelle. *Statistical Mechanics: Rigorous results*. Benjamin, New York, 1969.

[64] D. Ruelle. *Thermodynamic formalism*. Addison-Wesley,Reading, Massachusetts, 1978.

[65] M. Samuelides and B. Cessac. Random recurrent neural networks. *European Physical Journal - Special Topics*, 142:7–88, 2007.

[66] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply string correlated network states in a neural population. *Nature*, 440:1007– 1012, 2006.

[67] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[68] Thomas Schürmann and Peter Grassberger. Entropy estimation of symbol sequences. *Chaos*, 6(3):414–427, 1996.

[69] R. Segev, I. Baruchi, E. Hulata, and E. Ben-Jacob. Hidden neuronal correlations in cultured networks. *Physical Review Letters*, 92:118102, 2004.

[70] sessak V. and Remi Monasson. Small-correlation expansions for the inverse ising problem. *J. Phys. A*, 42:055001, 2009.

[71] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Physical Review Letters*, 35(26):1792+, December 1975.

[72] Jonathon Shlens, Greg D. Field, Jeffrey L. Gauthier, Martin Greschner, Alexander Sher, Alan M. Litke, and E. J. Chichilnisky. The structure of large-scale synchronized firing in primate retina. *The Journal of Neuroscience*, 29(15):5022–5031, April 2009.

[73] Jonathon Shlens, Greg D. Field, Jeffrey L. Gauthier, Matthew I. Grivich, Dumitru Petrusca, Alexander Sher, Alan M. Litke, and E. J. Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *The Journal of Neuroscience*, 26(32):8254–8266, August 2006.

[74] Ya Sinai. Gibbs measures in ergodic theory. *Russ. Math. Surveys*, 27(4):21–69, 1972.

[75] H. Sompolinsky, A. Crisanti, and H.J. Sommers. Chaos in random neural networks. *Phys. Rev. Lett.*, 61:259–262, 1988.

,,

[76] H. Sompolinsky and A. Zippelius. Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses. *Physical Review B*, 25(11):6860–6875, 1982.

[77] Plefka T. Convergence condition of the tap equations for the infinite-ranged ising spin glass model. *J. Phys. A*, 15:1971, 1982.

[78] Aonan Tang1, David Jackson, Jon Hobbs, Wei Chen, Jodi L. Smith, Hema Patel, Anita Prieto, Dumitru Petrusca, Matthew I. Grivich, Alexander Sher, Pawel Hottowy, Wladyslaw Dabrowski, Alan M. Litke, and John M. Beggs. A maximum entropy model applied to spatial and temporal correlations from cortical networks *In Vitro*. *The Journal of Neuroscience*, 28(2):505–518, January 2008.

[79] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of âĂŸ solvable model of a spin glass. *Philos. Mag.*, 35:593–601, 1977.

[80] Gašper Tkačik, Elad Schneidman, Michael J. Berry II, and William Bialek. Ising models for networks of real neurons. *arXiv*, page 4, 2006.

[81] Wilson Truccolo and John P. Donoghue. Nonparametric modeling of neural point processes via stochastic gradient boosting regression. *Neural Computation*, 19(3):672–705, 2007.

[82] Wilson Truccolo, Uri T. Eden, Matthew R. Fellows, John P. Donoghue, and Emery N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble and extrinsic covariate effects. *J Neurophysiol*, 93:1074–1089, 2005.

[83] Alessandro E. P. Villa, , Igor V. Tetko, Brian Hyland, and Abdellatif Najem. Spatiotemporal activity patterns of rat cortical neurons predict responses in a conditioned task. *Proc Natl Acad Sci USA*, 96(3):1106–1111, 1999.

Description

The following paper has been published at Journal of Statistical Physics. (3), 565-602 (2009).

" How Gibbs distributions may naturally arise from synaptic adaptation mechanisms. A model-based argumentation.

B. Cessac *, †H. Rostro †, J.C. Vasquez †, T. Viéville †

January 17, 2011

Abstract

This paper addresses two questions in the context of neuronal networks dynamics, using methods from dynamical systems theory and statistical physics: (i) How to characterize the statistical properties of sequences of action potentials ("spike trains") produced by neuronal networks ? and; (ii) what are the effects of synaptic plasticity on these statistics ? We introduce a framework in which spike trains are associated to a coding of membrane potential trajectories, and actually, constitute a symbolic coding in important explicit examples (the so-called gIF models). On this basis, we use the thermodynamic formalism from ergodic theory to show how Gibbs distributions are natural probability measures to describe the statistics of spike trains, given the empirical averages of prescribed quantities. As a second result, we show that Gibbs distributions naturally arise when considering "slow" synaptic plasticity rules where the characteristic time for synapse adaptation is quite longer than the characteristic time for neurons dynamics.

Keywords Neurons dynamics, spike coding, statistical physics, Gibbs distributions, Thermodynamic formalism.

1 Introduction.

Spike trains as a "neural code". Neurons activity results from complex and nonlinear mechanisms (Hodgkin & Huxley, 1952; Cronin, 1987; Dayan & Abbott, 2001; Gerstner & Kistler, 2002b), leading to a wide variety of dynamical behaviours (Cronin, 1987; Guckenheimer & Labouriau, 1993). This activity is revealed by the emission of action potentials or "spikes". While the shape of an action potential is essentially always the same for a given neuron, the succession of spikes emitted by this neuron can have a wide variety of patterns (isolated spikes, periodic spiking, bursting, tonic spiking, tonic bursting, etc ...) (Izhikevich, 2004; Brette & Gerstner, 2005; Touboul, 2008), depending on physiological parameters, but also on excitations coming either from other neurons or from external inputs. Thus, it seems natural to consider

*Laboratoire J. A. Dieudonné, U.M.R. C.N.R.S. N 6621, Université de Nice Sophia-Antipolis,France
†INRIA, 2004 Route des Lucioles, 06902 Sophia-Antipolis, France.

spikes as "information quanta" or "bits" and to seek the information exchanged by neurons in the structure of spike trains. Doing this, one switches from the description of neurons in terms of membrane potential dynamics, to a description in terms of spike trains. This point of view is used, in experiments, by the analysis of *raster plots*, i.e. the activity of a neuron is represented by a mere vertical bar each time this neuron emits a spike. Though this change of description raises many questions, it is commonly admitted in the computational neuroscience community that spike trains constitute a "neural code".

This raises however other questions. How is "information" encoded in a spike train: rate coding (Adrian & Zotterman, 1926), temporal coding (Theunissen & Miller, 1995), rank coding (Perrinet, Delorme, Samuelides, & Thorpe, 2001; Delorme, Perrinet, & Thorpe, 2001), correlation coding (Johnson, 1980) ? How to measure the information content of a spike train ? There is a wide literature dealing with these questions (Nirenberg & Latham, 2003; D. Johnson, 2004; Barbieri et al., 2004; Nemenman, Lewen, Bialek, & Steveninck, 2006; Arabzadeh, Panzeri, & Diamond, 2006; Sinanović & Johnson, 2006; Gao, Kontoyiannis, & Bienenstock, 2008; Osbone, Palmer, Lisberger, & Bialek, 2008), which are inherently related to the notion of *statistical characterizations* of spike trains, see (Rieke, Warland, Steveninck, & Bialek, 1996; Dayan & Abbott, 2001; Gerstner & Kistler, 2002b) and references therein for a review. As a matter of fact, a prior to handle "information" in a spike train is the definition of a suitable probability distribution that matches the empirical averages obtained from measures. Thus, in some sense that we make precise in this paper, the choice of a set of quantities to measure (observables) constrains the form of the probability characterizing the statistics of spike trains.

As a consequence, there is currently a wide debate on the canonical form of these probabilities. While Poisson statistics, based on the mere knowledge of frequency rates, are commonly used with some success in biological experiments (Georgopoulos, Kalaska, Caminiti, & Massey, 1982; Georgeopoulos, Merchant, Naselaris, & Amirikian, 2007), other investigation evidenced the role of spikes coincidence or correlations (Grammont & Riehle, 1999, 2003) and some people have proposed non Poisson probabilities, such as Ising-like Gibbs distributions, to interpret their data (Schneidman, Berry, Segev, & Bialek, 2006; Tkacik, Schneidman, Berry, & Bialek, 2006). It is important to note here that beyond the determination of the "right" statistical model there is the aim of identifying which type of information is used by the brain to interpret the spike trains that it receives, coming from different sources. As an example choosing a model where only frequency rates matters amounts to assuming that spikes coming from different neurons are essentially treated independently.

However, determining the form of the probability characterizing the statistics of spike trains is extremely difficult in real experiments. In the present paper, we focus on neural networks *models* considered as dynamical systems, with a good mathematical and numerical control on dynamics. In this context we argue that Gibbs distributions are indeed natural candidates whenever a set of quantities to measure (observables) has been prescribed. As a matter of fact Poisson distributions and Ising-like distributions are specific examples, but, certainly, do not constitute the general case.

Synaptic plasticity. The notion of neural code and information cannot be separated from the capacity of neuronal networks to evolve and adapt by *plasticity* mechanisms, and especially

synaptic plasticity. The latter occurs at many levels of organization and time scales in the nervous system (Bienenstock, Cooper, & Munroe, 1982). It is of course involved in memory and learning mechanisms, but it also alters excitability of brain area and regulates behavioural states (e.g. transition between sleep and wakeful activity). Therefore, understanding the effects of synaptic plasticity on neurons dynamics is a crucial challenge. On experimental grounds, different synaptic plasticity mechanisms have been exhibited from the Hebbian's ones (Hebb, 1949) to Long Term Potentiation (LTP) and Long Term Depression (LTD), and more recently to Spike Time Dependent Plasticity (STDP) (Markram, Lübke, Frotscher, & Sakmann, 1997; Bi & Poo, 2001) (see (Dayan & Abbott, 2001; Gerstner & Kistler, 2002a; Cooper, Intrator, Blais, & Shouval, 2004) for a review). Modeling these mechanisms requires both a bottom-up and top-down approach.

This issue is tackled, on theoretical grounds, by inferring "synaptic updates rules" or "learning rules" from biological observations (Malsburg, 1973; Bienenstock et al., 1982; Miller, Keller, & Stryker, 1989) and extrapolating, by theoretical or numerical investigations, the effects of such synaptic rule on such neural network *model*. This bottom-up approach relies on the belief that there are "canonical neural models" and "canonical plasticity rules" capturing the most essential features of biology. Unfortunately, this results in a plethora of canonical "candidates" and a huge number of papers and controversies. In an attempt to clarify and unify the overall vision, some researchers have proposed to associate learning rules and their dynamical effects to general principles, and especially to"variational" or "optimality" principles, where some functional has to be maximised or minimised (Dayan & Hausser, 2004; Rao & Sejnowski, 1991, 2001; Bohte & Mozer, 2007; Chechik, 2003; Toyoizumi, Pfister, Aihara, & Gerstner, 2005, 2007). Therefore, in these "top-down" approaches, plasticity rules "emerge" from first principles. Unfortunately, in most examples, the validations of these theories has been restricted to considering isolated neurons submitted to input spike trains with ad hoc statistics (typically, Poisson distributed with independent spikes (Toyoizumi et al., 2005, 2007)).

However, addressing the effect of synaptic plasticity in neuronal networks where dynamics is *emerging* from collective effects and where spikes statistics are *constrained* by this dynamics seems to be of central importance. This is the point of view raised in the present paper, where, again, we focus on models, which are simplifications of real neurons. Even in this case, this issue is subject to two main difficulties. On one hand, one must identify the generic collective dynamical regimes displayed by the model for different choices of parameters (including synaptic weights). On the other hand, one must analyse the effects of varying synaptic weights when applying plasticity rules. This requires to handle a complex interwoven evolution where neurons dynamics depends on synapses and synapses evolution depends on neuron dynamics. The first aspect has been addressed by several authors using mean-field approaches (see e.g. (Samuelides & Cessac, 2007) and references therein), "Markovian approaches" (Soula & Chow, 2007), or dynamical system theory (see (Cessac & Samuelides, 2007) and references therein). The second aspect has, up to our knowledge, been investigated theoretically in only a few examples with Hebbian learning (Daucé, Quoy, Cessac, Doyon, & Samuelides, 1998; Siri, Berry, Cessac, Delord, & Quoy, 2007, 2008) or discrete time Integrate and Fire models with an STDP like rule (Soula, 2005; Soula, Beslon, & Mazet, 2006) and is further addressed here.

What the paper is about. To summarize the previous discussion the study of neuronal networks at the current stage is submitted to two central questions:

"• How to characterize the statistics of spike trains in a network of neurons ?

- How to characterize the effects of synaptic plasticity on this network dynamics and especially on spike trains statistics ?

In this paper we suggest that these two questions are closely entangled and must both be addressed in the same setting. In order to have a good control on the mathematics we mainly consider the case of neural networks models where one has a full characterization of the generic dynamics (Cessac, 2008; Cessac & Viéville, 2008). Thus, our aim is not to provide general statements about biological neural networks. We simply want to have a good mathematical control of what is going on in specific models, with the hope that this analysis should shed some light on what happens (or *does not* happen) in "real world" neural systems. However, though part of the results used here are rigorous, this work relies also on "working assumptions" that we have not been able to check rigorously. These assumptions, which are essentially used to apply the standard theorems in ergodic theory and thermodynamic formalism, provide a logical chain which drives us to important conclusions that could be checked in experimental data.

The paper is organised as follows. In section 2 we introduce a framework in which spike trains are associated to a coding of membrane potential trajectories, and actually, constitute a symbolic coding in explicit examples. On this basis, we show how Gibbs distributions are natural probability measures to describe the statistics of spike trains, given the data of known empirical averages. Several authors have discussed the relevance of Gibbs distribution in the context of Hopfield model (see (Amit, 1989) and references therein) and more recently to spike trains (Wood, Roth, & Black, 2006; Kang & Amari, 2008). A breakthrough has been made in (Schneidman et al., 2006; Tkacik et al., 2006). Actually, our approach has been greatly influenced by these two papers, though our results hold in a wider context. In section 3, we discuss the effect of synaptic adaptation when considering "slow" synaptic plasticity rules where the characteristic time for synapse adaptation is quite a bit longer that the characteristic time for neurons dynamics. These rules are formulated in the context of thermodynamic formalism, where we introduce a functional, closely related to thermodynamic potentials like free energy in statistical physics, and called "topological pressure" in ergodic theory. In this setting we show that the variations of synaptic weights leads to variation of the topological pressure that can be smooth ("regular periods"), or singular ("phase transitions"). Phase transitions are in particular associated to a change of "grammar" inducing modifications in the set of spike trains that the dynamics is able to produce. We exhibit a functional, closely related to the topological pressure, that decreases during regular periods. As a consequence, when the synaptic weights converge to a fixed value, this functional reaches a minimum. This minimum corresponds to a situation where spike trains statistics are characterized by a Gibbs distribution, whose potential can be explicitly written. An example with numerical computations is presented in section 4.

2 Neuron dynamics.

2.1 Neural state.

We consider a set of N neurons. Each neuron i is characterized by its state, X_i, which belongs to some compact set $\mathcal{I} \in \mathbb{R}^M$. M is the number of variables characterizing the state of one neuron (we assume that all neurons are described by the same number of variables). A typical example is an integrate and fire model where $M = 1$ and $X_i = V_i$ is the membrane potential of neuron i and $\mathcal{I} = [V_{min}, V_{max}]$ (see section 2.5). Other examples are provided by conductances based models of Hodgkin-Huxley type[1] (Hodgkin & Huxley, 1952). Then $X_i = (V_i, m_i, n_i, h_i)$ where m_i, n_i are respectively the activation variable for Sodium and Potassium channels and h_i is the inactivation variable for the Sodium channel.

The evolution of these N neurons is given by a deterministic dynamical system of form:

$$\mathbf{X}(t+1) = \mathbf{F}_{\boldsymbol{\gamma}}\left[\mathbf{X}(t)\right] \tag{1}$$

where $\mathbf{X} = \{X_i\}_{i=1}^N$ represents the dynamical state of a network of N neurons at time t, while time is discrete (for a discussion on time discretisation in spiking neural networks see (Cessac & Viéville, 2008)). Thus $\mathbf{X} \in \mathcal{M} = \mathcal{I}^N$ where $\mathcal{M}$ is the phase space of (1), and $\mathbf{F}_{\boldsymbol{\gamma}}(\mathcal{M}) \subset \mathcal{M}$. The map $\mathbf{F}_{\boldsymbol{\gamma}} : \mathcal{M} \to \mathcal{M}$ depends on a set of parameters $\boldsymbol{\gamma} \in \mathbb{R}^P$. The typical case considered here is $\boldsymbol{\gamma} = \left(\mathcal{W}, \mathbf{I}^{(ext)}\right)$ where $\mathcal{W}$ is the matrix of synaptic weights and $\mathbf{I}^{(ext)}$ is some external current, assumed to be independent of time in the present paper (see section 2.5 for two explicit examples). Thus $\boldsymbol{\gamma}$ is a point in a $P = N^2 + N$ dimensional space of control parameters.

2.2 Natural partition.

Neurons are excitable systems. Namely, neuron i "fires" (emits a spike or action potential), whenever its state X_i belongs to some connected region $\mathcal{P}_1$ of its phase space. Otherwise, it is quiescent ($X \in \mathcal{P}_0 = \mathcal{I} \setminus \mathcal{P}_1$). In Integrate and Fire models neuron i fires whenever its membrane potential V_i exceeds some threshold θ. In this case, the corresponding region is $\mathcal{P}_1 = [\theta, V_{max}]$. In Fitzhugh-Nagumo (FitzHugh, 1955, 1961; Nagumo, Arimoto, & Yoshizawa, 1962) or Hodgkin-Huxley model (Hodgkin & Huxley, 1952), the firing corresponds to the crossing of a manifold called the threshold separatrix (Cronin, 1987) which separates the phase space into two connected regions $\mathcal{P}_0$ and $\mathcal{P}_1$. For N identical neurons this leads to a "natural partition" $\mathcal{P}$ of the product phase space $\mathcal{M}$. Call $\Lambda = \{0,1\}^N$, $\boldsymbol{\omega} = (\omega_i)_{i=1}^N \in \Lambda$. Then, $\mathcal{P} = \{\mathcal{P}_{\boldsymbol{\omega}}\}_{\omega \in \Lambda}$, where $\mathcal{P}_{\boldsymbol{\omega}} = \mathcal{P}_{\omega_1} \times \mathcal{P}_{\omega_2} \times \cdots \times \mathcal{P}_{\omega_N}$. Equivalently, if $\mathbf{X} \in \mathcal{P}_{\boldsymbol{\omega}}$, all neurons such that $\omega_i = 1$ are firing while neurons such that $\omega_k = 0$ are quiescent. We call therefore $\boldsymbol{\omega}$ a "spiking pattern".

2.3 Raster plots.

To each initial condition $\mathbf{X} \in \mathcal{M}$ we associate a "raster plot" $\omega = \{\boldsymbol{\omega}(t)\}_{t=0}^{+\infty}$ such that $\mathbf{X}(t) \in \mathcal{P}_{\boldsymbol{\omega}(t)}, \forall t \geq 0$. We write $\mathbf{X} \rightharpoonup \omega$. Thus, ω is the sequence of spiking patterns displayed by the

[1] Note that Hodgkin-Huxley equations are differential equations, while (1) corresponds to a discrete time evolution. We assume that we have discretized time with a time scale that can be arbitrary small. A detailed discussion on these aspects for conductance based IF models has been presented in (Cessac & Viéville, 2008). Some further comments are given below.

neural network when prepared with the initial condition $\mathbf{X}$. On the other way round, we say that an infinite sequence $\omega = \{\boldsymbol{\omega}(t)\}_{t=0}^{+\infty}$ is *an admissible raster plot* if there exists $\mathbf{X} \in \mathcal{M}$ such that $\mathbf{X} \rightharpoonup \omega$. We call $\Sigma_{\boldsymbol{\gamma}}$ the set of admissible raster plots for the set of parameters $\boldsymbol{\gamma}$.

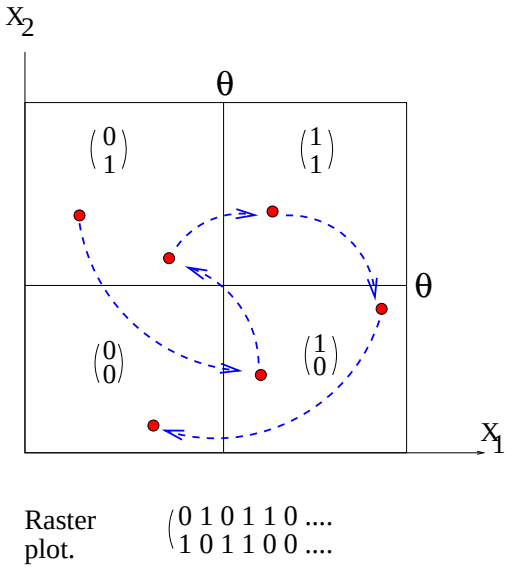

Figure 1: The phase space $\mathcal{M}$ (here represented for $N = 2$ neurons with constant threshold) is partitioned in 2^N parts $\mathcal{P}_{\boldsymbol{\omega}} = \prod_{i=1}^{N} \mathcal{P}_{\omega_i}$, $\boldsymbol{\omega}$ being a spiking pattern. In this way, one associates naturally to an orbit of (1) a raster plot.

The dynamics (1) induce a dynamics on the set of raster plot in the following way:

$$\begin{array}{ccc} \mathbf{X} & \stackrel{\mathbf{F}_{\boldsymbol{\gamma}}}{\longrightarrow} & \mathbf{F}_{\boldsymbol{\gamma}}(\mathbf{X}) \\ \downarrow & & \downarrow \\ \omega & \stackrel{\sigma_{\boldsymbol{\gamma}}}{\longrightarrow} & \sigma_{\boldsymbol{\gamma}}(\omega) \end{array}$$

In this construction $\sigma_{\boldsymbol{\gamma}}$, called the "left shift", shifts the raster plot left-wise at each time step of the dynamics. Thus, in some sense, raster plots provide a code for the orbits of (1). But, the correspondence may not be one-to-one. That is why we use the notation $\rightharpoonup$ instead of $\rightarrow$.

Let us introduce the following notation. If we are interested in a prescribed sequence of spiking patterns $\boldsymbol{\omega}(s), \ldots \boldsymbol{\omega}(t)$, from time s to time t, we denote by $[\omega]_{s,t}$, the set of raster plots whose spiking patterns from time s to time t match this sequence (cylinder set).

2.4 Asymptotic dynamics.

We are here mainly interested in the asymptotic behavior of (1) and set a few notations and notions. The ω-limit set, Ω, is the set of accumulation points of $\mathbf{F}_{\boldsymbol{\gamma}}^t(\mathcal{M})$. Since $\mathcal{M}$ is closed and invariant, we have $\Omega = \bigcap_{t=0}^{\infty} \mathbf{F}_{\boldsymbol{\gamma}}^t(\mathcal{M})$. In dissipative systems (i.e. a volume element in the phase space is dynamically contracted), the ω-limit set typically contains the attractors of the system. A compact set $\mathcal{A} \in \mathcal{M}$ is called an *attractor* for $\mathbf{F}_{\boldsymbol{\gamma}}$ if there exists a neighborhood $\mathcal{U}$

of $\mathcal{A}$ and a time $n > 0$ such that $\mathbf{F}_{\boldsymbol{\gamma}}^n(\mathcal{U}) \subset \mathcal{U}$ and $\mathcal{A} = \bigcap_{t=0}^{\infty} \mathbf{F}_{\boldsymbol{\gamma}}^t(\mathcal{U})$. In all examples considered in the present paper $\mathbf{F}_{\boldsymbol{\gamma}}$ is dissipative and the phase space is divided into finitely many attraction basins each of them containing an attractor (see Fig. 2). Simple examples of attractors are stable fixed points, or stable periodic orbits. More complex attractors such as chaotic attractors can be encountered as well. The attraction basins and attractors change when the parameters $\boldsymbol{\gamma}$ vary. These changes can be smooth (structural stability) or sharp (typically this arises at bifurcations points).

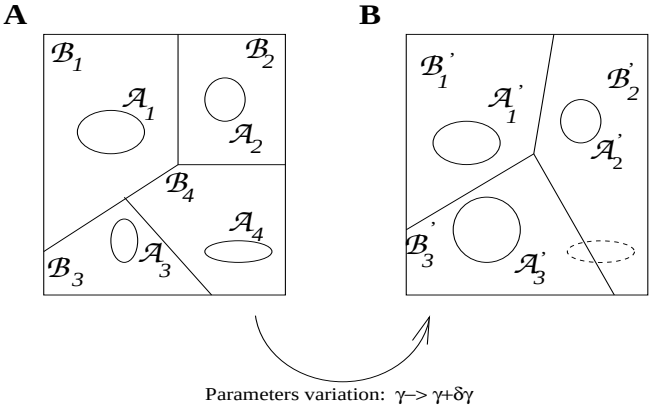

Figure 2: Schematic illustration of the attractor landscape for neural network models. [**A**] The phase space is partitioned into bounded domains $\mathcal{B}_l$ and for each initial condition in $\mathcal{B}_l$ the initial trajectory is attracted toward an attractor (e.g. a fixed point, a periodic orbit or a more complex attractor) $\mathcal{A}_l$. [**B**] If the parameters (external current, weights) change, the landscape is modified and several phenomena can occur: change in the basins shape, number of attractors, modification of the attractor as for $\mathcal{A}_3$ in this example; A point belonging to $\mathcal{A}_4$ in Fig. 2 A, can, after modification of the parameters, converge either to attractor $\mathcal{A}_2'$ or $\mathcal{A}_3'$

2.5 A generic example: generalized integrate and fire models

2.5.1 Quiescent stage of the neuron.

Among the various models of neural dynamics, generalized integrate and fire models play a central role, due to their (relative) simplicity, while the biological plausibility is well accepted (Rudolph & Destexhe, 2006; Gerstner & Kistler, 2002a). A representative class of such models is of form:

$$\frac{dV_k}{dt} = -\frac{1}{\tau_L}(V_k - E_L) + i_k^{(ext)} - i_k^{(syn)}(V_k, t, \left\{t_j^{(n)}\right\}_t), \tag{2}$$

defining the evolution of the membrane potential V_k of neuron k. Here $\tau_L = RC \simeq 10 - 20ms$ is the membrane time-constant related to the membrane resistance and its electric capacity, while $E_L \simeq -80mV$ is the related reversal potential. The term $i_k^{(ext)}$ is an external "current"[2] assumed to be time constant in this paper. In the end of this section we shall however consider the case where some noise is superimposed upon the constant external current. $t_j^{(n)}$ is the n-th

[2]This is a slight abuse of language since we have divided eq. (2) by the membrane capacity.

firing time[3] of neuron j and $\left\{t_j^{(n)}\right\}_t$ is the list of firing times of all neurons up to time t.

The synaptic currents reads:

"

$$i_k^{(syn)}(V_k, t, \left\{t_j^{(n)}\right\}_t) = (V_k - E^+) \sum_{j\in\mathcal{E}} g_{kj}(t, \left\{t_j^{(n)}\right\}_t) + (V_k - E^-) \sum_{j\in\mathcal{I}} g_{kj}(t, \left\{t_j^{(n)}\right\}_t),$$

where $E^\pm$ are reversal potential (typically $E^+ \simeq 0mV$ and $E^- \simeq -75mV$). $\mathcal{E}$ and $\mathcal{I}$ refers respectively to excitatory and inhibitory neurons and the $+$ $(-)$ sign is relative to excitatory (inhibitory) synapses. Note that conductances are always positive thus the sign of the post-synaptic potential (PSP) is determined by the reversal potentials $E^\pm$. At rest ($V_k \sim -70mV$) the $+$ term leads to a positive PSP while $-$ leads to a negative PSP.

Conductances depend on past spikes via the relation:

$$g_{kj}(t, \left\{t_j^{(n)}\right\}_t) = G_{kj} \sum_{n=1}^{M_j(t,\mathbf{V})} \alpha_j(t - t_j^{(n)}).$$

In this equation, $M_j(t, \mathbf{V})$ is the number of times neuron j has fired at time t and $\left\{t_j^{(n)}\right\}_t$ is the list of firing times of all neurons up to time t. G_{kj} are positive constants, proportional to the *synaptic efficacy*:

$$\begin{cases} W_{kj} = E^+ G_{kj} & \text{if} \quad j \in \mathcal{E}, \\ W_{kj} = E^- G_{kj} & \text{if} \quad j \in \mathcal{I}. \end{cases} \tag{3}$$

We use the convention $W_{kj} = 0$ if there is no synapse from j to k. Finally, α represents the unweighted shape (called a α-shape) of the post-synaptic potentials. Known examples of α-shapes are $\alpha(t) = Ke^{-t/\tau}H(t)$ or $\alpha(t) = Kte^{-t/\tau}H(t)$, where H is the Heaviside function.

Then, we may write (2) in the form (see (Cessac & Viéville, 2008) for more details):

$$\frac{dV_k}{dt} + g_k V_k = i_k,$$

with:

$$g_k(t, \left\{t_j^{(n)}\right\}_t) = \frac{1}{\tau_L} + \sum_{j=1}^{N} g_{kj}(t, \left\{t_j^{(n)}\right\}_t),$$

and:

$$i_k(t, \left\{t_j^{(n)}\right\}_t) = \frac{E_L}{\tau_L} + \sum_{j\in\mathcal{E}} W_{kj} \sum_{n=1}^{M_j(t,\mathbf{V})} \alpha_j(t - t_j^{(n)}) + \sum_{j\in\mathcal{I}} W_{kj} \sum_{n=1}^{M_j(t,\mathbf{V})} \alpha_j(t - t_j^{(n)}) + i_k^{(ext)}.$$

[3]For continuous time systems, the firing times of neuron k, for the trajectory $\mathbf{V}$, is defined by:

$$t_k^{(n)}(\mathbf{V}) = \inf\left\{t \,|t > t_k^{(n-1)}(\mathbf{V}),\ V_k(t) \geq \theta\right\},$$

where $t_k^{(0)} = -\infty$.

2.5.2 Fire regime.

The previous equations hold whenever the neuron is quiescent, i.e. whenever membrane potential is smaller than a threshold $\theta > 0$, usually depending on time (to account for characteristics such as refractory period of the neuron) and on the neuronal state. Here, however, we assume that θ is a constant (in fact without loss of generality (Gerstner & Kistler, 2002a)). When the membrane potential exceeds the threshold value, the neuron "fires" (emission of an action potential or "spike"). The spike shape depends on the model. In the present case, the membrane potential is reset instantaneously to a fixed value $V_{reset} \simeq E_L$, corresponding to the value of the membrane potential when the neuron is at rest. For simplicity we set $V_{reset} = 0$ without loss of generality. This modeling of the spike introduces a natural notion of spike time, but the price to pay is to introduce a discontinuity in dynamics. Moreover, since spike is instantaneous the set of possible spike times occurring within time interval is uncountable in the continuous case. Introducing a minimal time discretisation at scale δ removes this pathology.

2.5.3 Time discretisation.

Assuming that spike times are only known within a precision $\delta > 0$ (Cessac & Viéville, 2008) one shows that the dynamics of membrane potential, discretized at the time scale δ writes:

$$V_k(t+\delta) = \rho_k(t, t+\delta, [\omega]_{0,t})\, V_k(t) + J_k(t, [\omega]_{0,t}) \tag{4}$$

where:

$$\begin{aligned} J_k(t, [\omega]_{0,t}) &= \int_t^{t+\delta} i_k(s, [\omega]_{0,t})\, \rho_k(s, t+\delta, [\omega]_{0,t})\, ds, \\ \rho_k(s, t+\delta, [\omega]_{0,t}) &= e^{-\int_s^{t+\delta} g_k(s', [\omega]_{0,t})\, ds'}, \\ g_k(t, [\omega]_{0,t}) &= \frac{1}{\tau_L} + \sum_{j=1}^{N} G_{kj} \sum_{n=1}^{M_j(t,\omega)} \alpha^{\pm}(s - t_j^{(n)}), \end{aligned}$$

where $M_j(t, \omega)$ is the number of spikes emitted by neuron j up to time t, in the raster plot ω. In the sequel we assume that $\alpha^{\pm}(u) = 0$ whenever $|u| \geq \tau_M$, i.e. $g_k(t, [\omega]_{0,t})$ depends on past firing times over a finite time horizon τ_M. We also set $\delta = 1$.

2.5.4 Model I and model II.

A step further, two additional simplifications can be introduced:

1. Assuming that $\rho_k(t, t+\delta, [\omega]_{0,t}) = \rho$ is almost constant, which is equivalent to considering "current" synapses instead of "conductance" synapses, i.e. neglect the dependency of $i_k^{(syn)}$ with respect to V_k;

2. Considering a much simpler form for $J_k(t, [\omega]_{0,t})$, where each connected neuron simply increments the membrane potential during its firing state. This is equivalent to considering the post-synaptic profiles in the previous equations as instantaneous step-wise profiles.

Considering these assumptions leads the following dynamics of membrane potential:

$$\text{,,} \qquad \mathbf{F}_{\boldsymbol{\gamma},i}(\mathbf{V}) = \rho\, V_i\,(1 - Z[V_i]) + \sum_{j=1}^{N} W_{ij} Z[V_j] + I_i^{ext}; \qquad i = 1 \ldots N, \tag{5}$$

where $\mathbf{F}_{\boldsymbol{\gamma},i}$ is the i-th component of $\mathbf{F}_{\boldsymbol{\gamma}}$, $Z(x) = 1$ if $x \geq \theta$ and 0 otherwise, both the integrate and firing regime being integrated in this equation. It turns out that this corresponds exactly to the time discretisation of the standard integrate and fire neuron model, which as discussed in e.g.,(Izhikevich, 2003) provides a rough but realistic approximation of biological neurons behaviors. This model is called *model I* in the sequel whereas the model based on (4) is called *model II*.

2.5.5 Genericity results for models I and II.

The map $\mathbf{F}_{\boldsymbol{\gamma}}$ in models I and II is locally contracting (Cessac, 2008; Cessac & Viéville, 2008), but it is not smooth due to the sharp threshold in neurons firing definition. The *singularity set* where $\mathbf{F}_{\boldsymbol{\gamma}}$ is not continuous is:

$$\mathcal{S} = \{\mathbf{V} \in \mathcal{M} | \exists i = 1 \ldots N, \text{ such that } V_i = \theta\}.$$

This is the set of membrane potential vectors such that at least one of the neurons has a membrane potential exactly equal to the threshold. Because of this, dynamics exhibit sensitivity to perturbations for orbits which approach too close the singularity set.

Now, call

$$d(\Omega, \mathcal{S}) = \inf_{\mathbf{V} \in \Omega} \inf_{t \geq 0} \min_{i=1 \ldots N} |V_i(t) - \theta|. \tag{6}$$

Then, the following theorem holds (Cessac, 2008; Cessac & Viéville, 2008).

Theorem 1 *For a generic (in a metric and topological sense) set of parameters $\boldsymbol{\gamma}$, $d(\Omega, \mathcal{S}) > 0$. Consequently,*

1. *Ω is composed of finitely many periodic orbits with a finite period, Thus, attractors are generically stable period orbits.*
2. *There is a one-to-one correspondence between membrane potential trajectories and raster plots, except for a negligible set of $\mathcal{M}$.*
3. *There is a finite Markov partition.*

Remarks

1. Note however that, depending on parameters (synaptic weights, external current), some periods can be quite large (well beyond any accessible computational time) Also, the number of stable periodic orbits can grow exponentially with the number of neurons.
2. Result 2 means that spike trains provide a symbolic coding for the dynamics.

3. See section 2.6.1 for the implications of result 3.

These models constitute therefore nice examples where one has a good control on dynamics and" where one can apply the machinery of thermodynamic formalism (see Appendix). Note that since dynamics is eventually periodic, one may figure out that there is nothing non trivial to say from the respect of spike train statistics. The key point is that period are practically so large that there is any hope to see the periodicity in real experiments. Thus, one may do "as if" the system were "chaotic". This is a fortiori the case if one superimposes upon the external current a small amount of noise to the external current $\mathbf{I}^{(ext)}$.

Also, though dynamics of model I and II may look rather trivial compared to what is expected from biological neural networks, it might be that such models are able to approximate trajectories of a real neural network, for suitable values of $\boldsymbol{\gamma}$ (e.g. after a suitable adaptation of the synaptic weights) and provided N, the number of neurons, is sufficiently large. This type of properties are currently sought by computational neuroscientist with interesting indications that 'IF models are good enough" to approximate biological neurons spike trains (Jolivet, Rauch, Lescher, & Gerstner, 2006). See (Rostro-Gonzalez, Cessac, Vasquez, & Viéville, 2009) for a recent illustration of this.

2.6 Spikes dynamics and statistics.

2.6.1 Grammar.

Though dynamics (1) produces raster plots, it is important to remark that it is not able to produce *any* possible sequence of spiking patterns. This fundamental fact is most often neglected in computational neuroscience literature and leads, e.g. to severe overestimation of the system's entropy. Also, it plays a central role in determining spike train statistics.

There are therefore *allowed* and *forbidden* sequences depending on conditions involving the detailed form of $\mathbf{F}_{\boldsymbol{\gamma}}$ and on the parameters $\boldsymbol{\gamma}$ (synaptic weights and external current). Let $\boldsymbol{\omega}, \boldsymbol{\omega}'$ be two spiking patterns. The transition $\boldsymbol{\omega} \to \boldsymbol{\omega}'$ is *legal* or *admissible* if there exists a neural state $\mathbf{X}$ such that neurons fire according to the firing pattern $\boldsymbol{\omega}$ and, at the next time, according to the firing pattern $\boldsymbol{\omega}'$. Equivalently, $\mathbf{F}_{\boldsymbol{\gamma}}(\mathcal{P}_{\boldsymbol{\omega}}) \cap \mathcal{P}_{\boldsymbol{\omega}'} \neq \emptyset$. An admissible transition must satisfy *compatibility conditions* depending on synaptic weights and currents.

As an example, for model I, compatibility conditions write, for all $i = 1 \dots N$:

$$
\begin{array}{lll}
(a) & i \text{ is such that } \omega_i = 1 \text{ and } \omega'_i = 1 \quad \Leftrightarrow & \sum_{j=1}^N W_{ij}\omega_j + I_i^{ext} \geq \theta \\
(b) & i \text{ is such that } \omega_i = 1 \text{ and } \omega'_i = 0 \quad \Leftrightarrow & \sum_{j=1}^N W_{ij}\omega_j + I_i^{ext} < \theta \\
(c) & i \text{ is such that } \omega_i = 0 \text{ and } \omega'_i = 1 \quad \Leftrightarrow & \exists \mathbf{X} \in \mathcal{P}_{\boldsymbol{\omega}} \text{ such that, } \rho X_i + \sum_{j=1}^N W_{ij}\omega_j + I_i^{ext} \geq \theta \\
(d) & i \text{ is such that } \omega_i = 0 \text{ and } \omega'_i = 0 \quad \Leftrightarrow & \exists \mathbf{X} \in \mathcal{P}_{\boldsymbol{\omega}} \text{ such that, } \rho X_i + \sum_{j=1}^N W_{ij}\omega_j + I_i^{ext} < \theta
\end{array}
$$

Note that, while the two first conditions only depend on the partition element $\mathcal{P}_{\boldsymbol{\omega}}$, the two last ones depend on the point $\mathbf{X} \in \mathcal{P}_{\boldsymbol{\omega}}$. Basically, this means that the natural partition is not a Markov partition. The idea is then to find a refinement of the natural partition such that allowed transitions depend only on the partition-elements.

A nice situation occurs when the refined partition is *finite*. Equivalently, compatibility conditions can be obtained by a *finite* set of rules or a finite *grammar*. This is possible if there

exists some finite integer r with the following property: Construct blocks of spiking patterns $\boldsymbol{\omega}(0)\dots\boldsymbol{\omega}(r-1)$ of width r. There are at most 2^{Nr} such possible blocks (and in general quite a bit less *admissible* blocks). Label each of these blocks with a symbol $\alpha \in A$, where A is called an *alphabet*. Each block corresponds to a connected domain $\cap_{s=0}^{r-1}\mathbf{F}_{\boldsymbol{\gamma}}^{-s}[\mathcal{P}_{\boldsymbol{\omega}(s)}] \subset \mathcal{M}$ and to a cylinder set $[\omega]_{0,r-1} \subset \Sigma_{\boldsymbol{\gamma}}$. Define a *transition matrix* $G_{\boldsymbol{\gamma}} : A \times A \to \{0,1\}$, depending on $\boldsymbol{\gamma}$, with entries $g_{\alpha\beta}$, such that $g_{\alpha\beta} = 1$ if the transition $\beta \to \alpha$ is admissible, and 0 otherwise. To alleviate the notations in the next equation write $\alpha(t) = [\omega]_{t,t+r-1}$, $\forall t \geq 0$. If

$$\Sigma_{\boldsymbol{\gamma}} = \left\{\omega \mid g_{\alpha(t+1)\alpha(t)} = 1, \forall t \geq 0\right\},$$

then all admissible raster plots are obtained via the finite transition matrix $G_{\boldsymbol{\gamma}}$ which provides the grammar of spiking patterns sequences.

In models I, II such a finite grammar exists for the (generic) set of parameters $\boldsymbol{\gamma}$ such that $d(\Omega, \mathcal{S}) > 0$. Moreover, there are open domains in the space of parameters where the grammar is fixed and is not affected by small variations of synaptic weights or current.

For more general models, it might be that there is no finite grammar to describe all admissible raster plots. When dealing with experiments, one has anyway only access to finite raster plots and the grammar extracted from empirical observations has a finite number of rules (see (Collet, Galves, & Lopez, 1995; Chazottes, Floriani, & Lima, 1998; Chazottes, 1999) for nice applications of this idea in the field of turbulence). This empirical grammar is compatible with the dynamics at least for the time of observation.

The grammar depends on $\boldsymbol{\gamma}$. Assume now that we are continuously varying $\boldsymbol{\gamma}$. This corresponds to following some path in the space of control parameters. Then, this path generically crosses domains where grammar is fixed, while at the boundary of these domains, there is a grammar modification, and thus a change in the set of admissible raster plots that dynamics is able to produce. Synaptic adaptation, as described in section 3, corresponds precisely to this situation. Thus, we expect that the neural network exhibits "regular" periods of adaptation where spike trains before and after synaptic changes, correspond to the same grammar. Between these regular periods, sharp changes in raster plots structure are expected, corresponding somehow to displaying new transitions and new rules in the spike code. Actually, the ability of neural networks to display, after adaptation, a *specific* subset of admissible sequences, provided by a specific grammar, is an important issue of this paper.

2.6.2 Spike responses of neurons.

Neurons respond to excitations or stimuli by finite sequences of spikes. In model I and II a stimulus is typically the external current but more general forms (spike trains coming from external neurons) could be considered as well. Thus, the dynamical response R of a neuronal network to a stimuli S (which can be applied to several neurons in the network), is a sequence $\boldsymbol{\omega}(t)\dots\boldsymbol{\omega}(t+n)$ of spiking patterns. "Reading the neural code" means that one seeks a correspondence between responses and stimuli. However, the spike response does not only depend on the stimulus, but also on the network dynamics and therefore fluctuates randomly (note that this apparent randomness is provided by the dynamical evolution and does not require the invocation of an exogenous noise). Thus, the spike response is sought as a conditional probability

$P(R|S)$ (Rieke et al., 1996). Reading the code consists in inferring $P(S|R)$ e.g. via Bayesian approaches, providing a loose dictionary where the observation of a fixed spikes sequences R does not provide a unique possible stimulus, but a set of stimuli, with different probabilities. Having models for conditional probabilities $P(R|S)$ is therefore of central importance. For this, we need a good notion of statistics.

2.6.3 Performing statistics.

These statistics can be obtained in two different ways. Either one repeats a large number of experiments, submitting the system to the same stimulus S, and computes $P(R|S)$ by an average over these experiments. This approach relies on the assumption that the system has the same statistical properties during the whole set of experiments (i.e. the system has not evolved, adapted or undergone bifurcations meanwhile). In our setting this amounts to assuming that the parameters $\boldsymbol{\gamma}$ have not changed.

Or, one performs a time average. For example, to compute $P(R|S)$, one counts the number of times $n(R, T, \omega)$ when the finite sequence of spiking patterns R, appears in a spike train ω of length T, when the network is submitted to a stimulus S. Then, the probability $P(R|S)$ is estimated by:

$$P(R|S) = \lim_{T\to\infty} \frac{n(R, T, \omega)}{T}.$$

This approach implicitly assumes that the system is in a stationary state.

The empirical approach is often "in-between". One fixes a time window of length T to compute the time average and then performs an average over a finite number $\mathcal{N}$ of experiments corresponding to selecting different initial conditions. In any case the implicit assumptions are essentially impossible to control in real (biological) experiments, and difficult to prove in models. So, they are basically used as "working" assumptions.

To summarize, one observes, from $\mathcal{N}$ repetitions of the same experiment, $\mathcal{N}$ raster plots $\omega_m, m = 1 \ldots \mathcal{N}$ on a finite time horizon of length T. From this, one computes experimental averages allowing to estimate $P(R|S)$ or, more generally, to estimate the average value, $\langle\phi\rangle$, of some prescribed observable $\phi(\omega)$. These averages are estimated by :

$$\bar{\phi}^{(\mathcal{N},T)} = \frac{1}{\mathcal{N}T} \sum_{m=1}^{\mathcal{N}} \sum_{t=1}^{T} \phi(\sigma_{\boldsymbol{\gamma}}^t \omega_m). \tag{7}$$

Typical examples of such observables are $\phi(\omega) = \omega_i(0)$ in which case $\langle\phi\rangle$ is the firing rate of neuron i; $\phi(\omega) = \omega_i(0)\omega_j(0)$ then $\langle\phi\rangle$ measures the probability of spike coincidence for neuron j and i; $\phi(\omega) = \omega_i(\tau)\omega_j(0)$ then $\langle\phi\rangle$ measures the probability of the event "neuron j fires and neuron i fires τ time step later" (or sooner according to the sign of τ). In the same way $P(R|S)$ is the average of the indicatrix function $\chi_R(\omega) = 1$ if $\omega \in R$ and 0 otherwise. Note that in (7) we have used the shift $\sigma_{\boldsymbol{\gamma}}^t$ for the time evolution of the raster plot. This notation is more compact and more adapted to the next developments than the classical formula, reading, e.g., for firing rates $\frac{1}{\mathcal{N}T} \sum_{m=1}^{\mathcal{N}} \sum_{t=1}^{T} \phi(\boldsymbol{\omega}_m(t))$.

This estimation depends on T and $\mathcal{N}$. However, one expects that, as $\mathcal{N}, T \to \infty$, the empirical average $\bar{\phi}^{(\mathcal{N},T)}$ converges to the theoretical average $\langle \phi \rangle$, as stated e.g. from the law of large numbers. Unfortunately, one usually does not have access to these limits, and one is lead to extrapolate theoretical averages from empirical estimations. The main difficulty is that these observed raster plots are produced by an underlying dynamics which is usually not explicitly known (as it is the case in experiments) or impossible to fully characterize (as it is the case in most large dimensional neural networks models). Even for models I and II, where one has a full characterization of generic orbits, a numerical generation of raster plots can produce periodic orbits whose period is out of reach. Thus, one is constrained to propose ad hoc statistical models. This amounts to assuming that the underlying dynamics satisfies specific constraints.

2.6.4 Inferring statistics from an hidden dynamics.

We follow this track, proposing a generic method to construct statistical models from a prescribed set of observables. For this, we shall assume that the observed raster plots are generated by a *uniformly hyperbolic* dynamical system, where spike trains constitute a symbolic coding for dynamics (see the appendix for more details). This a technical assumption allowing us to develop the thermodynamic formalism on a safe mathematical ground. Basically, we assume that the observed raster plots can be generated by a dynamical system which is chaotic with exponential correlation decay, and that a finite Markov partition exists, obtained by a refinement of the natural partition.

This is obviously a questionable assumption, but let us give a few arguments in its favor. First, it has been argued and numerically checked in (Rostro-Gonzalez et al., 2009) that finite sequence of spiking patterns, produced by an hidden neural network, including data coming from biological experiments, can be exactly reproduced by a gIF model of type II, by adding hidden neurons and noise to the dynamics. This suggests that gIF models are somehow "dense" in the space of spiking neural networks dynamical systems, meaning that any finite piece of trajectory from a spiking neural network can be approached by a gIF model trajectory. Second, adding noise to model I, II makes them uniformly hyperbolic[4], though not continuous. So, we expect the theory developed below to apply to model I, II. Now, dealing with spike train coming from a neural network whose mathematical properties are unknown (this is especially the case with biological neural networks), and whose statistical properties are sought, the idea is to do as if these data were generated by a dynamical system like model I or II.

As a matter of fact, the choice of a statistical model always relies on assumptions. Here we make an attempt to formulate these assumptions in a compact way with the widest range of application. These assumptions are compatible with the statistical models commonly used in the literature like Poisson models or Ising like models à la Schneidman and collaborators(Schneidman

[4]Contraction is uniform in model I. In model II it can be bounded by a constant strictly lower that 1. Introducing noise amounts to adding, for each neuron, an unstable direction where randomness is generated by a chaotic one-dimensional system. One forms in this way a skew product where the unstable fiber is the random generator and the stable one corresponds to the contracting dynamics of membrane potentials (eq. (4), or (5)) Because of the discontinuity of the map, the main difficulty is to show that, in this extended system, generic points have a local stable manifold of sufficiently large diameter (Cessac & Fernandez, in preparation). See (Blanchard, Cessac, & Krueger, 2000; Cessac, Blanchard, Krueger, & Meunier, 2004) for an application of this strategy in the framework of Self-Organized Criticality.

et al., 2006), but lead also us to propose more general forms of statistics. Moreover, our approach incorporates additional elements such as the consideration of neurons dynamics, and the grammar. This last issue is, according to us, fundamental, and, to the best of our knowledge, has never been considered before in this field. Finally, this postulate allows us to propose algorithms that can be applied to real data with an posteriori check of the initial hypotheses (Cessac, Vasquez, & Viéville, 2009).

On this basis we propose the following definition. Fix a set ϕ_l, $l = 1 \dots K$, of observables, i.e. functions $\Sigma_{\boldsymbol{\gamma}} \to \mathbb{R}$ which associate real numbers to sequences of spiking patterns. Assume that the empirical average (7) of these functions has been computed, for a finite T and $\mathcal{N}$, and that $\bar{\phi_l}^{(T,\mathcal{N})} = C_l$.

A *statistical model* is a probability distribution ν on the set of raster plots such that:

1. $\nu(\Sigma_{\boldsymbol{\gamma}}) = 1$, i.e. the set of non admissible raster plots has a zero ν-probability.
2. ν is ergodic for the left-shift $\sigma_{\boldsymbol{\gamma}}$ (see section 6.1 in the appendix for a definition).
3. For all $l = 1 \dots K$, $\nu(\phi_l) = C_l$, i.e., ν is compatible with the empirical averages.

Note that item 2 amounts to assuming that statistics are invariant under time translation. On practical grounds, this hypothesis can be relaxed using sliding time windows. This issue is discussed in more details in (Cessac et al., 2009). Note also that ν depends on the parameters $\boldsymbol{\gamma}$.

Assuming that ν is ergodic has the advantage that one does not have to average *both* over experiments *and* time. It is sufficient to focus on time average for a single raster plot, via the time-empirical average:

$$\pi_\omega^{(T)}(\phi) = \frac{1}{T}\sum_{t=1}^{T} \phi(\sigma_{\boldsymbol{\gamma}}^t \omega)$$

defined within more details in the appendix, section 6.2.

Item 3 can be generalized as follows.

3'. $d(\pi_\omega^{(T)}, \nu) \to 0$, as $T \to \infty$,

where $d(\mu, \nu)$ is the relative entropy or Kullback-Leibler divergence between two measures μ, ν (see eq. (54) in the appendix.).

Dealing with real or numerical data, it is obviously not possible to extrapolate to $T \to \infty$, but the main idea here is to minimize the "distance" $d(\pi_\omega^{(T)}, \nu)$ between the empirical measure, coming from the data, and the statistical model. Especially, if several models can be proposed then, the "best" one minimizes the Kullback-Leibler divergence (see th. 2 in the Appendix). The main advantage is that $d(\pi_\omega^{(T)}, \nu)$ can be numerically estimated using the thermodynamic formalism. Note however that this approach may fail at phase transition points where $d(\mu, \nu) = 0$ does not necessarily imply $\mu = \nu$ (Chazottes, 1999). Phase transition can be numerically detected from empirical data (Comets, 1997).

2.6.5 Gibbs measures as statistical models.

Variational principle. Statistical physics naturally proposes a canonical way to construct a statistical model: "Maximizing the entropy under the constraints $\nu(\phi_l) = C_l$, $l = 1 \dots K$", (see (Jaynes, 1957) for a beautiful and deep presentation of what became, since then, a "folklore" result). In the context of thermodynamic formalism this amounts to solving the following variational principle (see eq. (43) in the Appendix).

$$P[\boldsymbol{\psi}] = \sup_{\nu \in m^{(inv)}} (h[\nu] + \nu[\boldsymbol{\psi}]),$$

where $m^{(inv)}$ is the set of invariant measures for σ and h the Kolomogorov-Sinai entropy or entropy rate (see Appendix for more details). The "potential" $\boldsymbol{\psi}$ is given by $\boldsymbol{\psi} = \sum_{l=1}^{K} \lambda_l \phi_l$ where the λ_l's are adjustable Lagrange multipliers. A measure $\nu_{\boldsymbol{\psi}}$ which realizes the supremum, i.e.

$$P[\boldsymbol{\psi}] = h[\nu_{\boldsymbol{\psi}}] + \nu_{\boldsymbol{\psi}}[\boldsymbol{\psi}],$$

is called, in this context, an "equilibrium state". The function $P[\boldsymbol{\psi}]$ is called the "topological pressure".

Gibbs states. Uniformly hyperbolic dynamical systems have equilibrium states. Moreover, in this case, equilibrium states are *Gibbs states* (and vice-versa). A Gibbs state, or Gibbs measure, is a probability measure such that, one can find some constants c_1, c_2 with $0 < c_1 \leq 1 \leq c_2$ such that for all $n \geq 1$ and for all $\omega \in \Sigma$:

$$c_1 \leq \frac{\nu_{\boldsymbol{\psi}}\left(\omega \in [\omega]_{0,n-1}\right)}{\exp(-nP[\boldsymbol{\psi}] + S^{(n)}\boldsymbol{\psi}(\omega))} \leq c_2,$$

(Cf eq. (51) in the Appendix), where $S^{(n)}\boldsymbol{\psi}(\omega) = \sum_{t=0}^{n-1} \boldsymbol{\psi}(\sigma_{\boldsymbol{\gamma}}^t \omega)$. Basically, this means that the probability that a raster plot starts with the bloc $[\omega]_{0,n-1}$ behaves like $\frac{\exp(S^{(n)}\boldsymbol{\psi}(\omega))}{Z_n}$. One recognizes the classical Gibbs form where space translation in lattice system is replaced by time translation (shift $\sigma_{\boldsymbol{\gamma}}^t$) and where the normalization factor Z_n is the partition function. Note that $P[\boldsymbol{\psi}] = \limsup_{n\to\infty} \frac{1}{n} \log Z_n$, so that $P[\boldsymbol{\psi}]$ is the formal analog of a thermodynamic potential (like free energy).

Topological pressure. The topological pressure is a convex function. Moreover, this is the generating function for the cumulants of the Gibbs distribution. Especially, the Lagrange multipliers λ_l can be tuned to a value λ_l^* such that $\nu_{\boldsymbol{\psi}}[\phi_l] = C_l$, (item 3) using:

$$\left.\frac{\partial P[\boldsymbol{\psi}]}{\partial \lambda_l}\right|_{\lambda_l = \lambda_l^*} = C_l. \tag{8}$$

It expresses that the Gibbs state is given the tangent of the pressure at λ_l^* (Keller, 1998). Note that we have here assumed that the topological pressure is differentiable, namely that we are away from a phase transition. Higher order moments are obtained in the same way, especially second order moments related to the central limit theorem obeyed by Gibbs distributions

(Bowen, 1975, 2008; Keller, 1998). It is also possible to obtain averages of more complex functions than moments (Ji, 1989). The topological pressure is obtained via the spectrum of the Ruelle-Perron-Frobenius operator and can be calculated numerically when $\boldsymbol{\psi}$ has a finite (and small enough) range (see appendix for more details).

Validating a Gibbs statistical model. The Kullback-Leibler divergence provides some notion of "distance" between two measures. For μ an invariant measure and $\nu_{\boldsymbol{\psi}}$ a Gibbs measure with a potential $\boldsymbol{\psi}$, both defined on the same set of sequences Σ, one has (see eq. (55) in the appendix):

$$d\left(\mu, \nu_{\boldsymbol{\psi}}\right) = P\left[\boldsymbol{\psi}\right] - \int \boldsymbol{\psi} d\mu - h(\mu).$$

Though $\pi_{\omega}^{(T)}$ is not invariant, this suggests to use this relation to compare different statistical models (i.e. when several choices of observables ϕ_l are possible) by choosing the one which minimizes the quantity:

$$d(\pi_{\omega}^{(T)}, \nu_{\boldsymbol{\psi}}) = P\left[\boldsymbol{\psi}\right] - \pi_{\omega}^{(T)}(\boldsymbol{\psi}) - h(\pi_{\omega}^{(T)}). \tag{9}$$

(see th. 2 in the appendix). The advantage is that this quantity can be numerically estimated (Cessac et al., 2009).

Probability of spiking patterns blocs. In this context, the probability of a spiking pattern block $R = [\boldsymbol{\omega}]_{0,n-1}$ of length n corresponding to the response R to a stimuli S "behaves like" (in the sense of eq. (51)):

$$P\left[R|S\right] = \nu\left[\omega \in R|S\right] \sim \frac{1}{Z_n\left[\boldsymbol{\lambda}^*(S)\right]} \exp\left[\sum_{l=1}^{K} \lambda_l^*(S) \sum_{t=0}^{n-1} \phi_l(\sigma_{\boldsymbol{\gamma}}^t \omega)\right], \tag{10}$$

where the conditioning is made explicit in the dependence of $\boldsymbol{\lambda}^*$ in the stimulus S. (Typically, referring to model I and II, S is an external current and is incorporated in $\boldsymbol{\gamma}$.) This equation holds as well when $S = 0$ (no stimulus). Obviously, for two different stimuli the probability $P(R|S)$ may drastically change. Indeed, different stimuli imply different parameters $\boldsymbol{\gamma}$ thus a different dynamics and different raster plots, with different statistical weights on a specific bloc R. The grammar itself can also change, in which case R may be forbidden for a stimulus and allowed for another one.

2.7 Examples.

2.7.1 Canonical examples.

Firing rates. If $\phi_l(\omega) = \omega_l(0)$, then $\pi_{\omega}^{(T)}(\phi_l) = r_l$ is the average firing rate of neuron l within the time period T. Then, the corresponding statistical model is a Bernoulli distribution where neuron l has a probability r_l to fire at a given time. The probability that neuron l fires k times within a time delay n is a binomial distribution and the inter-spike interval is Poisson distributed (Gerstner & Kistler, 2002a).

Spikes coincidence. If $\phi_l(\omega) \equiv \phi_{(i,j)}(\omega) = \omega_i(0)\,\omega_j(0)$ where, here, the index l is an enumeration for all (non-ordered) pairs (i,j), then the corresponding statistical models has the form of an Ising model, as discussed by Schneidman and collaborators in (Schneidman et al., 2006; Tkacik" et al., 2006). As shown by these authors in experiments on the salamander retina, the probability of spike blocs estimated from the "Ising" statistical model fits quite better to empirical date than the classical Poisson model.

Enlarged spikes coincidence. As a generalization one may consider the probability of co-occurrence of spikes from neuron i and j within some time interval τ. The corresponding functions are $\phi_l(\omega) = \omega_i(0)\omega_j(\tau)$ and the probability of a spike bloc R writes:

$$P\left[R|S\right] = \frac{1}{Z_n\left[\boldsymbol{\lambda}^*(S)\right]} \exp\left[\sum_{i,j} \lambda^*_{i,j}(S) \sum_{t=0}^{n-1} \omega_i(t)\,\omega_j(t+\tau)\right].$$

Further generalizations are considered below.

2.8 Conclusion.

The main conclusion of this section is that Gibbs measures constitute *optimal* (in the sense of entropy maximization) statistical models whenever a set of observable has been prescribed. It also allows to select a model between several choices by minimizing the Kullback-Leibler divergence between the empirical measure and the statistical models.

In all the examples presented above the statistical model is determined by the choice of an *a priori* form for the potential. In the next section, we explicitly compute the potential in neural networks with synaptic adaptation.

3 Synaptic plasticity.

3.1 Synapses update as an integration over spikes trains.

Synaptic plasticity corresponds to the evolution of synaptic efficacy (synaptic weights). More precisely, in our notations (eq. (3)), W_{ij} essentially provides the maximal amplitude of the post-synaptic potential induced, at the synapse linking j to i, when neuron j fires a spike. Synaptic weights evolve in time according to the spikes emitted by the pre- and post- synaptic neuron. In other words, the variation of W_{ij} at time t is a function of the spiking sequences of neurons i and j from time $t - T_s$ to time t, where T_s is time scale characterizing the width of the spike trains influencing the synaptic change. In most examples the synapse update writes:

$$\delta W_{ij}(t+1) = g\left(W_{ij}(t), [\omega_i]_{t-T_s,t}, [\omega_j]_{t-T_s,t}\right),\ t > T_s, \tag{11}$$

with $[\omega_i]_{t-T_s,t} = [\omega_i(t-T_s)\ldots\omega_i(t)]$. Thus, typically, synaptic adaptation results from an integration of spikes over the time scale T_s.

3.2 Spikes responses of synapses.

The synaptic variation δW_{ij} is the integrated response of the synapse from neuron j to neuron i when neuron j sends a spike sequence $[\omega_j]_{t-T_s,t}$ and neuron i fires according to $[\omega_i]_{t-T_s,t}$. This response is not a deterministic function, while g is. Thus (11) is an approximation. As a matter of fact, the explicit form of g is usually derived from phenomenological considerations as well as experimental results where synaptic changes can be induced by *specific* simulations conditions, defined through the firing frequency of pre- and post-synaptic neurons (Bliss & Gardner-Medwin, 1973; Dudek & Bear, 1993), the membrane potential of the post-synaptic neuron (Artola, Bröcher, & Singer, 1990), or spike timing (Levy & Stewart, 1983; Markram et al., 1997; Bi & Poo, 2001) (see (Malenka & Nicoll, 1999) for a review). Thus, these results are usually based on a repetition of experiments involving the excitation of pre- and post-synaptic neurons by specific spike trains. The phenomenological plasticity rules derived from these experiments are therefore of *statistical* nature. Namely, they do not tell us what will be the exact changes induced on synapses when this or this spike train is applied to pre- and post-synaptic neuron. Instead, they provide us the *average* synaptic change. Thus, the function $g(W_{ij}, [\omega_i]_{t-T_s,t}, [\omega_j]_{t-T_s,t})$ in (11) is typically a statistical average of the synaptic response when the spike train of neuron j (resp. i) is $[\omega_j]_{t-T_s,t}$ (resp. $[\omega_i]_{t-T_s,t}$), and the actual synaptic weight value is W_{ij}.

In this paper we investigate the effects of those synaptic plasticity rules when the characteristic time scale T_s is quite a bit larger than the time scale of evolution of the neurons. Namely, we consider *slow* adaptation rules. In this situation the synaptic weights update can be written as a function of the empirical average (42). We therefore consider adaptation rules of form:

$$g\left(W_{ij}, [\omega_i]_{t-T_s,t}, [\omega_j]_{t-T_s,t}\right) = \epsilon \pi_\omega^{(T)}\left[\phi_{ij}(W_{ij}, .)\right]. \tag{12}$$

where ϵ is a parameter that will be typically small. $\phi_{ij}(W_{ij}, \omega)$ is a function that we now make explicit.

Following (Gerstner & Kistler, 2002a) canonical form of adaptation rules use an expansion in singlet, pairs, triplets etc of spikes. Formally, one may write a generic form for these rules. Fix $T_s < \infty$ a positive integer. Let $\mathcal{L}$ be the finite set of ordered lists L in $\{-T_s, -T_s+1, \ldots, 0\}$ with the form $L = (t_1, t_2, \ldots, t_L)$ with $-T_s \leq t_1 < t_2 < \cdots < t_L \leq 0$. For $L \in \mathcal{L}$, $1 \leq i \leq N$, we call an *i-monomial* a function $m_{i,L}(\omega) = \omega_i(t_1) \ldots \omega_i(t_L)$. We call an (i,j) *polynomial* a function:

$$\phi_{ij}(W_{ij}, \omega) = \sum_{L_1, L_2 \in \mathcal{L}} h_{ijL_1L_2}(W_{ij})\, m_{i,L_1}(\omega)\, m_{j,L_2}(\omega), \tag{13}$$

where $h_{ijL_1L_2}(W_{ij})$ are smooth functions of W_{ij}. They can be constant, linear functions of W_{ij} ("multiplicative" rules) or nonlinear functions allowing for example to constrain W_{ij} within bounded values (e.g. "hard bound" or "soft bounds" rules). The form (13) is the most general form of synaptic adaptation rules considered in this paper, while g has the explicit form:

$$g\left(W_{ij},[\omega_i]_{t-T_s,t},[\omega_j]_{t-T_s,t}\right) =$$
$$\epsilon \sum_{k,l=0}^{\text{"}K\text{"}} \sum_{\substack{-T_s \le t_1 < \cdots < t_k \le 0 \\ -T_s \le s_1 < \cdots < s_l \le 0}} h_{ij;t_1,\ldots,t_k,s_1,\ldots,s_l}(W_{ij})\, \pi_\omega^{(T)}\left[\omega_i(t+t_1)\ldots\omega_i(t+t_k)\,\omega_j(t+s_1)\ldots\omega_j(t+s_l)\right]. \tag{14}$$

Though explicit, this formulation is heavy, and we use instead the form (12), (13).

3.3 Examples of adaptation rule instantiation.

Hebbian learning uses firing rates. A typical example corresponds to $g_{ij}(W_{ij},[\omega_i]_{t-T_s,t},[\omega_j]_{t-T_s,t}) = \epsilon\frac{1}{T_s}\sum_{s_1,s_2=t-T_s}^{t}(\omega_i(s_1)-r_i(s_1))(\omega_j(s_2)-r_j(s_2)$ (correlation rule) where $r_i(t)=\frac{1}{T_s}\sum_{s=t-T_s}^{t}\omega_i(s)$ is the frequency rate of neuron i in the raster plot ω, computed in the time windows $[t-T_s,t]$.

Spike-Time Dependent Plasticity as derived from Bi and Poo (Bi & Poo, 2001) provides the average amount of synaptic variation given the delay between the pre- and post-synaptic spike. Thus, "classical" STDP writes (Gerstner & Kistler, 2002a; Izhikevich & Desai, 2003):

$$g\left(W_{ij},[\omega_i]_{t-T_s,t},[\omega_j]_{t-T_s,t}\right) = \frac{\epsilon}{T_s}\sum_{s_1,s_2=t-T_s}^{t} f(s_1-s_2)\,\omega_i(s_1)\,\omega_j(s_2) \tag{15}$$

with:

$$f(x)=\begin{cases} A_-e^{\frac{x}{\tau_-}}, & x<0, \quad A_-<0; \\ A_+e^{-\frac{x}{\tau_+}}, & x>0, \quad A_+>0; \\ 0, & x=0; \end{cases} \tag{16}$$

where the shape of f has been obtained from statistical extrapolations of experimental data. Hence STDP is based on a second order statistics (spikes correlations). There is, in this case, an evident time scale $T_s=\max(\tau_-,\tau_+)$, beyond which f is essentially zero.

"Nearest neighbors" STDP (according to the terminology of(Izhikevich & Desai, 2003)) writes:

$$g\left(W_{ij},[\omega_i]_{t-T_s,t},[\omega_j]_{t-T_s,t}\right) = \frac{\epsilon}{T_s}\sum_{s=t-T_s}^{t} f(\tau_j(s)-s)\,\omega_i(\tau_j(s))\,\omega_j(s), \tag{17}$$

with $\tau_j(s)=\min_{t,\omega_j(t)=1}|t-s|$.

Generalized STDP As a last example, (Gerstner & Kistler, 2002a) propose a rule which corresponds to:

$$g(W_{ij},[\omega_i]_{t-T_s,t},[\omega_j]_{t-T_s,t}) = \epsilon\left[a_1^{pre}\sum_{s=t-T_s}^{t}\omega_j(s) + a_1^{post}\sum_{s=t-T_s}^{t}\omega_i(s) + \sum_{s_1,s_2=t-T_s}^{t} f(s_1-s_2)\omega_i(s_1)\omega_j(s_2)\right] \tag{18}$$

We capture thus the main standard synaptic adaptation rules with (13).

3.4 Coupled dynamics.

We consider now the following coupled dynamics. Neurons are evolving according to (1). We focus here on *slow* synapses dynamics. Namely, synaptic weights are constant for $T \geq T_s$ consecutive dynamics steps, where T is large. This defines an "adaptation epoch". At the end of the adaptation epoch, synaptic weights are updated according to (11). This has the consequence of modifying neurons dynamics and possibly spike trains. The weights are then updated and a new adaptation epoch begins. We denote by t the update index of neuron states (neuron dynamics) inside an adaptation epoch, while τ indicates the update index of synaptic weights (synaptic plasticity). Call $\mathbf{X}^{(\tau)}(t)$ the state of the neurons at time t within the adaptation epoch τ. Let $W_{ij}^{(\tau)}$ be the synaptic weights from neuron j to neuron at i in the τ-th adaptation epoch. At the end of each adaptation epoch, the neuron dynamics time indexes are reset, i.e. $x_i^{(\tau+1)}(0) = x_i^{(\tau)}(T), i = 1 \dots N$. The coupled dynamics writes:

$$\left\{ \begin{array}{ccc} \mathbf{X}^{(\tau)}(t+1) & = & \mathbf{F}_{\boldsymbol{\gamma}^{(\tau)}}(\mathbf{X}^{(\tau)}(t)) \\ \delta W_{ij}^{(\tau)} & \stackrel{\text{def}}{=} & W_{ij}^{(\tau+1)} - W_{ij}^{(\tau)} = g\left(W_{ij}^{(\tau)}, [\omega_i]_{t-T_s,t}, [\omega_j]_{t-T_s,t}\right) \end{array} \right. \tag{19}$$

Recall that $\boldsymbol{\gamma} = (\mathcal{W}, \mathbf{I}^{(ext)})$ (see section 2.1) and $\boldsymbol{\gamma}^{(\tau)}$ is the set of parameters at adaptation epoch τ. In the present setting the external current $\mathbf{I}^{(ext)}$ is kept fixed and only synaptic weights are evolving. Basically, $\mathbf{I}^{(ext)}$ is used as an *external stimulus*.

3.5 Statistical effects of synaptic plasticity.

Synaptic plasticity has several prominent effects. First, modifying the synaptic weights has an action on the dynamics resulting, either in smooth changes, or in sharp changes. In the first case, corresponding to parameters variations in a domain where the dynamical system is structurally stable, small variations of the synaptic weights induce smooth variations on the ω-limit set structure (e.g. points are slightly moved) and in the statistics of orbits. The grammar is not changed. On the opposite, when crossing bifurcations manifolds, the slightest change in one synaptic weight value results typically in a drastic reorganization of the ω-limit set structure where attractors and their attraction basin are modified (see Fig. 2 and (Daucé et al., 1998; Siri et al., 2007) for an illustration of this in the case of frequency rates neural networks with Hebbian learning). But this can also change the grammar and the set of admissible raster plots. Some forbidden transitions become allowed, some allowed transitions become forbidden. Finally, the spikes train statistics are also modified. Typically, the time-empirical average of the raster plot changes ($\pi^{(T)}_{\omega^{(\tau)}} \to \pi^{(T)}_{\omega^{(\tau+1)}}$) and the corresponding statistical model also evolves.

But synaptic adaptation is not a mere variation of synaptic weights. Indeed, synaptic weights are associated with the coupled dynamics (19), meaning that synaptic changes depends on neurons dynamics, itself depending on parameters. Thus, a synaptic adaptation process corresponds to following a path in the space of parameters $\boldsymbol{\gamma}$; this path is not determined a priori but evolves according to neurons dynamics. Though this path can belong to a unique domain of structural stability, this is not the case in general. At some point during the adaptation, this path crosses a bifurcation manifold inducing sharp changes in the dynamics and in the grammar[5]. Hence,

[5]There is here an analogy with phase transitions in statistical physics (Beck & Schloegl, 1995). Detecting and

after several changes of this type one can end up with a system displaying raster plots with a structure rather different from the initial situation. These changes depend obviously upon the detailed form of neuron dynamics (1) and upon the synaptic update mechanism (11); they are also" conditioned by parameters such as stimuli. We now Analise these effects in the light of thermodynamic formalism and Gibbs distributions.

3.5.1 Static synaptic weights.

Let us first consider the situation where $\delta\mathcal{W} = 0$, corresponding to synaptic weights that do not evolve. Typically, this is the case if synaptic weights matrix converges to an asymptotic value $\mathcal{W}^*$. From eq. (12) this corresponds to :

$$\pi_\omega^{(T)}\left[\phi_{ij}\right] = 0,\ \forall i,j \in \{1,\dots N\}\,. \tag{20}$$

This imposes a condition on the average value of ϕ_{ij}. Therefore, from section 2.6.5 *this imposes that the statistical model is a Gibbs measure ν with a potential of form*:

$$\boldsymbol{\psi}^* = \Phi + \boldsymbol{\lambda}^*.\boldsymbol{\phi}, \tag{21}$$

where $\boldsymbol{\psi}^* = \left(\psi_{ij}^*\right)_{i,j=1}^N$, $\boldsymbol{\phi} = (\phi_{ij})_{i,j=1}^N$, $\boldsymbol{\lambda}^* = \left(\lambda_{ij}^*\right)_{i,j=1}^N$ and $\boldsymbol{\lambda}^*.\boldsymbol{\phi} = \sum_{i,j=1}^N \lambda_{ij}^*\phi_{ij}$. The potential Φ in (21) is such that $\Phi(\omega) = 0$ if ω is admissible and $\Phi(\omega) = -\infty$ if it is forbidden, so that forbidden raster plots have zero probability. This is a way to include the grammar in the potential (see appendix). The statistical parameters λ_{ij}^*, are given by eq. (8) in section 2.6.5, and, making ϕ_{ij} explicit (eq. (13)):

$$\left.\frac{\partial P\left[\psi\right]}{\partial\lambda_{ij}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = \nu_{\boldsymbol{\psi}^*}(\phi_{ij}) = \sum_{L_1,L_2\in\mathcal{L}} h_{ijL_1L_2}(W_{ij}^*)\nu_{\boldsymbol{\psi}^*}\left[m_{i,L_1}m_{j,L_2}\right] = 0. \tag{22}$$

Since $m_{i,L_1}m_{j,L_2}$ are monomials, this equation thus imposes *constraints* on the probability

$$\nu_{\boldsymbol{\psi}^*}\left[\omega_i(t_1)\dots\omega_i(t_{L_1})\omega_j(s_1)\dots\omega_j(s_{L_2})\right],$$

of spikes n-uplets $\omega_i(t_1)\dots\omega_i(t_L)\omega_j(s_1)\dots\omega_j(s_L)$. Note that condition (22) corresponds to an *extremum* for the topological pressure as a function of $\boldsymbol{\lambda}$. Note also that the variational formulation of Gibbs (equilibrium) state writes, in this case $P\left[\boldsymbol{\psi}^*\right] = \sup_{\nu\in m^{(inv)}} h\left[\nu\right]$ since $\nu(\psi) = 0$. Hence, the corresponding Gibbs measure has *maximal entropy.*

Let us emphasize what we have obtained. The statistical model that fits with the condition (20) in section 2.6.5 is a Gibbs distribution such that the probability of a spin block R of length n is given by :

$$P\left[R|S\right] = \frac{1}{Z_n\left[\boldsymbol{\lambda}^*(S)\right]}\exp\left[\sum_{t=1}^n \boldsymbol{\psi}^*(\sigma_{\boldsymbol{\gamma}}^t\omega)\right].$$

characterizing these phase transitions from empirical data is possible provided that there are only a few control parameters (Comets, 1997).

where $\boldsymbol{\psi}^*$ depends on S via the statistical parameters $\boldsymbol{\lambda}^*$ and via Φ (grammar).

When the situation $\delta\mathcal{W} = 0$ corresponds to the asymptotic state for a synaptic adaptation process, this potential provides us the form of the statistical model *after adaptation*, and *integrates all past changes in the synaptic weights*. We now discuss this process within details.

3.6 Variational formulation of synaptic plasticity.

3.6.1 Synaptic weights update and related potentials.

Let us now formalize the coupled evolution (19) in the context of thermodynamic formalism. The main idea is to make the assumption that at each adaptation step, $\pi^{(T)}_{\omega^{(\tau)}}$ can be approximated by a Gibbs measure $\nu_{\boldsymbol{\psi}^{(\tau)}}$ with potential $\boldsymbol{\psi}^{(\tau)}$ and topological pressure $P\left[\boldsymbol{\psi}^{(\tau)}\right]$. In this case, when T is large, synaptic adaptation writes :

$$\delta W_{ij}^{(\tau)} = \epsilon \nu_{\boldsymbol{\psi}^{(\tau)}}\left[\phi_{ij}(W_{ij}^{(\tau)}, .)\right]. \tag{23}$$

The synaptic update results in a change of parameters $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}^{(\tau+1)} = \boldsymbol{\gamma}^{(\tau)} + \delta\boldsymbol{\gamma}^{(\tau)}$. This induces a variation of the potential $\boldsymbol{\psi}^{(\tau+1)} = \boldsymbol{\psi}^{(\tau)} + \delta\boldsymbol{\psi}^{(\tau)}$ and of the pressure $P\left[\boldsymbol{\psi}^{(\tau+1)}\right] = P\left[\boldsymbol{\psi}^{(\tau)}\right] + \delta P^{(\tau)}$. We now distinguish two situations both arising when synaptic weights evolve.

3.6.2 Smooth variations.

Let us first assume that these variations are smooth, i.e. one stays inside a domain where dynamics is structurally stable and the grammar $G_{\boldsymbol{\gamma}^{(\tau)}}$ is not modified. We assume moreover that the topological pressure is differentiable with respect to the variation $\delta\boldsymbol{\gamma}^{(\tau)}$.

Let us define:

$$\mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}(\mathcal{W}) = P\left[\boldsymbol{\psi}^{(\tau)} + (\mathcal{W} - \mathcal{W}^{(\tau)}).\boldsymbol{\phi}(\mathcal{W}^{(\tau)})\right] - P\left[\boldsymbol{\psi}^{(\tau)}\right], \tag{24}$$

so that $\mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}(\mathcal{W}^{(\tau)}) = 0$. Note that $\mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}(\mathcal{W})$ is *convex*, due to the convexity of the topological pressure (eq. (48) in appendix).

Then, using eq. (49) in appendix, the adaptation rule (23) can be written in the form:

$$\delta\mathcal{W}^{(\tau)} = \epsilon \nabla_{\mathcal{W}=\mathcal{W}^{(\tau)}} \mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}(\mathcal{W}). \tag{25}$$

Since, in this section, pressure P is assumed to be smooth, one has, using (49),(50):

$$\epsilon\left(P\left[\boldsymbol{\psi}^{(\tau)} + \delta\mathcal{W}^{(\tau)}.\boldsymbol{\phi}(\mathcal{W}^{(\tau)})\right] - P\left[\boldsymbol{\psi}^{(\tau)}\right]\right) = \delta\mathcal{W}^{(\tau)}.\left[I + \frac{\epsilon}{2}\kappa^{(\tau)}\right].\delta\mathcal{W}^{(\tau)} + O(\delta\mathcal{W}^{(\tau)^3}), \tag{26}$$

where $\kappa^{(\tau)}$ is the tensor with entries:

$$\kappa^{(\tau)}_{i_1,j_1,i_2,j_2} = C^{(\tau)}_{\phi_{i_1,j_1}\phi_{i_2,j_2}}(0) + 2\sum_{t=1}^{+\infty} C^{(\tau)}_{\phi_{i_1,j_1}\phi_{i_2,j_2}}(t); \qquad i_1, j_1, i_2, j_2 = 1 \ldots N, \tag{27}$$

and $C^{(\tau)}_{\phi_{i_1,j_1}\phi_{i_2,j_2}}(t) = \nu_{\boldsymbol{\psi}^{(\tau)}}\left[\phi_{i_1,j_1} \circ \sigma^t_{\boldsymbol{\gamma}^{(\tau)}} \phi_{i_2,j_2}\right] - \nu_{\boldsymbol{\psi}^{(\tau)}}\left[\phi_{i_1,j_1}\right] \nu_{\boldsymbol{\psi}^{(\tau)}}\left[\phi_{i_2,j_2}\right]$ is the correlation function of $\phi_{i_1,j_1}, \phi_{i_2,j_2}$ for the measure $\nu_{\boldsymbol{\psi}^{(\tau)}}$. Using the explicit form (13) of ϕ_{ij} one can see that $\kappa^{(\tau)}$, is a sum of time correlations between uplets of spikes. This is a version of the fluctuation-dissipation theorem where the response to a smooth variation of the potential $\boldsymbol{\psi}^{(\tau)}$ is given in terms of a series involving the time correlations of the perturbation (Ruelle, 1969, 1999). This series converges provided that dynamics is uniformly hyperbolic.

For sufficiently small ϵ, the matrix $I + \frac{\epsilon}{2}\kappa^{(\tau)}$ is positive and:

$$P\left[\boldsymbol{\psi}^{(\tau)} + \delta\mathcal{W}^{(\tau)}.\boldsymbol{\phi}(\mathcal{W}^{(\tau)})\right] \geq P\left[\boldsymbol{\psi}^{(\tau)}\right].$$

It follows that the variation $\delta^{(\tau)}\mathcal{F}_\phi = \mathcal{F}^{(\tau+1)}_{\boldsymbol{\phi}}(\mathcal{W}^{(\tau+1)}) - \mathcal{F}^{(\tau)}_{\boldsymbol{\phi}}(\mathcal{W}^{(\tau+1)})$ is given by:

$$\delta^{(\tau)}\mathcal{F}_{\boldsymbol{\phi}} = P\left[\boldsymbol{\psi}^{(\tau)}\right] - P\left[\boldsymbol{\psi}^{(\tau)} + \delta\mathcal{W}^{(\tau)}.\boldsymbol{\phi}(\mathcal{W}^{(\tau)})\right] = -\frac{1}{\epsilon}\delta\mathcal{W}^{(\tau)}.\left[I + \frac{\epsilon}{2}\kappa^{(\tau)}\right].\delta\mathcal{W}^{(\tau)} - O(\delta\mathcal{W}^{(\tau)^3})$$

$$\delta^{(\tau)}\mathcal{F}_{\boldsymbol{\phi}} = -\epsilon\nu_{\boldsymbol{\psi}^{(\tau)}}\left[\boldsymbol{\phi}\right]\left[I + \frac{\epsilon}{2}\kappa^{(\tau)}\right].\nu_{\boldsymbol{\psi}^{(\tau)}}\left[\boldsymbol{\phi}\right] - O(\delta\mathcal{W}^{(\tau)^3}) \tag{28}$$

This variation is therefore *negative* when ϵ is sufficiently small.

We come therefore to the following important conclusion. The adaptation rule (12) is a *gradient* system where the function $\mathcal{F}^{(\tau)}_{\boldsymbol{\phi}}$ *decreases* when iterating synaptic adaptation rules. Were the transition $\tau \to \tau+1$ to be smooth for all τ, would $\mathcal{F}^{(\tau)}_{\boldsymbol{\phi}}$ reach a minimum[6] at some $\mathcal{W}^*$ as $\tau \to \infty$. Such a minimum corresponds to $\nabla_{\mathcal{W}^*}\mathcal{F}^{(\tau)}_{\boldsymbol{\phi}} = 0$, thus to $\delta\mathcal{W} = 0$ according to eq. (25). Hence, this minimum corresponds to a *static distribution* for the synaptic weights. Therefore, according to section 3.5.1 the potential $\boldsymbol{\psi}^{(\tau)}$ would converge to the potential (21) as $\tau \to +\infty$. However, we cannot expect these transitions to be smooth for all τ.

3.6.3 Singular variations.

As we saw, during the synaptic adaptation process, the corresponding path in the space of parameters usually crosses bifurcations manifold inducing sharp changes in the dynamics. At those points pressure may not be smooth corresponding to a phase transition. These changes are not necessarily easy to detect numerically, though algorithms allowing to detect phase transitions from finite samples exist, based on rigorous results (Comets, 1997). Indeed, searching "at blind" for all possible kind of phase transitions in so large dimensional dynamical systems, without any idea of what can happen, seems to be desperate, even for model I and II, especially when thinking of what can already happen in one dimensional systems (Beck & Schloegl, 1995). So we shall not discuss within more details this aspect, focusing on grammar changes[7].

[6] Additional constraints are required ensuring that $\mathcal{F}^{(\tau)}_{\boldsymbol{\phi}}$ does not tend to $-\infty$. Typically, such constraints amount to bound the synaptic weights variation (soft or hard-bounds rules) by a suitable choice of functions $h_{ijL_1L_2}$ in eq. (13).

[7] We actually conjecture that the only possible phase transitions in model I and II are grammar changes occurring when crossing the set where $d(\Omega, \mathcal{S}) = 0$ (eq. (6)).

Indeed, from the point of view of neural network analysis, grammar changes implies changes in the type of raster plots that the network is able to produce. An interesting situation occurs when the set of admissible raster plots obtained after adaptation belongs to $\Sigma_{\boldsymbol{\gamma}^{(\tau)}} \cap \Sigma_{\boldsymbol{\gamma}^{(\tau+1)}}$. In this case, adaptation plays the role of a *selective mechanism* where the set of admissible raster plots, viewed as a neural code, is gradually reducing, producing after n steps of adaptation a set $\cap_{m=1}^{n} \Sigma_{\boldsymbol{\gamma}^{(m)}}$ which can be rather small. This has been observed by Soula et al. in (Soula et al., 2006) for model I with Spike Time Dependent Plasticity, though not analyzed in those terms. If we consider the situation where (1) is a neural network submitted to some stimulus, where a raster plot ω encodes the spike response to the stimulus, then $\Sigma_{\boldsymbol{\gamma}}$ is the set of all possible raster plots encoding this stimulus. Adaptation results in a reduction of the possible coding, thus reducing the variability in the possible responses.

3.6.4 Evolution of the potential.

Iterating the adaptation process, one expects to have periods of smooth variations, punctuated by sharp transitions where potential and grammar change. We qualify them under the generic name of "phase transitions" without further specifications. We write $\tau \equiv (l, n)$ where l indexes the phase transition and n the number of epoch since the last phase transition. The adaptation process corresponds now to a sequence $\boldsymbol{\psi}^{(\tau)} \equiv \boldsymbol{\psi}^{(l,n)}$ of Gibbs potentials. Thus, $\boldsymbol{\psi}^{(l,0)}$ is the potential arising just after the l-th transition, with the convention that $\boldsymbol{\psi}^{(l,0)}(\omega) = -\infty$ if ω is forbidden, such that $\boldsymbol{\psi}^{(l,0)}$ characterizes the grammar after the l-th phase transition. We call a *regular period* the succession of adaptation epochs between two phase transitions.

By definition, during a regular period the synaptic update $\mathcal{W}^{(\tau+1)} = \mathcal{W}^{(\tau)} + \delta\mathcal{W}^{(\tau)}$ induces a smooth variation of the potential $\boldsymbol{\psi}^{(\tau+1)} = \boldsymbol{\psi}^{(\tau)} + \delta\boldsymbol{\psi}^{(\tau)}$, and the variational principle (43) selects a new Gibbs measure $\nu_{\boldsymbol{\psi}^{(\tau+1)}}$ with a potential:

$$\boldsymbol{\psi}^{(\tau+1)} = \boldsymbol{\psi}^{(\tau)} + \delta\boldsymbol{\psi}^{(\tau)}, \tag{29}$$

It follows that:

$$\boldsymbol{\psi}^{(l,n)} = \boldsymbol{\psi}^{(l,0)} + \sum_{m=0}^{n} \delta\psi^{(\tau_l+m)}, \tag{30}$$

where τ_l is the epoch where the l-th phase transition aroused. The explicit form of $\delta\boldsymbol{\psi}^{(\tau)}$ (as well as $\boldsymbol{\psi}^{(\tau)}$) depends on the detailed form of the dynamics, and there is little hope to determine it in general, except when the adaptation process converges to a solution with $\delta\mathcal{W} = 0$ (see section 3.5.1). According to section 3.6.2, the function $\mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}$ (eq. (24)) decreases during regular periods.

When a phase transition occurs, the relation (29) does not hold anymore. We now write:

$$\boldsymbol{\psi}^{(l+1,0)} = \boldsymbol{\psi}^{(l,n_l)} + \delta\boldsymbol{\psi}_{reg}^{(l+1,0)} + \delta\boldsymbol{\psi}_{sing}^{(l+1,0)} \tag{31}$$

where n_l is the last epoch before the phase transition $l+1$. $\delta\boldsymbol{\psi}_{reg}^{(l+1,0)}$ contains the regular variations of the potential while $\delta\boldsymbol{\psi}_{sing}^{(l+1,0)}$ contains the singular variations corresponding to a

change of grammar.

We end up with the following picture. During regular periods the grammar does not evolve, the potential changes according to (30), (31), and the function $\mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}$ decreases. Then, there is a sharp change in the grammar and the potential. The function $\mathcal{F}_{\boldsymbol{\phi}}^{(\tau)}$ may also have singularities and may sharply increase. If the adaptation rule converges, the potential (30) converges to $\boldsymbol{\psi}^* = \Phi + \boldsymbol{\lambda}^*.\boldsymbol{\phi}$ (eq. (21)). This potential Φ characterizes the grammar and the statistical weight of allowed raster plots is given by $\boldsymbol{\lambda}^*.\boldsymbol{\phi}$. The potential $\boldsymbol{\psi}^*$ contains all changes in the grammar and statistics arising during the evolution (30,31). Thus it contains the history of the system and the evolution of the grammar.

4 A numerical example.

4.1 Model.

4.1.1 Adaptation rule.

As an example we consider an adaptation rule inspired from (15) with an additional term $r_d W_{ij}^{(\tau)}$, $-1 < r_d < 0$, corresponding to passive LTD.

$$\delta W_{ij}^{(\tau)} = \epsilon \left[r_d W_{ij}^{(\tau)} + \frac{1}{T} \sum_{t=T_s}^{T+T_s} \omega_j^{(\tau)}(t) \sum_{u=-T_s}^{T_s} f(u)\, \omega_i^{(\tau)}(t+u) \right], \tag{32}$$

where $f(x)$ is given by (16) and with:

$$T_s \stackrel{\text{def}}{=} 2\max(\tau_+, \tau_-).$$

Set :

$$S_i(\omega) = \sum_{u=-T_s}^{T_s} f(u)\, \omega_i(u), \tag{33}$$

$$H_{ij}(\omega) = \omega_j(0) S_i(\omega),$$

$\mathbf{H} = \{H_{ij}\}_{i,j=1}^N$, and:

$$\phi_{ij}(W_{ij}, \omega) = r_d W_{ij} + H_{ij}(\omega),$$

with $\boldsymbol{\phi} = \{\phi_{ij}\}_{i,j=1}^N$, where ϕ_{ij} is a finite range potential with range $2T_s$. Then (32) has the form (12), $\delta W_{ij}^{(\tau)} = \epsilon \pi_{\omega^{(\tau)}}^{(T)} [\phi_{ij}(W_{ij}, .)]$.

4.1.2 Effects of the adaptation rule on the synaptic weights distribution.

The term $S_i(\omega)$ (eq. 33) can be either negative, inducing Long Term Depression, or positive inducing Long Term Potentiation. In particular, its average with respect to the empirical measure $\pi_{\omega^{(\tau)}}^{(T)}$ reads:

$$\pi_{\omega^{(\tau)}}^{(T)}(S_i) = \eta r_i(\tau) \tag{34}$$

where:

,,
$$\eta = \left[A_- e^{-\frac{1}{\tau_-}} \frac{1 - e^{-\frac{T_s}{\tau_-}}}{1 - e^{-\frac{1}{\tau_-}}} + A_+ e^{-\frac{1}{\tau_+}} \frac{1 - e^{-\frac{T_s}{\tau_+}}}{1 - e^{-\frac{1}{\tau_+}}} . \right] \tag{35}$$

and where $r_i(\tau) = \pi^{(T)}_{\omega^{(\tau)}}(\omega_i)$ is the frequency rate of neuron i in the τ-th adaptation epoch.

The term η neither depend on ω nor on τ, but only on the adaptation rule parameters $A_-, A_+, \tau_-, \tau_+, T_s$. Equation (34) makes explicit 3 regimes.

- **Cooperative regime.** If $\eta > 0$ then $\pi^{(T)}_{\omega^{(\tau)}}(S_i) > 0$. Then synaptic weights have a tendency to become more positive. This corresponds to a cooperative system (Hirsch, 1989). When iterating adaptation, dynamics become trivial with neurons firing at each time step or remaining quiescent forever.
- **Competitive regime.** On the opposite if $\eta < 0$ synaptic weights become negative. This corresponds to a competitive system (Hirsch, 1989).
- **Intermediate regime.** The intermediate regime corresponds to $\eta \sim 0$. Here no clear cut tendency can be distinguished from the average value of S_i and spikes correlations have to be considered as well.

4.1.3 Static weights.

Thanks to the soft bound term $r_d W_{ij}$ the synaptic adaptation rule admits a static solution given by:

$$W_{ij} = -\frac{\pi^{(T)}_{\omega^{(\tau)}}\left[\omega_j(0) S_i(\omega)\right]}{r_d}. \tag{36}$$

Note that this equation can have several solutions.

Using the same decomposition as the one leading to (34) we obtain:

$$\pi^{(T)}_{\omega^{(\tau)}}\left[\omega_j(0) S_i(\omega)\right] = A_- \sum_{u=-T_s}^{-1} e^{\frac{u}{\tau_-}} \pi^{(T)}_{\omega^{(\tau)}}\left[\omega_j(0)\, \omega_i(u)\right] + A_+ \sum_{u=1}^{T_s} e^{-\frac{u}{\tau_+}} \pi^{(T)}_{\omega^{(\tau)}}\left[\omega_j(0)\, \omega_i(u)\right].$$

Note that $\omega_j(0)\, \omega_i(u) \geq 0$, thus the first term is negative and the second one is positive (see eq. (16)). The sign of W_{ij} depend on the parameters A_-, A_+, T_s, but also on the relative strength of the terms $\pi^{(T)}_{\omega^{(\tau)}}\left[\omega_j(0)\, \omega_i(u)\right]$.

4.1.4 Convergence to the static weights solution.

The synaptic adaptation rule (32) defines a mapping $\mathcal{Q}$ on the set of synaptic weights:

$$W_{ij}^{(\tau+1)} = W_{ij}^{(\tau)}(1 + \epsilon r_d) + \epsilon \pi^{(T)}_{\omega^{(\tau)}}\left[H_{ij}\right] = \mathcal{Q}_{ij}(\mathcal{W}^{(\tau)})$$

with $|1 + \epsilon r_d| < 1$, where $\pi^{(T)}_{\omega^{(\tau)}}$ depends on $\mathcal{W}^{(\tau)}$. Thus,

,,
$$\frac{\partial \mathcal{Q}_{ij}}{\partial W_{kl}} = (1+\epsilon r_d)\delta_{ij,kl} + \epsilon \frac{\partial \pi^{(T)}_{\omega^{(\tau)}}[H_{ij}]}{\partial W_{kl}},$$

where $\delta_{ij,kl} = 1$ if $i = k, j = l$ and 0 otherwise. The second term is a linear response characterizing the variation of $\pi^{(T)}_{\omega^{(\tau)}}[H_{ij}]$ with respect to small variations of W_{kl}. Approximating $\pi^{(T)}_{\omega^{(\tau)}}$ by a Gibbs distribution with a potential $\boldsymbol{\psi}^{(\tau)}$, it writes $C_{\boldsymbol{\psi}^{(\tau)},H_{ij}}(0) + 2\sum_{t=0}^{+\infty} C_{\boldsymbol{\psi}^{(\tau)},H_{ij}}(t)$ where $C_{\boldsymbol{\psi}^{(\tau)},H_{ij}}(t)$ is the time-t correlation function between $\boldsymbol{\psi}^{(\tau)}$ and H_{ij}. When $\nu_{\boldsymbol{\psi}^{(\tau)}}$ is smooth with respect to $\mathcal{W}$, the derivative is dominated by the first term and $\mathcal{Q}$ is contracting for sufficiently small ϵ. This is however not a sufficient condition for the convergence of (32) because $\mathcal{Q}$ is not continuous everywhere. It has jumps whenever $\boldsymbol{\gamma}$ crosses the boundary of a structural stability domain.

If the static solution is contained in a open ball where $\mathcal{Q}$ is continuous the contraction property ensures the convergence to a static solution for any initial condition in this ball (Brouwer theorem).

In this paper we postulate the convergence and check it numerically.

4.1.5 Spike train statistics in a static weights regime.

As emphasized in section 2.6.5 and 3.5.1, when the synaptic adaptation rule converges to a fixed point, the corresponding statistical model is a Gibbs measure with a potential:

$$\boldsymbol{\psi}^* = \Phi + \boldsymbol{\lambda}^*.\boldsymbol{\phi},$$

where Φ contains the grammar and $\boldsymbol{\lambda}$ are free statistical parameters. The value $\boldsymbol{\lambda}^*$ of these parameters in the potential $\boldsymbol{\psi}^*$ is determined by the relation:

$$\left.\frac{\partial P[\boldsymbol{\psi}]}{\partial \lambda_{ij}}\right|_{\boldsymbol{\lambda}^*} = r_d W^*_{ij} + \nu_{\boldsymbol{\psi}^*}[H_{ij}] = 0,\ \forall i,j\,, \tag{37}$$

where the pressure is given by:

$$P[\boldsymbol{\psi}] = r_d \boldsymbol{\lambda}.\mathcal{W} + \lim_{T\to\infty} \frac{1}{T} \log \sum_{\omega\in\Sigma^{(T)}} e^{\boldsymbol{\lambda}.S_T\mathbf{H}(\omega)}.$$

This procedure provides us the explicit form of the raster plot probability distribution when the adaptation rule converges. But the price to pay is that we have to determine *simultaneously* the N^2 parameters λ^*_{ij} on which the Gibbs measure depends. Focusing on the joint probability of a small set of neurons (pairs, triplets) this constraint can be relaxed in order to be numerically tractable.

4.2 Numerical checks.

The main goal of section 4 is to provide an example of the theoretical concepts developed in this paper. The emphasis is however not put on numerical results which will be the main topic of a forthcoming paper. Therefore, we focus here on numerical simulations for model I only, since

the simulations for model II are quite more computer-time consuming, and we consider only one case of η value corresponding to the intermediate regime defined above.

4.2.1 Implementation.

Neurons dynamics. Previous numerical explorations have shown that a Model I- network of N neurons, with synapses taken randomly from a distribution $\mathcal{N}(0, \frac{C^2}{N})$, where C is a control parameter, exhibits a dynamics with very large periods in specific regions of values of the space (ρ, C) (Cessac, 2008; Cessac & Viéville, 2008). On this basis, we choose $N = 100,\ \rho = 0.95,\ C = 4.0, N = 100$. The external current $\mathbf{I}^{(ext)}$ in eq. (5) is given by $I_i^{ext} = 0.06 + 0.01\,\mathcal{N}(0,1)$. Note that fixing a sufficiently large average value for this current avoids a situation where neurons stops firing after a certain time ("neural death").

STDP implementation. An efficient implementation of a STDP rule does not only depend on the analytic form of the rule but also on the respective time scales characterizing neurons and synapses evolutions. In "on-line-protocols", where the synaptic changes occur at the same time scale as neurons dynamics the so-called recursive approach appears to be the more appropriated (see (Zou, 2006) and references therein). In our case, we use instead an offline protocol, where we register the dynamics of the system on a long time windows, then compute the STDP modification, then let the system evolve again to its attractor (see section 3). The offline protocols are very expensive in machine-time, especially when using long spike trains to get a reliable time average $(\pi^{(T)}_{\omega^{(\tau)}}(\phi_{ij}))$. For this reason we need to add some words about our implementation.

We register spike trains in a binary code. Indeed, this is the cheapest way in memory requirements, though it might be expensive for accessing specific spike times. Also, bit-wise operations are faster than their equivalents on other types of data. Finally, there exist very fast methods for computing the number of bits on any specific variable of type of integer (The faster for large number of iterations is a look-up table of precomputed values, but in-line methods - using parallel methods based on masks- are not so far in terms of performances. For details and speed comparison see the G. Manku website http://infolab.stanford.edu/ manku/bitcount/bitcount.html). Numerical comparison of this method with the direct one that records the dynamics in Boolean arrays and compute STDP spike by spike shows enormous performance difference growing exponentially as the length of trains increases.

Computation of $\delta^{(\tau)}\mathcal{F}_\phi$. To check the result in section 3.6.2, stating that $\mathcal{F}^{(\tau)}_{\boldsymbol{\phi}}(\mathcal{W})$ is a decreasing function of τ (i.e. $\delta^{(\tau)}\mathcal{F}_{\boldsymbol{\phi}}$ is negative) during regular periods, we use the following method (Chazottes, 1999).

Fix a potential $\boldsymbol{\psi}$. If T is the length of the experimental raster plot, one divides it into k blocs of length n, such that $T = k \times n$. Call:

$$P_{n,k}(\omega) = \frac{1}{n} \log \left(\frac{1}{k} \sum_{j=0}^{k-1} \exp(S_j^{(n)}(\omega)) \right), \tag{38}$$

with $S_j^{(n)}(\omega) = \sum_{t=jn}^{jn+n-1} \boldsymbol{\psi}(\sigma^t\omega)$. Then, the following result can be proved (Chazottes, 1999):

„ $$P[\psi] = \lim_{n\to+\infty} \lim_{k\to+\infty} P_{n,k}(\omega), \tag{39}$$

for $\nu_{\boldsymbol{\psi}}$ almost-every ω. This computation requires the knowledge of $\boldsymbol{\psi}$.

The result in (Chazottes, 1999) can be extended straightforwardly to the following case. If $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2$ are two potentials for the same grammar, and $\delta\boldsymbol{\psi} = \boldsymbol{\psi}_2 - \boldsymbol{\psi}_1$ then:

$$P[\boldsymbol{\psi}_2] - P[\boldsymbol{\psi}_1] = \lim_{n\to+\infty} \lim_{k\to+\infty} \frac{1}{n} \log \left(\frac{1}{k} \sum_{j=0}^{k-1} \exp(\sum_{t=jn}^{jn+n-1} \delta\boldsymbol{\psi}(\sigma^t\omega)) \right) \tag{40}$$

Since $\delta^{(\tau)}\mathcal{F}_\phi = P\left[\boldsymbol{\psi}^{(\tau)}\right] - P\left[\boldsymbol{\psi}^{(\tau)} + \delta\mathcal{W}^{(\tau)}.\boldsymbol{\phi}(\mathcal{W}^{(\tau)})\right]$,

$$\delta^{(\tau)}\mathcal{F}_\phi = -\lim_{n\to+\infty} \lim_{k\to+\infty} \frac{1}{n} \log \left(\frac{1}{k} \sum_{j=0}^{k-1} \exp\Big(\sum_{t=jn}^{jn+n-1} \delta\mathcal{W}^{(\tau)}.\boldsymbol{\phi}(\mathcal{W}^{(\tau)}, \sigma^t(\omega^{(\tau)}))\Big) \right) \tag{41}$$

Expanding this equation in power series of ϵ one recovers the expansion (26). In particular, subtracting to (41) the leading term (of order $\delta\mathcal{W}^{(\tau)}.\delta\mathcal{W}^{(\tau)}$) one gets an estimation of the "linear response" term $\kappa^{(\tau)}_{i_1,j_1,i_2,j_2}$ in eq. (27). However, on practical grounds, the computation of (41) requires very long time series to reduce significantly the finite fluctuations of the empirical average. Similar problems are encountered in the computation of linear response from other methods (Cessac, 2007).

This result holds for any adaptation rule (any polynomial $\boldsymbol{\phi}$). In the case of the adaptation rule (32), this reduces to computing:

$$\delta^{(\tau)}\mathcal{F}_\phi = -r_d\delta\mathcal{W}^{(\tau)}.\mathcal{W}^{(\tau)} - \lim_{n\to+\infty} \lim_{k\to+\infty} \frac{1}{n} \log \left(\frac{1}{k} \sum_{j=0}^{k-1} \exp\Big(\sum_{t=jn}^{jn+n-1} \epsilon\delta\mathcal{W}^{(\tau)}.\mathbf{H}(\mathcal{W}^{(\tau)}, \sigma^t(\omega^{(\tau)}))\Big) \right).$$

4.2.2 Simulation results.

We have run simulations for STDP parameters values: $r_d = 0.99$, $\epsilon = 0.001$, $\tau_+ = 16$, $\tau_- = 32$, $A_+ = 1.0$, and $T_s = 64$, corresponding to standard values (Izhikevich & Desai, 2003). We fix $\eta = -0.01$ corresponding to the intermediate regime discussed in section 4.1.2. This fixes the value of A_- via eq. (35). The length of spike trains is $T = 3941 = 4096 - 128 = 128 \times 32 - 2 \times T_s$, where 32 is the number of bits in long integer, in our (machine-dependent) implementation. An extended description will be published elsewhere. The main results are summarized in fig. 3.

In this figure (Top) we have represented the evolution of the distribution of synaptic weights $\mathcal{W}^{(\tau)}$ (see captions). For this value of η, the histogram evolves to a unimodal distribution with some nodes having strong synaptic weights. In fig. 3 (center) we depict the raster plots at the beginning and at the end of synaptic adaptation process. In fig. 3 (bottom-left) we have shown the evolution of the Frobenius norm for $\delta\mathcal{W}^{(\tau)}$ (i.e, $\sum_{i,j} |\delta\mathcal{W}^{(\tau)}_{ij}|^2$), which converges to 0. This shows the convergence of the rule in this case. We have also plotted the average weights

,,

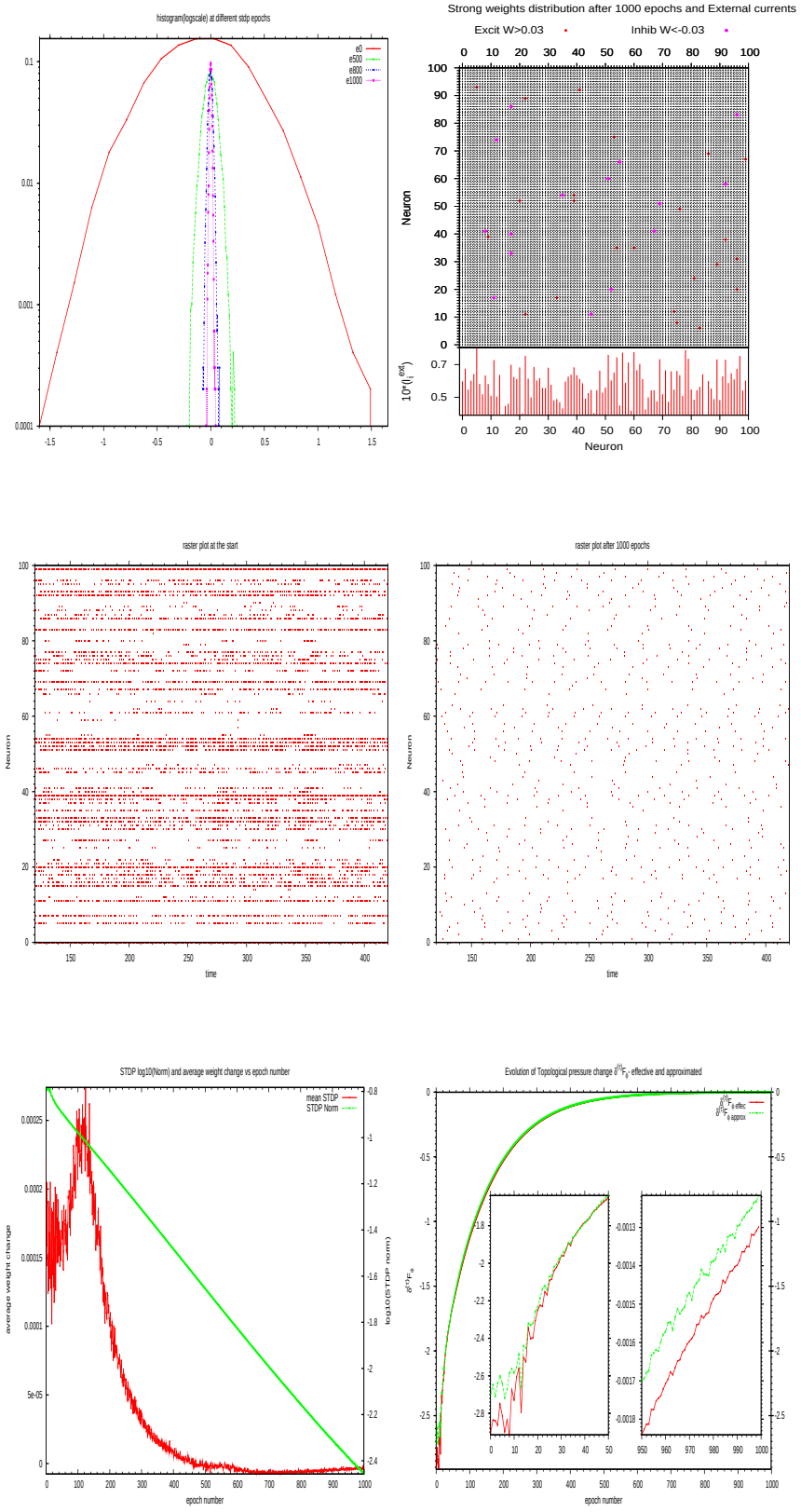

Figure 3: (color on line). The parameters are $N = 100, r_d = -0.4, \eta = -0.01, \epsilon = 0.01$. **Top**:(left) Histogram in log scale of synaptic connectivity in the network after 1000 epochs of STDP;(right) map of strongest connections, either inhibitory (pink dots) or excitatory (red diamonds) (i.e, $|W_{ij}| > 0.03$). We also include at the bottom of this map the external current of each neuron (multiplied by 10). **Middle**:(left) A comparison of raster plots at the beginning and (right) the end of the synapses evolution. **Bottom**:(left) Evolution of the mean synaptic change $\frac{1}{N^2}\sum_{i,j}\delta\mathcal{W}_{ij}^{(\tau)}$(red) and the Frobenius norm for STDP matrix on log scale i.e, $\log_{10}\left(\sum_{i,j}|\delta\mathcal{W}_{ij}^{(\tau)}|^2\right)$ (green); (right) Variations of Topological Pressure (with 2 zoom graphics showing sharp changes) after each epoch as calculated explicitly in eq. (41) with 512 trains of size 512 (red), and first order term approximated by $\frac{1}{\epsilon}\delta\mathcal{W}^{(\tau)}.\delta\mathcal{W}^{(\tau)}$(green).

modification $\frac{1}{N^2}\sum_{i,j}\delta W_{ij}$. Finally in fig. 3 (bottom -right) is represented $\delta^{(\tau)}\mathcal{F}_\phi$ computed with 512 block of size 512. We see clearly that this quantity is negative as expected from eq. (28) and tends to 0. We have also represented the term $-\frac{1}{\epsilon}\delta\mathcal{W}^{(\tau)}.\delta\mathcal{W}^{(\tau)}$ corresponding to the first order expansion in eq. (28).

5 Conclusion.

In this paper, we have considered the questions of characterizing the statistics of spike trains generated by neuronal networks, and how these statistics evolve during synaptic adaptation, from the angle of dynamical systems theory and ergodic theory. We have introduced a framework where notions such as raster plots, empirical averages of observables, and synaptic plasticity rules, coming from neuroscience, can be formalized in the language of dynamical systems. In this context, we propose a strategy to produce optimal statistical models of spike trains, based on the maximization of statistical entropy where expectation of observables must be compatible with their empirical averages. These models are Gibbs measures. Considering adaptation rules with slow dynamics, we also suggest that such Gibbs measures may result from related adaptation rules whenever synaptic weights converge to a fix value. In this case, synaptic weights resume the whole history of the neuronal network contained in the structure of the generating function of cumulants (the topological pressure) and in the structure of allowed/forbidden raster plots (grammar).

To our opinion this theoretical paper is a beginning. Our hope is now to apply this strategy to data coming from biological neuronal networks, while only formal neural networks were considered here. At the present stage there is indeed a crucial issue to be able to characterize spike train statistics and to discriminate statistical models. For example, it is still an open problem to decide, for a given spike train in a given experiment, whether firing rate are sufficient for the statistical characterization (leading to uncorrelated Poisson models) or if higher order orders statistics such as correlations are also necessary. This is more than an academic question. Beyond the choice of such or such statistical model, there are hypotheses on how neuronal network convert stimuli from the external word into an action. Discriminating statistical models is the way to select the best hypothesis.

Our approach opens up the possibility of producing numerical methods for such a discrimination, e.g. based on Kullback-Leibler divergence minimization obtained from eq. (9). Some of these methods are already available as open source code on http://enas.gforge.inria.fr/. They are based on the theory developed in the present paper, relying itself on ergodic theory and thermodynamic formalism. At the present stage, the overwhelming richness of thermodynamic formalism has not really been considered in neuroscience. We hope that this paper will attract the attention of the neuroscience community on this point.

Acknowledgments. This work has been supported by the INRIA ARC MACACC and the Doeblin CNRS federation. We thank the Reviewer for helpful remarks, useful references and constructive criticism.

6 Appendix

In this appendix we present the thermodynamic formalism material used in the paper. Though all "quantities defined below depend on $\boldsymbol{\gamma}$, we have dropped this dependence, which is not central here, in order to alleviate notations. We follow (Keller, 1998; Parry & Pollicott, 1990; Chazottes et al., 1998; Chazottes, 1999; Chazottes & Keller, 2009) in the presentation of Gibbs measures in ergodic theory. In accordance with the body of the text Σ is the set of admissible raster plots and we denote by $\mathcal{B}$ the related Borel sigma-field.

6.1 Ergodic measures.

Fix ϕ a continuous function. The time average of ϕ along the orbit of $\mathbf{X}$ is given by

$$\pi_{\mathbf{X}}(\phi) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \phi(\mathbf{X}(t)).$$

If μ is an invariant measure[8], then, according to Birkhoff theorem, this limit exists for μ-almost every $\mathbf{X}$. Standard theorems in ergodic theory ensure that (1) has at least one invariant measure (Katok & Hasselblatt, 1998) but there are typically many. Among them, ergodic measures play a distinguished role. An invariant measure μ is ergodic if any real measurable invariant function is μ-almost surely constant. As a corollary of Birkhoff's theorem $\pi_{\mathbf{X}}(\phi) = \mu(\phi)$, for μ-almost every $\mathbf{X}$, where $\mu(\phi) = \int \phi d\mu$ is the expectation of ϕ with respect to μ. Namely, the empirical average $\pi_{\mathbf{X}}(\phi)$ is equal to the "ensemble" average $\mu(\phi)$ whatever the initial condition, provided it is chosen in the support of μ. Any invariant measure can be written as a convex decomposition of ergodic measures.

6.2 Raster plots statistics.

Fix $n > 0$, a set of times $t_1, \dots t_n$, and a set of prescribed spiking patterns $\boldsymbol{\omega}(t_1) \dots \boldsymbol{\omega}(t_n)$. The set of raster plots $\omega' \in \Sigma$ such that $\boldsymbol{\omega}'(t_k) = \boldsymbol{\omega}(t_k), k = 1 \dots n$ contains therefore raster plots where the spiking patterns at prescribed times $t_1 \dots t_n$ are imposed (cylinder set). By (countable) intersection and union of cylinder sets one can generate all possible *events* in Σ, such as "neuron i is firing at time t_1", or "neuron i_1 is firing at time t_1, and neuron i_2 is firing at time t_2, ... and neuron i_n is firing at time t_n", etc ....

The statistical properties of raster plots are inherited from the statistical properties of orbits of (1), via the correspondence $\nu[C] = \mu\left[\{\mathbf{X} \rightharpoonup \omega, \omega \in C\}\right]$, where C is a cylinder set. On practical grounds, these statistical properties are obtained by empirical average. The probability of a cylinder set C is given by:

$$\pi_{\omega}^{(T)}(C) = \frac{1}{T} \sum_{t=1}^{T} \chi(\sigma^t \in C) \tag{42}$$

[8] Namely $\mu(\mathbf{F}_{\boldsymbol{\gamma}}{}^{-1}A) = \mathbf{F}_{\boldsymbol{\gamma}}(A)$ for any measurable set $A \subset \mathcal{B}$.

where χ is the indicatrix function, while the time average of some function $\phi : \Sigma \to \mathbb{R}$ is given by:

" $$\pi_\omega^{(T)}(\phi) = \frac{1}{T}\sum_{t=1}^{T}\phi(\sigma^t\omega).$$

When ν is ergodic the following holds.

$$\nu = \pi_\omega \stackrel{\text{def}}{=} \lim_{T\to\infty}\pi_\omega^{(T)}(.),$$

for ν-almost every ω, where $\pi_\omega(.)$ is called the *empirical measure* for the raster plot ω.

6.3 Thermodynamic formalism.

6.3.1 Working assumptions.

In this paper, we have adopted a pragmatic approach based on the idea that statistical models obtained via the applications of principles in thermodynamic formalism (which are basically the same as those used in statistical physics) could constitute good prototypes for the analysis of real spike trains statistics. In this way, our goal is not to prove theorems but instead to use already proved theorems. However, theorems have hypotheses that we now explicit and whose plausible validity is discussed.

Basically, to apply the machinery described below we need first to assume that raster plots constitute a symbolic coding for the orbits of (1), i.e. there is a one-to-one correspondence between membrane potential trajectories and raster plots, except for a negligible subset of initial conditions. We also assume that symbolic trajectories (raster plots) are generated by a finite grammar. This hypothesis is discussed in section 2.6.1, and it can actually be rigorously established for models I and II without noise. The case with noise is under current investigations. The corresponding transition matrix G defines a sub-shift of finite type or topological Markov chain. Actually, part of the results presented below are extensions of standard results on Markov chains. We also assumed that G is *primitive* i.e. $\exists n_0 > 0$ such that $\forall i, j$, $\forall n > n_0$ $(G^n)_{ij} > 0$. This amounts to assuming that the dynamics on each attractor is topologically mixing[9].

All these hypotheses hold if the dynamical system generating the raster plots is continuous and uniformly hyperbolic. Following the definition of (Parry & Pollicott, 1990), a map $\mathbf{F} : \mathcal{M} \to \mathcal{M}$ is uniformly hyperbolic on a set Λ if:

1. There exists a splitting of the tangent space in unstable (E^u) and stable (E^s) subspaces such that there exists $C, \mu > 0$ with $\|D\mathbf{F}^t\|_{E^s}, \|D\mathbf{F}^{-t}\|_{E^u} \leq C\mu^t$, $t > 0$.

2. Λ contains a dense orbit.

3. The periodic orbits in Λ are dense (and Λ consists of more than a single closed orbit).

4. There exists an open set $U \supset \Lambda$ such that $\Lambda = \bigcap_{t=-\infty}^{\infty} F^t U$.

[9] A dynamical system $(\mathbf{F}, \mathcal{M})$ is topologically mixing if for any pair of open sets $U, V \subset \mathcal{B}$, there exists $n_0 \geq 0$ such that, $\forall n > n_0$, $\mathbf{F}^n U \cap V \neq \emptyset$

Adding noise to model I, II renders them uniformly hyperbolic (see footnote 4 in the text) but they are not continuous due to the singularity of the firing threshold. Hence, additional work is necessary to prove that there is a finite Markov partition (Cessac & Fernandez, in preparation). Throughout this Appendix we assume that all these hypotheses hold.

6.3.2 Potentials.

Fix $0 < \Theta < 1$. We define a metric on Σ, the set of raster plots generated by the grammar G, by $d_\Theta(\omega, \omega') = \Theta^p$, where p is the largest integer such that $\boldsymbol{\omega}(t) = \boldsymbol{\omega}(t'), 0 \leq t \leq p-1$. (The related topology thus structures the fact that raster plots coincide until $t = p-1$).

For a function $\boldsymbol{\psi} : \Sigma \to \mathbb{R}$ and $n \geq 1$ define

$$var_n \boldsymbol{\psi} = \sup \left\{ |\boldsymbol{\psi}(\omega) - \boldsymbol{\psi}(\omega')| : \boldsymbol{\omega}(t) = \boldsymbol{\omega}'(t), 0 \leq i < n \right\}.$$

We denote by $C(\Sigma)$ the space of functions $\boldsymbol{\psi}$ such that $var_n \boldsymbol{\psi} \to 0$ as $n \to \infty$. This is the set of continuous real functions on Σ.

A function $\boldsymbol{\psi} \in C(\Sigma)$ is called a *potential*. A potential is *regular*[10] if $\sum_{n=0}^{+\infty} var_n(\boldsymbol{\psi}) < \infty$. Equivalently, $\boldsymbol{\psi}$ is Hölder. For a positive integer r a *range-r* potential is a potential such that $\boldsymbol{\psi}(\omega) = \boldsymbol{\psi}(\omega')$, if $\boldsymbol{\omega}(t) = \boldsymbol{\omega}'(t), 0 \leq t < r$. That is, $\boldsymbol{\psi}$ depends only on the r first spiking patterns $\boldsymbol{\omega}(0), \ldots, \boldsymbol{\omega}(r-1)$. Thus, a finite range potential is necessarily regular. Examples are $\omega_{i_1}(0)$ (range-1 potential), $\omega_{i_1}(0)\omega_{i_2}(k)$ (range-k potential), $\omega_{i_1}(0)\omega_{i_2}(k)\omega_{i_3}(l)$ (range-$\max(k,l)$ potential), etc, ... The potential ϕ_{ij} introduced in section 3.2 is a range-T_s potential. More generally, all potential introduced in the paper have finite range.

A specific example of potential, used in this paper, corresponds to integrating the grammar into a potential Φ such that $\Phi(\omega) = 0$ is ω is allowed, and $\Phi(\omega) = -\infty$ otherwise. In this case, however, the potential is not continuous anymore but only upper semi-continuous.

6.3.3 Equilibrium states

Let ν be a σ-invariant measure and $\Sigma^{(n)}$ the set of admissible cylinders of lenght n. Call:

$$h^{(n)}[\nu] = -\sum_{\omega \in \Sigma^{(n)}} \nu\left([\omega]_{0,n-1}\right) \log \nu\left([\omega]_{0,n-1}\right).$$

Then,

$$h[\nu] = \lim_{n \to \infty} \frac{h^{(n)}[\nu]}{n}$$

is the entropy of ν. Let $m^{(inv)}$ be the set of invariant measures for σ, then the *pressure* of a potential $\boldsymbol{\psi}$ is defined by:

$$P[\boldsymbol{\psi}] = \sup_{\nu \in m^{(inv)}} \left(h[\nu] + \nu[\boldsymbol{\psi}] \right). \tag{43}$$

[10] This condition is analogous to the potential decay ensuring the existence of a thermodynamic limit in statistical mechanics (Ruelle, 1969; Meyer, 1980). In the present context it ensures that the Gibbs measure related to the potential $\boldsymbol{\psi}$ is unique and that it is also the unique equilibrium state.

This supremum is attained- not necessarily at a unique measure- where:

$$P\left[\boldsymbol{\psi}\right] = h\left[\nu_{\boldsymbol{\psi}}\right] + \nu_{\boldsymbol{\psi}}\left[\boldsymbol{\psi}\right].$$

"

$\nu_{\boldsymbol{\psi}}$ is called an *equilibrium state*, as it maximises some version of the (neg) free energy. Equivalently, a measure ν is an equilibrium state for a potential $\boldsymbol{\psi} \in C(\Sigma)$ if and only if, for all potential $\boldsymbol{\eta} \in C(\Sigma)$:

$$P\left[\boldsymbol{\psi} + \boldsymbol{\eta}\right] - P\left[\boldsymbol{\psi}\right] \geq \nu\left[\boldsymbol{\psi}\right] \tag{44}$$

This means that ν is a tangent functional for P at $\boldsymbol{\psi}$. In particular $\boldsymbol{\psi}$ has a unique equilibrium state if and only if P is differentiable at $\boldsymbol{\psi}$, i.e. $\lim_{t\to 0} \frac{1}{t}\left(P\left[\boldsymbol{\psi} + t\boldsymbol{\phi}\right] - P\left[\boldsymbol{\psi}\right] = \nu(\boldsymbol{\psi})\right)$, for all $\boldsymbol{\phi} \in C(\Sigma)$.

6.3.4 Properties of the topological pressure.

We review some important properties of the topological pressure used in this paper. Basically, the topological pressure has the same properties of thermodynamic potentials like free energy. First,

$$P\left[\boldsymbol{\psi}\right] = \limsup_{n\to\infty} \frac{1}{n} \log(Z_n(\boldsymbol{\psi})), \tag{45}$$

where $Z_n(\psi)$ is a partition function:

$$Z_n(\boldsymbol{\psi}) = \sum_{\omega\in\Sigma^{(n)}} e^{S^{(n)}\boldsymbol{\psi}(\omega)}, \tag{46}$$

and

$$S^{(n)}\boldsymbol{\psi}(\omega) = \sum_{t=0}^{n-1} \boldsymbol{\psi}(\sigma^t\omega). \tag{47}$$

Also, the topological pressure is a convex functional of the potentials i.e.

$$P\left[\alpha\boldsymbol{\psi}_1 + (1-\alpha)\boldsymbol{\psi}_2\right] \leq \alpha P\left[\boldsymbol{\psi}_1\right] + (1-\alpha)P\left[\boldsymbol{\psi}_2\right]; \quad \alpha \in [0,1]. \tag{48}$$

Like the free energy, the topological pressure is a generating functional (when it is differentiable). For example,

$$\left.\frac{\partial P\left[\boldsymbol{\psi} + \alpha\boldsymbol{\phi}\right]}{\partial\alpha}\right|_{\alpha=0} = \nu_{\boldsymbol{\psi}}[\boldsymbol{\phi}], \tag{49}$$

and:

$$\left.\frac{\partial^2 P\left[\boldsymbol{\psi} + \alpha_1\boldsymbol{\phi}_1 + \alpha_2\boldsymbol{\phi}_2\right]}{\partial\alpha_1\partial\alpha_2}\right|_{\alpha_1=\alpha_2=0} = C_{\boldsymbol{\phi}_1\boldsymbol{\phi}_2}(0) + 2\sum_{t=0}^{+\infty} C_{\boldsymbol{\phi}_1\boldsymbol{\phi}_2}(t), \tag{50}$$

where:

$$C_{\boldsymbol{\phi}_1\boldsymbol{\phi}_2}(t) = \nu_{\boldsymbol{\psi}}\left(\boldsymbol{\phi}_1 \circ \sigma^t\boldsymbol{\phi}_2\right) - \nu_{\boldsymbol{\psi}}(\boldsymbol{\phi}_1)\nu_{\boldsymbol{\psi}}(\boldsymbol{\phi}_2),$$

is the correlation function of ϕ_1, ϕ_2 at time t.

6.3.5 Gibbs measure.

Assume that the transition matrix of the shift σ is primitive and that $\boldsymbol{\psi}$ is regular. Then, there is a unique ergodic measure $\nu_{\boldsymbol{\psi}}$, called a Gibbs measure, for which one can find some constants c_1, c_2 with $0 < c_1 \leq 1 \leq c_2$ such that for all $n \geq 1$ and for all $\omega \in \Sigma$:

$$c_1 \leq \frac{\nu_{\boldsymbol{\psi}}\left(\omega \in [\omega]_{0,n-1}\right)}{\exp(-nP\left[\boldsymbol{\psi}\right] + S^{(n)}\boldsymbol{\psi}(\omega))} \leq c_2. \tag{51}$$

where $S^{(n)}\boldsymbol{\psi}(\omega)$ is given by (47). $\nu_{\boldsymbol{\psi}}$ is also the unique equilibrium state.

6.3.6 Ruelle-Perron-Frobenius operator.

The Ruelle-Perron-Frobenius (RPF) operator for the potential $\boldsymbol{\psi}$, denoted by $L_{\boldsymbol{\psi}}$, acts on functions $g \in C(\Sigma)$ as $L_{\boldsymbol{\psi}}g(\omega) = \sum_{\omega',\, \omega=\sigma\omega'} e^{\boldsymbol{\psi}(\omega')}g(\omega')$. This operator has a unique maximal eigenvalue $s = e^{P[\boldsymbol{\psi}]}$ associated to a right eigenfunction $b_{\boldsymbol{\psi}}$ and a left eigenfunction $\rho_{\boldsymbol{\psi}}$ (probability measure) such that $L_{\boldsymbol{\psi}}b_{\boldsymbol{\psi}}(\omega) = sb_{\boldsymbol{\psi}}(\omega)$ and $\int L_{\boldsymbol{\psi}}vd\rho_{\boldsymbol{\psi}} = s\int vd\rho_{\boldsymbol{\psi}}$, for all $v \in C(\Sigma)$. The remaining part of the spectrum is located in a disk in the complex plane, of radius strictly lower than s. Finally, for all $v \in C(\Sigma)$

$$\frac{1}{s^n}L^n_{\boldsymbol{\psi}}v \to b_{\boldsymbol{\psi}}\int vd\rho_{\boldsymbol{\psi}}.$$

The Gibbs measure is $\nu_{\boldsymbol{\psi}} = b_{\boldsymbol{\psi}}\rho_{\boldsymbol{\psi}}$.

Given $\boldsymbol{\psi}$ and knowing its right eigenfunction $b_{\boldsymbol{\psi}}$ and pressure $P\left[\boldsymbol{\psi}\right]$ one can construct a new potential:

$$\boldsymbol{\Psi}(\omega) = \boldsymbol{\psi}(\omega) - \log(b_{\boldsymbol{\psi}}(\sigma\omega)) + \log(b_{\boldsymbol{\psi}}(\omega)) - P\left[\boldsymbol{\psi}\right], \tag{52}$$

such that $L_{\boldsymbol{\Psi}}\mathbf{1} = 1$, where $\mathbf{1}$ is the constant function $\mathbf{1}(\omega) = 1$. Such a potential is called "normalised". Its pressure is zero.

If $\boldsymbol{\psi}$ has a finite range-r then the RPF operator reduces to the transition matrix of a $(r+1)$-step Markov chain. Thus, $b_{\boldsymbol{\psi}}$ and $\rho_{\boldsymbol{\psi}}$ can be easily determined. The Gibbs measure is none other than the invariant mesure of this chain. This can be used to generate a raster plot distributed according to a Gibbs distribution with a potential $\boldsymbol{\psi}$. Moreover, for $\nu_{\boldsymbol{\psi}}$-almost-every raster plot ω:

$$\lim_{T\to+\infty} \frac{\pi_\omega^{(T)}\left[j\boldsymbol{\omega}(1)\ldots\boldsymbol{\omega}(r-1)\right]}{\pi_\omega^{(T)}\left[\boldsymbol{\omega}(1)\ldots\boldsymbol{\omega}(r)\right]} = e^{\boldsymbol{\psi}(j\boldsymbol{\omega}(1)\ldots\boldsymbol{\omega}(r-1))} \tag{53}$$

This allows to obtain of $\boldsymbol{\psi}$ numerically (Chazottes et al., 1998).

6.3.7 Kullack-Leibler divergence.

There is a last important property. Let μ be an invariant measure and $\nu_{\boldsymbol{\psi}}$ a Gibbs measure with a potential $\boldsymbol{\psi}$, both defined on the same set of sequences Σ. Let

$$d(\mu, \nu_{\boldsymbol{\psi}}) = \limsup_{n\to\infty} \frac{1}{n} \sum_{[\omega]_{0,n-1}} \mu\left([\omega]_{0,n-1}\right) \log\left[\frac{\mu\left([\omega]_{0,n-1}\right)}{\nu_{\boldsymbol{\psi}}\left([\omega]_{0,n-1}\right)}\right]. \tag{54}$$

be the relative entropy (or Kullack-Leibler divergence) between μ and ν. Then,

$$d\left(\mu, \nu_{\boldsymbol{\psi}}\right) = P\left[\boldsymbol{\psi}\right] - \mu(\boldsymbol{\psi}) - h(\mu). \tag{55}$$

If $\mu = \nu$, $d(\mu, \nu) = 0$. The converse may not be true if the potential is not regular.

If a raster ω is typical for the Gibbs measure $\nu_{\boldsymbol{\psi}}$ then one expects that $\pi_\omega^{(T)}$ becomes closer to $\nu_{\boldsymbol{\psi}}$ than any other Gibbs measure (for another potential $\boldsymbol{\psi}'$) as T growths. This provides a criterion to compare two Gibbs measures (and to discrimate between several statistical models). Indeed, the following theorem holds (Chazottes et al., 1998).

Theorem 2 *For any pair of distinct*[11] *regular potentials $\boldsymbol{\phi}, \boldsymbol{\psi}$, there exists an integer $N \equiv N(\boldsymbol{\phi}, \boldsymbol{\psi}, \omega)$ such that, for all $T \geq N$,*

$$d\left(\pi_\omega^{(T)}, \nu_{\boldsymbol{\psi}}\right) < d\left(\pi_\omega^{(T)}, \nu_{\boldsymbol{\phi}}\right) \tag{56}$$

for $\nu_{\boldsymbol{\psi}}$-almost every ω.

This says that $d\left(\pi_\omega^{(T)}, \nu_{\boldsymbol{\psi}}\right)$ becomes, for sufficiently large T, smaller than the Kullback-Leibler divergence between $\pi_\omega^{(T)}$ and any other Gibbs measure.

[11] Non cohomologous.

References

Adrian, E., & Zotterman, Y. (1926). The impulses produced by sensory nerve endings: Part ii: " The response of a single end organ. *J Physiol (Lond.)*, *61*, 151-71.

Amit, D. J. (1989). *Modeling brain function—the world of attractor neural networks*. New York, NY, USA: Cambridge University Press.

Arabzadeh, E., Panzeri, S., & Diamond, M. (2006). Deciphering the spike train of a sensory neuron: Counts and temporal patterns in the rat whisker pathway. *The Journal of Neuroscience*, *26*(36), 9216-9226.

Artola, A., Bröcher, S., & Singer, W. (1990). Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, *347*(6288), 69–72.

Barbieri, R., Frank, L. M., Nguyen, D. P., Quirk, M. C., Wilson, M. A., & Brown, E. N. (2004). Dynamic analyses of information encoding in neural ensembles. *Neural Computation*, *16*, 277-307.

Beck, C., & Schloegl, F. (1995). *Thermodynamics of chaotic systems: An introduction*. Cambridge: Cambridge University Press.

Bi, G., & Poo, M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, *24*, 139–166.

Bienenstock, E. L., Cooper, L., & Munroe, P. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience*, *2*(1), 32-48.

Blanchard, P., Cessac, B., & Krueger, T. (2000). What can one learn about self-organized criticality from dynamical system theory ? *Journal of Statistical Physics*, *98*, 375–404.

Bliss, T., & Gardner-Medwin, A. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the unanaesthetised rabbit following stimulation of the perforant path. *J Physiol*, *232*, 357-374.

Bohte, S. M., & Mozer, M. C. (2007). Reducing the variability of neural responses: A computational theory of spike-timing-dependent plasticity. *Neural Computation*, *19*(2), 371–403.

Bowen, R. (1975). *Equilibrium states and the ergodic theory of anosov diffeomorphisms* (Vol. 470). New York: Springer-Verlag.

Bowen, R. (2008). *Equilibrium states and the ergodic theory of anosov diffeomorphisms. second revised version.* Springer-Verlag.

Brette, R., & Gerstner, W. (2005). Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of Neurophysiology*, *94*, 3637–3642.

Cessac, B. (2007). Does the complex susceptibility of the hénon map have a pole in the upper-half plane ? a numerical investigation. *Nonlinearity*, *20*, 2883–2895.

Cessac, B. (2008). A discrete time neural network model with spiking neurons. rigorous results on the spontaneous dynamics. *J. Math. Biol.*, *56*(3), 311-345.

Cessac, B., Blanchard, P., Krueger, T., & Meunier, J. (2004). Self-organized criticality and thermodynamic formalism. *Journal of Statistical Physics*, *115*(516), 1283–1326.

Cessac, B., & Samuelides, M. (2007). From neuron to neural networks dynamics. *EPJ Special topics: Topics in Dynamical Neural Networks*, *142*(1), 7–88.

Cessac, B., Vasquez, J., & Viéville, T. (2009). Parametric estimation of spike train statistics. *submitted*.

Cessac, B., & Viéville, T. (2008, jul). On dynamics of integrate-and-fire neural networks with adaptive conductances. *Frontiers in neuroscience*, *2*(2).

Chazottes, J. (1999). *Entropie relative, dynamique symbolique et turbulence.* Unpublished doctoral dissertation, Université de Provence - Aix Marseille I.

Chazottes, J., Floriani, E., & Lima, R. (1998). Relative entropy and identification of gibbs measures in dynamical systems. *J. Statist. Phys.*, *90*(3-4), 697-725.

Chazottes, J., & Keller, G. (2009). Pressure and equilibrium states in ergodic theory. In E. of Complexity & S. Science (Eds.), (chap. Ergodic Theory). Springer.

Chechik, G. (2003). Spike-timing-dependent plasticity and relevant mutual information maximization. *Neural Computation*, *15*(7), 1481–1510.

Collet, P., Galves, A., & Lopez, A. (1995). Maximum likelihood and minimum entropy identification of grammars. *Random and Computational Dynamics*, *3*(3/4), 241-250.

Comets, F. (1997). Detecting phase transition for gibbs measures. *Ann. Appl. Probab.*, *7*(2), 545-563.

Cooper, L., Intrator, N., Blais, B., & Shouval, H. (2004). *Theory of cortical plasticity.* World Scientific, Singapore.

Cronin, J. (1987). *Mathematical aspects of hodgkin-huxley theory.* Cambridge University Press.

Daucé, E., Quoy, M., Cessac, B., Doyon, B., & Samuelides, M. (1998). Self-organization and dynamics reduction in recurrent networks: stimulus presentation and learning. *Neural Networks*, *11*, 521–33.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience : Computational and mathematical modeling of neural systems.* MIT Press.

Dayan, P., & Hausser, M. (2004). *Plasticity kernels and temporal statistics* (Vol. 16; S. Thrun, L. Saul, & B. Schoelkopf, Eds.). Cambridge MA: MIT Press.

Delorme, A., Perrinet, L., & Thorpe, S. (2001). Networks of integrate-and-fire neurons using rank order coding b: Spike timing dependent plasticity and emergence of orientation selectivity. *Neurocomputing*, *38-40*, 539-45.

Dudek, S., & Bear, M. F. (1993). Bidirectional long-term modification of synaptic effectiveness in the adult and immature hippocampus. *J Neurosci.*, *13*(7), 2910–2918.

FitzHugh, R. (1955). Mathematical models of threshold phenomena in the nerve membrane. *Bull. Math. Biophysics*, *17*, 257–278.

FitzHugh, R. (1961). Impulses and physiological states in models of nerve membrane. *Biophys. J.*, *1*, 445-466.

Gao, Y., Kontoyiannis, I., & Bienenstock, E. (2008). Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, *10*(2), 71-99.

Georgeopoulos, A. P., Merchant, H., Naselaris, T., & Amirikian, B. (2007). Mapping of the preferred direction in the motor cortex. *PNAS*, *104*(26), 11068-11072.

Georgopoulos, A., Kalaska, J., Caminiti, R., & Massey, J. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primary motor cortex. *J Neurosci*, *2*(1527-1537).

Gerstner, W., & Kistler, W. (2002b). *Spiking neuron models.* Cambridge University Press.

Gerstner, W., & Kistler, W. M. (2002a). Mathematical formulations of hebbian learning. *Biological Cybernetics*, *87*, 404–415.

Grammont, F., & Riehle, A. (1999). Precise spike synchronization in monkey motor cortex ” involved in preparation for movement. *Exp. Brain Res.*, *128*, 118–122.

Grammont, F., & Riehle, A. (2003). Spike synchronization and firing rate in a population of motor cortical neurons in relation to movement direction and reaction time. *Biol Cybern*, *88*, 360-373.

Guckenheimer, J., & Labouriau, I. S. (1993). Bifurcation of the hodgkin-huxley equations: A new twist. *Bull. Math. Biol.*, *55*, 937-952.

Hebb, D. (1949). *The organization of behavior: a neuropsychological theory.* Wiley, NY.

Hirsch, M. (1989). Convergent activation dynamics in continuous time networks. *Neur. Networks*, *2*, 331–349.

Hodgkin, A., & Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, *117*, 500–544.

Izhikevich, E. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(6), 1569–1572.

Izhikevich, E. (2004, September). Which model to use for cortical spiking neurons? *IEEE Trans Neural Netw*, *15*(5), 1063–1070.

Izhikevich, E., & Desai, N. (2003). Relating stdp to bcm. *Neural Computation*, *15*, 1511–1523.

Jaynes, E. (1957). Information theory and statistical mechanics. *Phys. Rev.*, *106*(620).

Ji, C. (1989). Estimating functionals of one-dimensional gibbs states. *Probab. Theory Related Fields*, *82*(2), 155-175.

Johnson. (1980). Sensory discrimination: neural processes preceding discrimination decision. *J Neurophysiol*, *43*(6), 1793-1815.

Johnson, D. (2004). Neural population structure and consequences for neural coding. *Journal of Computational Neuroscience*, *16*(1), 69-80.

Jolivet, R., Rauch, A., Lescher, H.-R., & Gerstner, W. (2006). *”integrate-and-fire models with adaptation are good enough”*. MIT Press, Cambridge.

Kang, K., & Amari, S. ichi. (2008). Discrimination with spike times and isi distributions. *Neural Computation*, *20*, 1411-1426.

Katok, A., & Hasselblatt, B. (1998). *Introduction to the modern theory of dynamical systems.* Kluwer.

Keller, G. (1998). *Equilibrium states in ergodic theory.* Cambridge University Press.

Levy, W., & Stewart, D. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, *8*(4), 791-797.

Malenka, R. C., & Nicoll, R. A. (1999). Long-term potentiation - a decade of progress ? *Science*, *285*(5435), 1870 - 1874.

Malsburg, C. von-der. (1973). Self-organisation of orientation sensitive cells in the striate cortex. *Kybernetik*, *14*, 85–100.

Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic ap and epsp. *Science*, *275*(213).

Meyer, D. (1980). *The ruelle-araki transfer operator in classical statistical mechanics* (Vol. 123; L. N. in Physics, Ed.). Springer-Verlag.

Miller, K., Keller, J., & Stryker, M. (1989). Ocular dominance column development: analysis and simulation. *Science*, *245*(4918), 605–615.

Nagumo, J., Arimoto, S., & Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. *Proc.IRE*, *50*, 2061–2070.

Nemenman, I., Lewen, G., Bialek, W., & Steveninck, R. de Ruyter van. (2006). Neural coding of a natural stimulus ensemble: Information at sub-millisecond resolution. *PLoS Comp Bio*, *4*, e1000025.

Nirenberg, S., & Latham, P. (2003). Decoding neuronal spike trains: how important are correlations. *Proceeding of the Natural Academy of Science*, *100*(12), 7348–7353.

Osbone, L., Palmer, S., Lisberger, S., & Bialek, W. (2008). Combinatorial coding in neural populations. *arXiv.org:0803.3837*.

Parry, W., & Pollicott, M. (1990). *Zeta functions and the periodic orbit structure of hyperbolic dynamics* (Vols. 187–188). Asterisque.

Perrinet, L., Delorme, A., Samuelides, M., & Thorpe, S. (2001). Networks of integrate-and-fire neuron using rank order coding a: How to implement spike time dependent hebbian plasticity. *Neurocomputing*, *38*.

Rao, R., & Sejnowski, T. J. (1991). *Predictive sequence learning in recurrent neocortical circuits* (Vol. 12; S. Solla, T. Leen, & K. Muller, Eds.). Cambridge MA, MIT Press.

Rao, R., & Sejnowski, T. J. (2001). Spike-timing-dependent hebbian plasticity as temporal difference learning. *Neural Comput.*, *13*(10), 2221–2237.

Rieke, F., Warland, D., Steveninck, R. de Ruyter van, & Bialek, W. (1996). *Spikes, exploring the neural code*. The M.I.T. Press.

Rostro-Gonzalez, H., Cessac, B., Vasquez, J., & Viéville, T. (2009). Back-engineering of spiking neural networks parameters. *Journal of Computational Neuroscience*. (submitted)

Rudolph, M., & Destexhe, A. (2006). Analytical integrate and fire neuron models with conductance-based dynamics for event driven simulation strategies. *Neural Computation*, *18*, 2146–2210.

Ruelle, D. (1969). *Statistical mechanics: Rigorous results*. Benjamin, New York.

Ruelle, D. (1999). Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics. *J. Statist. Phys.*, *95*, 393-468.

Samuelides, M., & Cessac, B. (2007). Random recurrent neural networks. *European Physical Journal - Special Topics*, *142*, 7–88.

Schneidman, E., Berry, M., Segev, R., & Bialek, W. (2006). Weak pairwise correlations imply string correlated network states in a neural population. *Nature*, *440*, 1007– 1012.

Sinanović, A., & Johnson, D. (2006). Toward a theory of information processing. *signal processing*. (submitted)

Siri, B., Berry, H., Cessac, B., Delord, B., & Quoy, M. (2007). Effects of hebbian learning on the dynamics and structure of random networks with inhibitory and excitatory neurons. *Journal of Physiology, Paris*, *101*(1-3), 138–150. (e-print: arXiv:0706.2602)

Siri, B., Berry, H., Cessac, B., Delord, B., & Quoy, M. (2008). A mathematical analysis of the effects of hebbian learning rules on the dynamics and structure of discrete-time random recurrent neural networks. *Neural Comp.*, *20*(12), 12. (e-print: arXiv:0705.3690v1)

Soula, H. (2005). *Dynamique et plasticité dans les réseaux de neurones à impulsions*. Unpublished doctoral dissertation, INSA Lyon.

Soula, H., Beslon, G., & Mazet, O. (2006). Spontaneous dynamics of asymmetric random recurrent spiking neural networks. *Neural Computation*, *18*(1).

Soula, H., & Chow, C. C. (2007). Stochastic dynamics of a finite-size spiking neural networks. *Neural Computation*, *19*, 3262–3292.

Theunissen, F., & Miller, J. (1995). Temporal encoding in nervous systems: A rigorous definition. *Journal of Computational Neuroscience*, *2*, 149162.

Tkacik, G., Schneidman, E., Berry, M., & Bialek, W. (2006). Ising models for networks of real neurons. *arXiv*, *q-bio/0611072*.

Touboul, J. (2008). Bifurcation analysis of a general class of nonlinear integrate-and-fire neurons. *SIAM Journal on Applied Mathematics*, *68*(4), 1045-1079.

Toyoizumi, T., Pfister, J.-P., Aihara, K., & Gerstner, W. (2005). Generalized bienenstock-cooper-munro rule for spiking neurons that maximizes information transmission. *Proceedings of the National Academy of Science*, *102*, 5239–5244.

Toyoizumi, T., Pfister, J.-P., Aihara, K., & Gerstner, W. (2007). Optimality model of unsupervised spike-timing dependent plasticity: Synaptic memory and weight distribution. *Neural Computation*, *19*, 639–671.

Wood, F., Roth, S., & Black, M. (2006). Modeling neural population spiking activity with gibbs distributions. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 1537–1544). Cambridge, MA: MIT Press.

Zou, Q. (2006). *Modèles computationnels de la plasticité impulsionnelle: synapses, neurones et circuits.* Unpublished doctoral dissertation, Université Paris VI.

Bibliography

[1] Abeles, M. and G. L. Gerstein: 1988, 'Detecting Spatiotemporal Firing Patterns Among Simultaneously Recorded Single Neurons'. *Journal of Neurophysiology* **60**(3), 909–924. [6]

[2] Adrian, E. and Y. Zotterman: 1926, 'The impulses produced by sensory nerve endings: Part II: The response of a single end organ'. *J Physiol (Lond.)* **61**, 151–71. [5]

[3] Amari, S.-I.: 2010, 'Information Geometry of Multiple Spike Trains'. In: S. Grün and S. Rotter (eds.): *Analysis of Parallel Spike trains*, Vol. 7 of *Springer Series in Computational Neuroscience*. Springer, part 11, pp. 221–253. DOI: 10.1007/978-1-4419-5675. [8, 10, 46]

[4] Amari, S.-I. and H. Nagaoka: 2000, *Methods of Information Geometry*. Oxford Univ. Press. [147]

[5] Arabzadeh, E., S. Panzeri, and M. Diamond: 2006, 'Deciphering the Spike Train of a Sensory Neuron: Counts and Temporal Patterns in the Rat Whisker Pathway'. *The Journal of Neuroscience* **26**(36), 9216–9226. [5]

[6] Artola, A., S. Bröcher, and W. Singer: 1990, 'Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex.'. *Nature* **347**(6288), 69–72. [125]

[7] Ay, N. and T. Wennekers: 2003, 'Dynamical properties of strongly interacting Markov chains'. *Neural Networks* **16**(10), 1483 – 1497. [33, 150]

[8] B., C., V. J.C, and N. H.: 2010, 'Spike trains statistics in Integrate and Fire Models: exact results'. In: *proceedings of the Neurocomp Conference*. Lyon, France. [121]

[9] Bai, Z., J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst (eds.): 2000, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM. [59, 63, 151]

[10] Barbieri, R., L. M. Frank, D. P. Nguyen, M. C. Quirk, M. A. Wilson, and E. N. Brown: 2004, 'Dynamic Analyses of Information Encoding in Neural Ensembles'. *Neural Computation* **16**, 277–307. [5]

[11] BenArous, G. and A. Guionnet: 1995, 'Large deviations for Langevin spin glass dynamics'. *Probability Theory and Related Fields* **102**, 455–509. [147, 148]

[12] Bi, G. and M. Poo: 2001, 'Synaptic Modification by Correlated Activity: Hebb's Postulate Revisited'. *Annual Review of Neuroscience* **24**, 139–166. [47, 123, 125, 126]

[13] Bienenstock, E. L., L. Cooper, and P. Munroe: 1982, 'Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex'. *The Journal of Neuroscience* **2**(1), 32âĂŞ–48. [123]

[14] Bliss, T. and A. Gardner-Medwin: 1973, 'Long-lasting potentiation of synaptic transmission in the Dentate Area of the unanaesthetised rabbit following stimulation of the perforant path.'. *J Physiol* **232**, 357–374. [125]

[15] Bohte, S. M. and M. C. Mozer: 2007, 'Reducing the Variability of Neural Responses: A Computational Theory of Spike-Timing-Dependent Plasticity.'. *Neural Computation* **19**(2), 371–403. [123]

[16] Bollt, E. M., T. Stanford, Y.-C. Lai, and K. Życzkowski: 2000, 'Validity of Threshold-Crossing Analysis of Symbolic Dynamics from Chaotic Time Series'. *Phys. Rev. Lett.* **85**(16), 3524–3527. [80]

[17] Bowen, R.: 2008, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms. Second revised version.*, Vol. 470 of *Lect. Notes.in Math.* Springer-Verlag. [28, 29, 117]

[18] Bressloff, P. C. and S. Coombes: 2003, *Synchronization of synaptically-coupled neural oscillators in: Epilepsy as a dynamic disease*, Chapt. 7. J.Milton and P. Jung, Springer-Verlag. [46]

[19] Brette, R. and W. Gerstner: 2005, 'Adaptive exponential integrate-and-fire model as an effective description of neuronal activity'. *Journal of Neurophysiology* **94**, 3637–3642. [4]

[20] Brillinger, D. R.: 1988, 'Maximum likelihood analysis of spike trains of interacting nerve cells'. *Biol Cybern* **59**(3), 189–200. [16]

[21] Brown, E. N., R. E. Kass, and P. P. Mitra: 2004, 'Multiple neural spike train data analysis: state-of-the-art and future challenges'. *Nature Neuroscience* **7**(5), 456–461. [6]

[22] Buhl, M. and M. B. Kennel: 2007, 'Globally enumerating unstable periodic orbits for observed data using symbolic dynamics'. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **17**(3), 033102. [80]

[23] Cessac, B.: 2007, 'Does the complex susceptibility of the Hénon map have a pole in the upper-half plane ? A numerical investigation.'. *Nonlinearity* **20**, 2883–2895. [135]

[24] Cessac, B.: 2008, 'A discrete time neural network model with spiking neurons. Rigorous results on the spontaneous dynamics'. *J. Math. Biol.* **56**(3), 311–345. [119, 120, 131, 134, 143]

[25] Cessac, B.: 2010, 'A discrete time neural network model with spiking neurons II. Dynamics with noise'. *J.Math. Biol., accepted*. [12, 17, 19, 23, 99, 120, 121, 143, 144, 145, 147, 149]

[26] Cessac, B., H.Rostro-Gonzalez, J. Vasquez, and T. Viéville: 2009, 'How Gibbs distribution may naturally arise from synaptic adaptation mechanisms: a model based argumentation'. *J. Stat. Phys, In press*. [9, 55, 113, 143, 145]

[27] Cessac, B. and M. Samuelides: 2007, 'From neuron to neural networks dynamics.'. *EPJ Special topics: Topics in Dynamical Neural Networks* **142**(1), 7–88. [124]

[28] Cessac, B. and T. Viéville: 2008, 'On Dynamics of Integrate-and-Fire Neural Networks with Adaptive Conductances'. *Frontiers in neuroscience* **2**(2). [16, 113, 119, 120, 134]

[29] Chazottes, J.: 1999, 'Entropie Relative, Dynamique Symbolique et Turbulence'. Ph.D. thesis, Université de Provence - Aix Marseille I. [37, 57, 112, 134, 135]

[30] Chazottes, J., E. Floriani, and R. Lima: 1998, 'Relative entropy and identification of Gibbs measures in dynamical systems'. *J. Statist. Phys.* **90**(3-4), 697–725. [35, 57, 112]

[31] Chazottes, J. and G. Keller: 2009, *Pressure and Equilibrium States in Ergodic Theory*, Chapt. Ergodic Theory. Encyclopedia of Complexity and System Science, Springer. to appear. [29]

[32] Chazottes, J.-R. and E. Ugalde: 2003, 'Projection of Markov Measures May Be Gibbsian'. *Journal of Statistical Physics* **111**, 1245–1272. [81]

[33] Chazottes, J.-R. and E. Ugalde: 2011, 'On the preservation of gibbsianness under amalgamation of symbols'. *Entropy of Hidden Markov Processes and Connections to Dynamical Systems*. [81]

[34] Chechik, G.: 2003, 'Spike-Timing-Dependent Plasticity and Relevant Mutual Information Maximization'. *Neural Computation* **15**(7), 1481–1510. [123]

[35] Chiu, S. N. and K. I. Liu: 2009, 'Generalized Cramér-von Mises goodness-of-fit tests for multivariate distributions'. *Computational Statistics and Data Analysis* **53**(11), 3817–3834. [41]

[36] Chornoboy, E. S., L. P. Schramm, and A. F. Karr: 1988, 'Maximum likelihood identification of neural point process systems'. *Biol Cybern* **59**(4-5), 265–275. [16]

[37] Cirrone, G., S. Donadio, S. Guatelli, A. Mantero, B. Mascialino, S. Parlati, A. Pfeiffer, M. Pia, A. Ribon, and P. Viarengo: 2004, 'A Goodness-of-Fit Statistical Toolkit'. *IEEE Transactions on Nuclear Science* **51**(5), 2056–2063. [64]

[38] Cocco, S., S. Leibler, and R. Monasson: 2009, 'Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods'. *PNAS* **106**(33), 14058–14062. [7, 26]

[39] Cochran, W. G.: 1934, 'The distribution of quadratic forms in a normal system, with applications to the analysis of covariance'. *Mathematical Proceedings of the Cambridge Philosophical Society* **30**(2), 178–191. [39]

[40] Cooper, L., N. Intrator, B. Blais, and H. Shouval: 2004, *Theory of cortical plasticity*. World Scientific, Singapore. [123]

[41] Cressie, N. and T. R. C. Read: 1984, 'Multinomial Goodness-of-Fit Tests'. *Journal of the Royal Statistical Society. Series B (Methodological)* **46**(3), 440–464. [40]

[42] Cronin, J.: 1987, *Mathematical aspects of Hodgkin-Huxley theory*. Cambridge University Press. [4, 114]

[43] Csiszar, I.: 1984, 'Sanov Property, Generalized I-Projection and a Conditional Limit Theorem'. *Ann. Prob* **12**(3), 768–793. [9, 32, 34, 53]

[44] Cunningham, J. P., B. M. Yu, K. V. Shenoy, and M. Sahani: 2008, 'Inferring neural firing rates from spike trains using Gaussian processes'. In: J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.): *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press, pp. 329–336. [82]

[45] Dalevi, D., D. Dubhashi, and M. Hermansson: 2006, 'A New Order Estimator for Fixed and Variable Length Markov Models with Applications to DNA Sequence Similarity'. *Statistical Applications in Genetics and Molecular Biology* **5**(1). [35]

[46] Daucé, E., M. Quoy, B. Cessac, B. Doyon, and M. Samuelides: 1998, 'Self-Organization and dynamics reduction in recurrent networks: stimulus presentation and learning'. *Neural Networks* **11**, 521–33. [124]

[47] Daw, C. S., C. E. A. Finney, and E. R. Tracy: 2003, 'A review of symbolic analysis of experimental data'. *Review of Scientific Instruments* **74**(2), 915–930. [80]

[48] Dayan, P. and L. Abbott: 2001, *Theoretical Neuroscience : Computational and Mathematical Modeling of Neural Systems*. MIT Press. [2, 4, 5, 123]

[49] Dayan, P. and M. Hausser: 2004, *Plasticity kernels and temporal statistics*, Vol. 16 of *Advances in Neural information processing systems*. Cambridge MA: MIT Press. [123]

[50] Delorme, A., L. Perrinet, and S. Thorpe: 2001, 'Networks of integrate-and-fire neurons using Rank Order Coding B: Spike timing dependent plasticity and emergence of orientation selectivity'. *Neurocomputing* **38-40**, 539–45. [5]

[51] Dembo, A. and O. Zeitouni: 1993, *Large Deviation Techniques and Applications*. Springer. [36, 37]

[52] Demmel, J. W.: 1997, *Applied Numerical Linear Algebra*. SIAM, 1 edition. [25, 50, 59, 151]

[53] Dhadialla, P. S., I. E. Ohiohenuan, A. Cohen, and S. Strickland: 2009, 'Maximum-entropy network analysis reveals a role for tumor necrosis factor nerve development and function'. *Proceedings of the National Academy of Sciences of the United States* **106**(30), 12494–12499. [10]

[54] Diesmann, M., M.-O. Gewaltig, and A. Aertsen: 1999, 'Stable propagation of synchronous spiking in cortical neural networks.'. *Nature* **402**, 529–533. [46]

[55] Dudek, S. and M. F. Bear: 1993, 'Bidirectional long-term modification of synaptic effectiveness in the adult and immature hippocampus.'. *J Neurosci.* **13**(7), 2910–2918. [125]

[56] Ecker, A. S., P. Berens, G. A. Keliris, M. Bethge, N. K. Logothetis, and A. S. Tolias: 2010, 'Decorrelated Neuronal Firing in Cortical Microcircuits'. *Science* **327**(5965), 584–587. [5, 79]

[57] Endres, D. and M. W. Oram: 2010, 'Feature extraction from spike trains with Bayesian binning: 'Latency is where the signal starts"'. *Journal of Computational Neuroscience* **29**(1-2), 149–169. [82]

[58] Erb, I. and N. Ay: 2004, 'Multi-Information in the Thermodynamic Limit'. *Journal of Statistical Physics* **115**, 949–976. 10.1023/B:JOSS.0000022375.49904.ea. [33]

[59] Falconer, K.: 1997, *Techniques in Fractal Geometry*. Wiley. [41]

[60] Faugeras, O., J. Touboul, and B. Cessac: 2008, 'A constructive mean field analysis of multi population neural networks with random synaptic weights and stochastic inputs'. *Frontiers in Neuroscience*. submitted. [147, 148]

[61] Ferrari, F. and A. Wyner: 2003, 'Estimation of General Stationary Processes by Variable Length Markov Chains'. *Scandinavian Journal of Statistics* **30**, 459–480. [35, 145, 151]

[62] FitzHugh, R.: 1955, 'Mathematical models of threshold phenomena in the nerve membrane'. *Bull. Math. Biophysics* **17**, 257–278. [113]

[63] FitzHugh, R.: 1961, 'Impulses and physiological states in models of nerve membrane'. *Biophys. J.* **1**, 445–466. [113]

[64] Franke, F., M. Natora, C. Boucsein, M. H. J. Munk, and K. Obermayer: 2010, 'An online spike detection and spike classification algorithm capable of instantaneous resolution of overlapping spikes'. *Journal of Computational Neuroscience* **29**(1-2), 127–148. [79]

[65] Frigge, M., D. C. Hoaglin, and B. Iglewicz: 1989, 'Implementation of the Boxplot'. *The American Statistician* **43**(1), 50–54. [39, 74]

[66] Gantmacher, F. R.: 1998, *the theory of matrices*. Providence, RI. [19, 118, 142]

[67] Gao, Y., I. Kontoyiannis, and E. Bienenstock: 2008, 'Estimating the Entropy of Binary Time Series: Methodology, Some Theory and a Simulation Study'. *Entropy* **10**(2), 71–99. [5, 48, 55]

[68] Georgeopoulos, A. P., H. Merchant, T. Naselaris, and B. Amirikian: 2007, 'Mapping of the preferred direction in the motor cortex'. *PNAS* **104**(26), 11068–11072. [5]

[69] Georgopoulos, A., J. Kalaska, R. Caminiti, and J. Massey: 1982, 'On the relations between the direction of two-dimensional arm movements and cell discharge in primary motor cortex'. *J Neurosci* **2**(1527-1537). [5]

[70] Gerstner, W. and W. Kistler: 2002a, *Spiking Neuron Models*. Cambridge University Press. [4, 5, 119]

[71] Gerstner, W. and W. M. Kistler: 2002b, 'Mathematical formulations of Hebbian learning.'. *Biological Cybernetics* **87**, 404–415. [123, 125, 126, 127]

[72] Gerstner, W. and W. M. Kistler: 2002c, *Spiking Neuron Models*. Cambridge Press. [124]

[73] Globerson, A., E. Stark, E. Vaadia, and N. Tishby: 2009, 'The minimum information principle and its application to neural code analysis'. *Proceedings of the National Academy of Sciences of the United States* **106**(9), 3490–3495. [13]

[74] Grammont, F. and A. Riehle: 1999, 'Precise spike synchronization in monkey motor cortex involved in preparation for movement'. *Exp. Brain Res.* **128**, 118–122. [5, 45, 46]

[75] Grammont, F. and A. Riehle: 2003, 'Spike synchronization and firing rate in a population of motor cortical neurons in relation to movement direction and reaction time'. *Biological cybernetics* **88**(5), 260–373. [5]

[76] Grassberger, P.: 1989, 'Estimating the Information Content of Symbol Sequences and Efficient Codes'. *IEEE Transactions on Information Theory* **35**. [48, 53, 55]

[77] Grosse, I., P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H. E. Stanley: 2002, 'Analysis of symbolic sequences using the Jensen-Shannon divergence'. *Phys. Rev. E* **65**(4), 041905. [82]

[78] Grün, S., M. Diesmann, and A. Aertsen: 2002, 'Unitary Events in Multiple Single-Neuron Spiking Activity: I. Detection and Significance'. *Neural Computation* **14**(1), 43–80. [6]

[79] Guckenheimer, J. and I. S. Labouriau: 1993, 'Bifurcation of the Hodgkin-Huxley equations: A new twist'. *Bull. Math. Biol.* **55**, 937–952. [4]

[80] Hebb, D.: 1949, *The organization of behavior: a neuropsychological theory.* Wiley, NY. [123]

[81] Hertz, J.: 1997, *Theoretical Aspects of Neural Computation.*, Chapt. Modelling synfire processing., pp. 135–144. Wong K-Y M. King I. and Yeung D-Y (eds), Springer-Verlag. [149]

[82] Herzel, H. and I. GroÃ§e: 1995, 'Measuring correlations in symbol sequences'. *Physica A: Statistical and Theoretical Physics* **216**(4), 518 –542. [82]

[83] Hirsch, M.: 1989, 'Convergent activation dynamics in continuous time networks'. *Neur. Networks* **2**, 331–349. [132]

[84] Hodgkin, A. and A. Huxley: 1952, 'A quantitative description of membrane current and its application to conduction and excitation in nerve.'. *Journal of Physiology* **117**, 500–544. [2, 4, 112, 113]

[85] Ince, R. A., R. Senatore, E. Arabzadeh, F. Montani, M. E. Diamond, and S. Panzeri: 2010a, 'Information-theoretic methods for studying population codes'. *Neural Networks* **23**(6), 713 – 727. Analysis and Modeling of Massively Parallel Neural Signals. [61, 150]

[86] Ince, R. A. A., M. A., R. S. Petersen, and S. Panzeri: 2010b, 'Open source tools for the information theoretic analysis of neural data'. *Frontier in Neuroinformatics* **1**(4), 62–70. [61, 63]

[87] Ince, R. A. A., F. Montani, E. Arabzadeh, M. E. Diamond, and S. Panzeri: 2009a, 'On the presence of high-order interactions among somatosensory neurons and their effect on information transmission'. *Journal of Physics: Conference Series* **197**(1), 012013. [61]

[88] Ince, R. A. A., R. S. Petersen, D. C. Swan, and S. Panzeri: 2009b, 'Python for information theoretic analysis of neural data'. *Frontier in Neuroinformatics* **3**(4), 1–15. [60, 61, 152]

[89] Izhikevich, E.: 2003, 'Simple Model of Spiking Neurons'. *IEEE Transactions on Neural Networks* **14**(6), 1569–1572. [120]

[90] Izhikevich, E.: 2004, 'Which model to use for cortical spiking neurons?'. *IEEE Trans Neural Netw* **15**(5), 1063–1070. [4]

[91] Izhikevich, E. and N. Desai: 2003, 'Relating STDP to BCM'. *Neural Computation* **15**, 1511–1523. [126, 127, 135]

[92] Jaynes, E.: 1957, 'Information theory and statistical mechanics'. *Phys. Rev.* **106**, 620. [8, 9, 28, 31]

[93] Johnson, D.: 2004, 'Neural Population Structure and Consequences for Neural Coding'. *Journal of Computational Neuroscience* **16**(1), 69–80. [5]

[94] Johnson, D. H. and A. Swami: 1983, 'The transmission of signals by auditory-nerve fiber discharge patterns'. *J. Acoust. Soc. Am* **74**(2), 493–501. [16]

[95] Johnson, K. O.: 1980, 'Sensory discrimination: neural processes preceding discrimination decision.'. *J Neurophysiol* **43**(6), 1793–1815. [5]

[96] Jolivet, R., A. Rauch, H.-R. Lescher, and W. Gerstner: 2006, *"Integrate-and-Fire models with adaptation are good enough"*. MIT Press, Cambridge. [119]

[97] Kahle, T., E. Olbrich, J. Jost, and N. Ay: 2009, 'Complexity measures from interaction structures'. *Phys. Rev. E* **79**(2), 026201. [32]

[98] Kass, R. E. and V. Ventura.: 2001, 'A spike-train probability model.'. *Neural Comput.* **13**(8), 1713–1720. [16]

[99] Kass, R. E., V. Ventura, and E. N. Brown: 2005, 'Statistical Issues in the Analysis of Neuronal Data'. *Journal of Neurophysiology* **94**(1), 8–25. [6]

[100] Katok, A. and B. Hasselblatt: 1998, *Introduction to the modern theory of dynamical systems*. Kluwer. [115]

[101] Kayser, C., N. K. Logothetis, and S. Panzeri: 2010, 'Millisecond encoding precision of auditory cortex neurons'. *Proceedings of the National Academy of Sciences* **107**(39), 16976–16981. [82]

[102] Keller, G.: 1998, *Equilibrium States in Ergodic Theory*. Cambridge University Press. [20, 25, 28, 29, 31, 37, 51, 112, 141]

[103] Kelly, R. C., M. A. Smith, J. M. Samonds, A. Kohn, A. B. Bonds, J. A. Movshon, and T. Sing Lee: 2007, 'Comparison of Recordings from Microelectrode Arrays and Single Electrodes in the Visual Cortex'. *J. Neurosci.* **27**(2), 261–264. [79]

[104] Kempton, T.: 2011, 'Factors of Gibbs Measures for Subshifts of Finite Type'. *submitted to the Bulletin of the London Mathematical Society.* [81]

[105] Kempton, T. and M. Pollicot: 2011, 'Factors of Gibbs Measures for Full Shifts'. *Entropy of Hidden Markov Processes and Connections to Dynamical Systems*. [81]

[106] Kennel, M. B. and M. Buhl: 2003, 'Estimating Good Discrete Partitions from Observed Data: Symbolic False Nearest Neighbors'. *Phys. Rev. Lett.* **91**(8), 084102. [80]

[107] Kolomiets, B., E. Dubus, M. Simonutti, S. Rosolen, J. A. Sahel, and S. Picaud: 2010, 'Late histological and functional changes in the P23H rat retina after photoreceptor loss'. *Neurobiology of Disease* **38**(1), 47 – 58. [100]

[108] Laskey, K. and L. Martignon: 1996, 'Bayesian Learning of Loglinear Models for Neural Connectivity'. In: *Proceedings of the Proceedings of the Twelfth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-96)*. San Francisco, CA, pp. 373–380, Morgan Kaufmann. [7]

[109] Latham, P., A. Roth, M. Hausser, and M. London: 2006, 'Requiem for the spike?'. *Soc. Neurosc. Abstr.* **32**. [149]

[110] Lemeshko, B. and S. Lemeshko: 2005, 'Statistical distribution convergence and homogeneity test power for Smirnov and LehmannâĂŤRosenblatt tests'. *Measurement Techniques* **48**, 1159–1166. 10.1007/s11018-006-0038-3. [41]

[111] Lesne, A., J.-L. Blanc, and L. Pezard: 2009, 'Entropy estimation of very short symbolic sequences'. *Phys. Rev. E* **79**(4), 046208. [81]

[112] Levy, W. and D. Stewart: 1983, 'Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus.'. *Neuroscience* **8**(4), 791–âĂŞ797. [125]

[113] Li, W.: 1990, 'Mutual information functions versus correlation functions'. *Journal of Statistical Physics* **60**, 823–837. 10.1007/BF01025996. [81]

[114] Lindsey, B., K. Morris, R. Shannon, and G. Gerstein: 1997, 'Repeated patterns of distributed synchrony in neuronal assemblies'. *Journal of Neurophysiology* **78**, 1714–1719. [8]

[115] London, M., A. Shcreibman, and I. Segev: 2002, 'Estimating Information Theoretic Quantities of spike-trains using the Context Tree Weighting algorithm'. *Nature neuroscience* **5**. Appendix to: The information efficacy of a synapse. [55]

[116] Maillard, G.: 2007, 'Introduction to chains with complete connections'. Ecole Federale Polytechnique de Lausanne. [144]

[117] Malenka, R. C. and R. A. Nicoll: 1999, 'Long-Term Potentiation - A Decade of Progress ?'. *Science* **285**(5435), 1870 – 1874. [125]

[118] Malouf, R.: 2002, 'A comparison of algorithms for maximum entropy parameter estimation'. In: *proceedings of the 6th conference on Natural language learning - Volume 20*. Stroudsburg, PA, USA, pp. 1–7, Association for Computational Linguistics. [62]

[119] Markram, H., J. Lübke, M. Frotscher, and B. Sakmann: 1997, 'Regulation of synaptic efficacy by coincidence of postsynaptic AP and EPSP'. *Science* **275**(213). [123, 125]

[120] Marre, O., S. E. Boustani, Y. Frégnac, and A. Destexhe: 2009, 'Prediction of Spatiotemporal Patterns of Neural Activity from Pairwise Correlations'. *Physical Review Letters* **102**(13), 4. [8, 10, 22, 27, 37, 39, 46, 122, 147]

[121] Martignon, L., G. Deco, K. Laskey, M. Diamond, W. Freiwald, and E. Vaadia: 2000, 'Neural Coding: Higher-Order Temporal Patterns in the Neurostatistics of Cell Assemblies'. *Neural Computation* **12**(11), 2621–2653. [7, 8, 46, 83]

[122] Martignon, L., H. von Hasseln, S. Grün, A. Aertsen, and G. Palm: 1995, 'Detecting Higher-Order interactions among the spiking events in a group of neurons'. *Biological Cybernetics* **73**(1), 69–81. [7, 8]

[123] Mascialino, B., A. Pfeiffer, M. Pia, A. Ribon, and P. Viarengo: 2006, 'New Developments of the Goodness-of-Fit Statistical Toolkit'. *IEEE Transactions on Nuclear Science* **53**(6), 3834–3841. [64]

[124] Meyer, D.: 1980, *The Ruelle-Araki transfer operator in classical statistical mechanics*, Vol. 123. Springer-Verlag. [116]

[125] Mezard, M. and T. Mora: 2009, 'Constraint satisfaction problems and neural networks: A statistical physics perspective'. *Journal of Physiology-Paris* **103**(1–2), 107–113. [146]

[126] Mézard, M., G. Parisi, and M. Virasoro: 1987, *Spin-glass theory and beyond*. World scientific Singapore. [147]

[127] Miller, K., J. Keller, and M. Stryker: 1989, 'Ocular dominance column development: analysis and simulation'. *Science* **245**(4918), 605–615. [123]

[128] Morvai, G. and B. Weiss: 2008, 'Estimating the Lengths of Memory Words'. *IEEE Transactions on Information Theory* **54**(8), 3804–3807. [35]

[129] Nagumo, J., S. Arimoto, and S. Yoshizawa: 1962, 'An active pulse transmission line simulating nerve axon'. *Proc.IRE* **50**, 2061–2070. [113]

[130] Negaev, A.: 2002, 'An asymptotic formula for the Neyman-Pearson risk in discriminating between two markov chains'. *Journal of Mathematical Sciences* **111**(3), 3582–3591. [37, 41]

[131] Nemenman, I., W. Bialek, and R. de Ruyter van Steveninck: 2004, 'Entropy and information in neural spike trains: Progress on the sampling problem'. *Phys. Rev. E* **69**(5), 056111. [82]

[132] Nemenman, I., G. Lewen, W. Bialek, and R. de Ruyter van Steveninck: 2006, 'Neural coding of a natural stimulus ensemble: Information at sub-millisecond resolution.'. *PLoS Comp Bio* **4**, e1000025. [5]

[133] Nirenberg, S. and P. Latham: 2003, 'Decoding neuronal spike trains: how important are correlations'. *Proceeding of the Natural Academy of Science* **100**(12), 7348–7353. [5]

[134] Ohiorhenuan, I. E., F. Mechler, K. P. Purpura, A. M. Schmid, Q. Hu, and J. D. Victor: 2010, 'Sparse coding and high-order correlations in fine-scale cortical networks'. *Nature* **466**(7), 617–621. [8, 141]

[135] Okatan, M., M. A.Wilson, and E. N. Brown: 2005, 'Analyzing Functional Connectivity Using a Network Likelihood Model of Ensemble Neural Spiking Activity.'. *Neural Computation* **17**(9), 1927–1961. [16]

[136] Opper, M. and D. Saad: 2001, *Advanced Mean Field Methods: Theory and Practice*. MIT. Press. [147]

[137] Osbone, L., S. Palmer, S. Lisberger, and W. Bialek: 2008, 'Combinatorial coding in neural populations'. *arXiv.org:0803.3837*. [5]

[138] Osborne, L. C., S. E. Palmer, S. G. Lisberger, and W. Bialek: 2008, 'The Neural Basis for Combinatorial Coding in a Cortical Population Response'. *The Journal of Neuroscience* **28**(50), 13522–13531. [7]

[139] Parry, W. and M. Pollicott: 1990, *Zeta functions and the periodic orbit structure of hyperbolic dynamics*, Vol. 187–188. Asterisque. [112]

[140] Paugam-Moisy, H., R. Martinez, and S. Bengio: 2008, 'Delay learning and polychronization for reservoir computing'. *Neurocomputing* **71**, 1143–1158. [149]

[141] Perrinet, L., A. Delorme, M. Samuelides, and S. Thorpe: 2001, 'Networks of integrate-and-fire neuron using rank order coding A: How to implement spike time dependent Hebbian plasticity.'. *Neurocomputing* **38**. [5]

[142] Piccardi, C.: 2006, 'On parameter estimation of chaotic systems via symbolic time-series analysis'. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **16**(4), 043115. [38]

[143] Plefka, T.: 1982, 'Convergence condition of the TAP equations for the infinite-ranged Ising spin glass model'. *J. Phys. A* **15**, 1971. [147]

[144] Politis, D. N.: 2003, 'The Impact of Bootstrap Methods on Time Series Analysis'. *Statistical Science* **18**(2), 219–230. [39]

[145] Pouget, A. and G. C. DeAngelis: 2008, 'Paying attention to correlated neural activity'. *Nature Neuroscience* **11**(12), 1371–1372. [5]

[146] Pouzat, C. and A. Chaffiol: 2009a, 'Automatic Spike Train Analysis and Report Generation. An Implementation with R, R2HTML and STAR'. *J Neurosci Methods* **181**, 119–144. [6]

[147] Pouzat, C. and A. Chaffiol: 2009b, 'On Goodness of Fit Tests For Models of Neuronal Spike Trains Considered as Counting Processes'. [6]

[148] Pouzat, C. and A. Chaffiol: 2009c, 'On Goodness of Fit Tests For Models of Neuronal Spike Trains Considered as Counting Processes'. *http://arxiv.org/abs/0909.2785v1*. [16]

[149] QueirÃşs, S. M. D.: 2009, 'On a comparative study between dependence scales determined by linear and non-linear measures'. *Physica D: Nonlinear Phenomena* **238**(7), 764 – 770. [81]

[150] Rao, R. and T. J. Sejnowski: 1991, *Predictive sequence learning in recurrent neocortical circuits*, Vol. 12 of *Advances in Neural Information and Processing Systems*. Cambridge MA, MIT Press. [123]

[151] Rao, R. and T. J. Sejnowski: 2001, 'Spike-timing-dependent Hebbian plasticity as temporal difference learning.'. *Neural Comput.* **13**(10), 2221–2237. [123]

[152] Renart, A., J. de la Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. D. Harris: 2010, 'The Asynchronous State in Cortical Circuits'. *Science* **327**(5965), 587–590. [5]

[153] Rieke, F., D. Warland, R. de Ruyter van Steveninck, and W. Bialek: 1996, *Spikes, Exploring the Neural Code*. The M.I.T. Press. [5]

[154] Rieke, F., D. Warland, R. de Ruyter van Steveninck, and W. Bialek: 1997, *Spikes: Exploring the Neural Code*. Bradford Books. [5]

[155] Rosenthal, J.: 2000, 'Parallel computing and Monte Carlo algorithms'. *Far East J. Theor. Stat.* **4**, 207–236. [146]

[156] Rostro-Gonzalez, H., B. Cessac, J. Vasquez, and T. Viéville: 2009, 'Back-engineering of spiking neural networks parameters'. *Journal of Computational Neuroscience*. submitted. [119]

[157] Roudi, Y., E. Aurell, and J. A. Hertz: 2009a, 'Statistical physics of pairwise probability models'. *Frontiers in Computational Neuroscience* p. 15. [7, 26, 32]

[158] Roudi, Y. and J. Hertz: 2010, 'Mean Field Theory For Non-Equilibrium Network Reconstruction'. *arXiv* p. 11. [8, 9, 46, 145]

[159] Roudi, Y., S. Nirenberg, and P. Latham: 2009b, 'Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't.'. *PLOS Computational Biology* **5**(5). [8, 141, 149]

[160] Roudi, Y., J. Tyrcha, and J. A. Hertz: 2009c, 'Ising model for Neural data: Model quality and approximate methods for extracting functional connectivity'. *Physical Review E* p. 051915. [7, 26, 145]

[161] Ruelle, D.: 1969, *Statistical Mechanics: Rigorous results*. Benjamin, New York. [29, 116, 130, 141]

[162] Ruelle, D.: 1999, 'Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics.'. *J. Statist. Phys.* **95**, 393–468. [130]

[163] Samuelides, M. and B. Cessac: 2007, 'Random Recurrent Neural Networks'. *European Physical Journal - Special Topics* **142**, 7–88. [123, 147, 148]

[164] Sarig, O.: 2010, *Thermodynamic Formalism for countable Markov shifts*. http://www.wisdom.weizmann.ac.il/ sarigo/TDFnotes.pdf. [117]

[165] Schaub, M. T. and S. R. Schultz: 2010, 'The Ising decoder: reading out the activity of large neural ensembles'. *arXiv:1009.1828*. [7, 141]

[166] Schneidman, E., M. Berry, R. Segev, and W. Bialek: 2006, 'Weak pairwise correlations imply string correlated network states in a neural population'. *Nature* **440**, 1007– 1012. [5, 6, 10, 26, 31, 32, 33, 45, 46, 83, 141, 146, 147]

[167] Schürmann, T. and P. Grassberger: 1996, 'Entropy estimation of symbol sequences'. *Chaos* **6**(3), 414–427. [55]

[168] Scott, D. W.: 1992, *Multivariate Density Estimation. Theory, Practice, and Visualization*. Wiley. [82]

[169] Segev, R., I. Baruchi, E. Hulata, and E. Ben-Jacob: 2004a, 'Hidden neuronal correlations in cultured networks'. *Physical Review Letters* **92**, 118102. [8]

[170] Segev, R., J. Goodhouse, J. Puchalla, and M. J. Berry II: 2004b, 'Recording spikes from a large fraction of the ganglion cells in a retinal patch'. *Nat Neurosci* **7**, 1155 – 1162. [87]

[171] Seneta, E.: 2006, *Non-negative Matrices and Markov Chains*. Springer. [19, 118, 142]

[172] Sessak, V. and R. Monasson: 2009, 'Small-correlation expansions for the inverse ising problem.'. *J. Phys. A* **42**, 055001. [147]

[173] Sherrington, D. and S. Kirkpatrick: 1975, 'Solvable Model of a Spin-Glass'. *Physical Review Letters* **35**(26), 1792+. [147]

[174] Shimazaki, H. and S. Shinomoto: 2007, 'A Method for Selecting the Bin Size of a Time Histogram'. *Neural Computation* **19**(6), 1503–1527. [82]

[175] Shlens, J., G. D. Field, J. L. Gauthier, M. Greschner, A. Sher, A. M. Litke, and E. J. Chichilnisky: 2009, 'The Structure of Large-Scale Synchronized Firing in Primate Retina'. *The Journal of Neuroscience* **29**(15), 5022–5031. [7, 8, 10, 141]

[176] Shlens, J., G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. J. Chichilnisky: 2006, 'The structure of Multi-Neuron Firing Patterns in Primate Retina'. *The Journal of Neuroscience* **26**(32), 8254–8266. [6, 10, 46, 83]

[177] S.I., A.: 2001, 'Information geometry on hierarchy of probability distributions'. *IEEE Trans. Information Theory* **47**(5), 1701–1711. [61]

[178] Sinanović, A. and D. Johnson: 2007, 'Toward a theory of information processing'. *signal processing* **87**(6). [5]

[179] Siri, B., H. Berry, B. Cessac, B. Delord, and M. Quoy: 2007, 'Effects of Hebbian learning on the dynamics and structure of random networks with inhibitory and excitatory neurons.'. *Journal of Physiology, Paris* **101**(1-3), 138–150. e-print: arXiv:0706.2602. [124]

[180] Siri, B., H. Berry, B. Cessac, B. Delord, and M. Quoy: 2008, 'A Mathematical Analysis of the Effects of Hebbian Learning Rules on the Dynamics and Structure of Discrete-Time Random Recurrent Neural Networks'. *Neural Comp.* **20**(12), 2937–2966. e-print: arXiv:0705.3690v1. [124]

[181] Sompolinsky, H., A. Crisanti, and H. Sommers: 1988, 'Chaos in random neural networks'. *Phys. Rev. Lett.* **61**, 259–262. [147]

[182] Sompolinsky, H. and A. Zippelius: 1982, 'Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses'. *Physical Review B* **25**(11), 6860–6875. [147]

[183] Sorensen, D. C.: 2002, 'Numerical methods for large eigenvalue problems'. *Acta Numerica* **11**, 519–584. [57]

[184] Soula, H.: 2005, 'Dynamique et plasticité dans les réseaux de neurones à impulsions'. Ph.D. thesis, INSA Lyon. [124]

[185] Soula, H., G. Beslon, and O. Mazet: 2006, 'Spontaneous dynamics of asymmetric random recurrent spiking neural networks'. *Neural Computation* **18**(1). [119, 124]

[186] Soula, H. and C. C. Chow: 2007, 'Stochastic Dynamics of a Finite-Size Spiking Neural Networks'. *Neural Computation* **19**, 3262–3292. [123]

[187] Steele, M., C. Hurst, and J. Chaseling: 2008, 'The power of chi-square type goodness-of-fit test statistics'. *Far East journal of theoretical statistics* **26**(1), 109–119. [41]

[188] Stevenson, I. H. and K. P. Kording: 2011, 'How advances in neural recording affect data analysis'. *Nature Neuroscience* **14**(2). [6, 7]

[189] Stevenson, I. H., J. M. Rebesco, L. E. Miller, and K. P. Körding: 2008, 'Inferring functional connections between neurons'. *Current Opinion in Neurobiology* **18**, 582–588. [13]

[190] Takekawa, T., Y. Isomura, and T. Fukai: 2010, 'Accurate spike sorting for multi-unit recordings.'. *Eur J Neurosci* **31**(2), 263–72. [79]

[191] Tang, A., D. Jackson, J. Hobbs, W. Chen, J. L. Smith, H. Patel, A. Prieto, D. Petrusca, M. I. Grivich, A. Sher, P. Hottowy, W. Dabrowski, A. M. Litke, and J. M. Beggs: 2008, 'A Maximum Entropy Model Applied to Spatial and Temporal Correlations from Cortical Networks *In Vitro*'. *The Journal of Neuroscience* **28**(2), 505–518. [7, 8, 39, 40, 83]

[192] Tang, X. Z., E. R. Tracy, A. D. Boozer, A. deBrauw, and R. Brown: 1995, 'Symbol sequence statistics in noisy chaotic signal reconstruction'. *Phys. Rev. E* **51**(5), 3871–3889. [38]

[193] Theunissen, F. and J. Miller: 1995, 'Temporal Encoding in Nervous Systems: A Rigorous Definition.'. *Journal of Computational Neuroscience* **2**, 149âĂŤ162. [5]

[194] Thouless, D. J., P. W. Anderson, and R. G. Palmer: 1977, 'Solution of âĂŸ solvable model of a spin glass'. *Philos. Mag.* **35**, 593–601. [147]

[195] Tkačik, G., J. S. Prentice, V. Balasubramanian, and E. Schneidman: 2010, 'Optimal population coding by noisy spiking neurons'. *PNAS* **107**(32), 14419–14424. [7]

[196] Tkačik, G., E. Schneidman, M. J. Berry II, and W. Bialek: 2009, 'Spin glass models for a network of real neurons'. *arXiv* p. 15. [5, 7, 8, 10]

[197] Touboul, J.: 2008, 'Bifurcation Analysis of a General Class of Nonlinear Integrate-and-Fire Neurons'. *SIAM Journal on Applied Mathematics* **68**(4), 1045–1079. [4]

[198] Touboul, J. and O. Faugeras: 2007, 'The spikes trains probability distributions: a stochastic calculus approach'. *Journal of Physiology, Paris* **101/1-3**, 78–98. [13]

[199] Touchette, H.: 2009, 'The large deviation approach to statistical mechanics'. *Physics Reports* **478**(1-3), 1 – 69. [37]

[200] Toyoizumi, T., J.-P. Pfister, K. Aihara, and W. Gerstner: 2005, 'Generalized Bienenstock-Cooper-Munro rule for spiking neurons that maximizes information transmission'. *Proceedings of the National Academy of Science* **102**, 5239–5244. [123]

[201] Toyoizumi, T., J.-P. Pfister, K. Aihara, and W. Gerstner: 2007, 'Optimality Model of Unsupervised Spike-Timing Dependent Plasticity: Synaptic Memory and Weight Distribution'. *Neural Computation* **19**, 639–671. [123]

[202] Truccolo, W. and J. P. Donoghue: 2007, 'Nonparametric modeling of neural point processes via stochastic gradient boosting regression.'. *Neural Computation* **19**(3), 672–705. [16]

[203] Truccolo, W., U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown: 2005, 'A Point Process Framework for Relating Neural Spiking Activity to Spiking History, Neural Ensemble and Extrinsic Covariate Effects.'. *J Neurophysiol* **93**, 1074–1089. [16]

[204] Truccolo, W., L. R. Hochberg, and J. P. Donoghue: 2010, 'Collective dynamics in human and monkey sensorimotor cortex: predicting single neuron spikes'. *Nature Neuroscience* **13**, 105–111. [13]

[205] Vasquez, J.-C., B. Cessac, H. Rostro-Gonzalez, and T. Viéville: 2009, 'How Gibbs Distributions may naturally arise from synaptic adaptation mechanisms'. In: B. Central (ed.): *BMC Neuroscience 2009*, Vol. 10 of *(Suppl 10)*. p. 213. [143]

[206] Villa, A. E. P., , I. V. Tetko, B. Hyland, and A. Najem: 1999, 'Spatiotemporal activity patterns of rat cortical neurons predict responses in a conditioned task'. *Proc Natl Acad Sci USA* **96**(3), 1106–1111. [8]

[207] von-der Malsburg, C.: 1973, 'Self-organisation of orientation sensitive cells in the striate cortex'. *Kybernetik* **14**, 85–100. [123]

[208] Wu, C.: 1986, 'Jackknife, bootstrap and other resampling methods in regression analysis.'. *Annals of Statistics* **14**, 1261–1295. [39]

[209] Xiao, Y., A. Gordon, and A. Yakovlev: 2006, 'A C++ Program for the CramÃľr-Von Mises Two-Sample Test'. *Journal of Statistical Software* **17**(8), 1–15. [64]

[210] Yan, J., M. Cowles, S. Wang, and M. Armstrong: 2007, 'Parallelizing MCMC for Bayesian spatiotemporal geostatistical models'. *Statistics and Computing* **17**, 323–335. [146]

[211] Yu, S., D. Huang, W. Singer, and D. NikoliÄĞ: 2008, 'A Small World of Neuronal Synchrony'. *Cerebral Cortex* **18**(12), 2891–2901. [7]

[212] Yuval Peres, P. S. a.: 2005, 'Two new Markov order estimators'. *arXiv:math/0506080v1 [math.ST]*. [35]

[213] Zhu, J. and E. P. Xing: 2009, 'Maximum Entropy Discrimination Markov Networks'. *Journal of Machine Learning Research* pp. 2531–2569. [10]

[214] Zou, Q.: 2006, 'Modèles computationnels de la plasticité impulsionnelle: synapses, neurones et circuits'. Ph.D. thesis, Université Paris VI. [134]