Parametric Estimation of Gibbs distributions as generalized maximum-entropy models for the analysis of spike train statistics.

Juan Carlos Vasquez, Thierry Viéville, Bruno Cessac

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Parametric Estimation of Gibbs distributions as generalized maximum-entropy models for the analysis of spike train statistics.

J. C. Vasquez , T. Viéville *, B. Cessac *

N° 7561

February 2011

Domaine 5

ISSN 0249-6399 ISRN INRIA/RR--7561--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SOPHIA ANTIPOLIS - MÉDITERRANÉE

Parametric Estimation of Gibbs distributions as generalized maximum-entropy models for the analysis of spike train statistics.

J. C. Vasquez *, T. Viéville *, B. Cessac * †

Domaine : STIC pour les sciences de la vie et de l'environnement
Équipes-Projets NeuroMathComp & CORTEX

Rapport de recherche n° 7561 — February 2011 — 51 pages

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 "free-energy" density and 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 allows 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 and the perspectives regarding spike-train code analysis are also discussed.

Key-words: Spike train analysis , Higher-order correlation , Statistical Physics , Gibbs Distributions , Maximum Entropy estimation

* 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.

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Estimation paramétrique des distributions de Gibbs comme modèles generalisés d'entropie maximale pour l'analyse de la statistique d'un train de spike

Résumé : Nous proposons une généralisation des modèles d'entropie maximale existantes utilisées pour l'analyse statistique de trains de spikes. Nous apportons ici une méthode simple pour estimer les distributions de Gibbs, généraliser les approches existantes basées sur le modèle d'Ising, ou estimer en une seule étape des chaînes de Markov pour un potentiel paramétrique arbitraire. Notre méthode permet de prendre en compte les effets de mémoire dans la dynamique. Il fournit directement la densité de "l'énergie libre" et la divergence de Kullback-Leibler entre les statistiques empiriques et le modèle statistique. Il ne se limite pas une forme de potentiel de Gibbs spécifique, mais permet de le chosiir, et ne nécessite pas l'hypothèse de "detailed balance". En outre, il permet la comparaison des différents modèles statistiques et offre un contrôle des effets de l'échantillonnage de taille finie, inhérente aux statistiques empiriques, en se basant sur des résultats de grandes déviation. Une validation numérique de la méthode est proposée et les perspectives en matière d'analyse de trains de spike sont également discutées.

Mots-clés : Analyse statistique de spike , Correlation d'ordre supérieur , Physique statistique , Distribution de Gibbs , Estimation d'entropie maximale

1 Introduction

Processing and encoding of information in neuronal dynamics is a very active research field [60], although still much of the role of neural assemblies and their internal interactions remains unknown [56]. 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 main difficulty with such type of approaches is that they rely on a largely controversial assumption, Poissonian statistics (see [58, 59, 70]), 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 [70]. They have shown that a model taking into account pairwise synchronizations between neurons in a small assembly (10-40 retinal ganglion cells) describes most (80-90%) 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 [77]. 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 synchronization 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 [44, 39, 45] (it was referred as "log-linear models"). The use of Ising model in neural data analysis (especially of visual stimuli) has been largely exploited by several other authors [19, 53, 76, 81]. 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 [84], 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 [83], they use this framework to study the optimal population coding and the stimuli representation. Additionally, in [65, 63], 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 [69], 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 [62], it has been shown that although Ising model is good for small populations, this is an artifact of the way data is binned and of 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 ([76, 84, 50]) 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, [45] 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, [44] 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 presents strong limitations, they have applied this methods successfully to experimental data from multi-units recordings in the prefrontal 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, 87, 72], and recently [81] have shown the insufficiency of Ising model to predict the temporal statistics of the neural activity. As a consequence, a few authors ([43, 32, 64]) have attempted to define time-dependent Gibbs distributions on the base of a Markovian approach (1-step time pairwise correlations). In particular, in [43] 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 [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 (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 paper, '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 [21]. 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.
- (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 paper 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. We implement this algorithm.

- (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 [64]. 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 [10] using the present formalism which extends to time-dependent parameters, as discussed in section (4.3.5). We provide the tool for this purpose.

In this paper, we propose therefore a generalization of the maximal entropy models used for spike trains statistics analysis. In short, what we bring is:

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 plugin to existing software such as Neuron or MVA-Spike.

The paper is organized as follows. The section 2 presents the theoretical framework of our approach. We propose a global approach to spike train analysis considering *spatio-temporal* and time-*causal* structure of spike trains emitted by neural networks. 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 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.

In section 3, we propose a numerical method based on the presented framework 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 [70, 77, 76] or 1-step time pairwise correlations models as [43, 32], 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.

In section 4 we perform a large battery of tests enabling us to experimentally validate the method. First, 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 [12]. 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 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 [11] 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 and partially overcome in section 4.3.5.

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 [11]. 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 $\Sigma \stackrel{\text{def}}{=} \{0,1\}^{\mathbb{Z}}$ 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 $Prob\left(\omega(t) \,|\, [\omega]_{t-R}^{t-1}\right)$, 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 [35, 7, 18, 36, 86, 51, 85, 57] 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 [13] it is clearly desirable, for practical purposes to work with finite memory R. In this way, $Prob\left(\omega(t) \,|\, [\omega]_{t-R}^{t-1}\right)$ 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:

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

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

$$\mathscr{A} \stackrel{\text{def}}{=} \{0, \ldots, M-1\}, \tag{2}$$

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:

$$\mathscr{H} \stackrel{\text{def}}{=} \mathbb{R}^M. \tag{3}$$

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 $\mathscr{M}$, with entries $\mathscr{M}_{w'w}$ such that:

$$\mathscr{M}_{w'w} = \begin{cases} Prob\left(\omega(t) \mid [\omega]_{t-R}^{t-1}\right), & \text{if} \quad w' \sim [\omega]_{t-R}^{t-1},\, w \sim [\omega]_{t-R+1}^{t}\,; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

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 $\mathscr{M}$ has M^2 entries, it is a *sparse matrix* since each line has, at most, 2^N non-zero entries.

In the same way, the probability of spike blocks of length R is represented by a M-dimensional vector $P \in \mathscr{H}$ 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 $\sum_{w' \in \mathscr{A}} P_{w'} = 1$.

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.

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

- *Normalization.*

$$\forall w' \in \mathscr{A}, \quad \sum_{w \in \mathscr{A}} M_{w'w} = 1, \tag{5}$$

- *Invariant probability distribution.*

  A probability distribution μ is invariant for the Markov chain if

$$\mu = \mu M. \tag{6}$$

  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.*

  For any integer $n > R$, for any sequence $[\omega]_1^n$,

$$Prob\left([\omega]_1^n\right) = \mu_{w_1} \prod_{k=1}^{n-1} M_{w_k w_{k+1}}. \tag{7}$$

  On the opposite, for blocks of size $0 < n < R$ then

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

  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 [73, 27] 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 $\mathscr{L}$ with $\mathscr{L}_{w'w} \geq 0$, but *we do not assume the normalization property* (5). We assume $\mathscr{L}$ to be primitive, i.e. $\exists n > 0$, such that, $\forall w, w'$, $\mathscr{L}^n_{w'w} > 0$ where $\mathscr{L}^n$ is the n-th power of $\mathscr{L}$. In the context of Markov chains (i.e. when $\mathscr{L}$ 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:

[2] The matrix $\mathscr{L}$ defined in (12) below is primitive by construction. But $\mathscr{L}$ 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 [13] 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.

Theorem 1 *$\mathscr{L}$ 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 $\mathscr{L}b\rangle = sb\rangle$, $\langle b\mathscr{L} = s\langle b$. Those vectors can be chosen such that $\langle b.b\rangle = 1$ 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. As a consequence, for any vector v in $\mathscr{H}$ not orthogonal to $\langle b$,*

$$\frac{1}{s^n}\mathscr{L}^n v \to b\rangle\langle b\,.\,v, \tag{8}$$

as $n \to +\infty$.

When this matrix is *normalized* (prop. (5)) then the following additional properties hold.

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

$$\mu_w = b_w\rangle\langle b_w. \tag{9}$$

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 $\mathscr{L}$ in theorem 1 is not normalized it is always possible to associate it with the (normalized) probability transition of a Markov chain $\mathscr{M}$ by the relation :

$$\mathscr{M}_{w'w} = \mathscr{L}_{w'w}\frac{1}{s}\frac{b_w\rangle}{b_{w'}\rangle}, \tag{10}$$

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

As a consequence, the probability (7) of a spike sequence reads:

$$Prob\,([\omega]_1^n) = \mu_{w_1}\frac{1}{s^n}\frac{b_{w_n}\rangle}{b_{w_1}\rangle}\prod_{k=1}^{n-1}\mathscr{L}_{w_k w_{k+1}}. \tag{11}$$

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[3] ψ which associate to a raster plot ω a real number. 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 (1)

[3]Some regularity conditions, associated with a sufficiently fast decay of the potential at infinity, are also required, that we do not state here [38].

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 $\mathscr{L}(\psi)$ given by:

$$\mathscr{L}_{w'w}(\psi) = e^{\psi_{w'w}}, \tag{12}$$

It is primitive by construction[4] so it obeys Perron-Frobenius theorem. We define:

$$P(\psi) = \log s, \tag{13}$$

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 $\mathscr{L}(\psi)$ 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 (11) of a spike sequence then reads:

$$Prob([\omega]_1^n) = \mu_{w_1} \frac{1}{e^{nP(\psi)}} \frac{b_{w_n}\rangle}{b_{w_1}\rangle} e^{\sum_{k=1}^{n-1} \psi_{w_k w_{k+1}}}. \tag{14}$$

This is an example of *a Gibbs distribution* [5] *associated with the potential* ψ. From now on we note μ_ψ this probability and since the probabilities of events are referring to μ_ψ we note $\mu_\psi([\omega]_1^n)$ instead of $Prob([\omega]_1^n)$.

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

$$Z(w') = \sum_{w \in \mathscr{A}} e^{H_{w'w}} \tag{16}$$

[4] Take two symbols w', w correspond to blocks $w' \sim \alpha'_1 \dots \alpha'_R$, $w \sim \alpha_1 \dots \alpha_R$ where $\alpha'_k, \alpha_k \in \mathscr{A}, k = 1 \dots R$. 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 $w_1 = w'$, $w_2 = \alpha'_2 \dots, \alpha'_R\, \alpha_1, \dots, w_{R+1} = \alpha'_R\, \alpha_1 \dots \alpha_{R-1}$, $w_{R+2} = w$. 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 $\mathscr{L}_{w_k w_{k+1}}(\psi)$ is positive and therefore the matrix element $\mathscr{L}^{R+2}_{w'w}$ is positive.

[5] According to [5] which is the more general definition that we now, although equivalent definitions exist (see e.g. [68]), μ is a Gibbs distribution for the potential ψ if for any $n > 0$ there are some constants with $0 < c_1 < c_2$ such that the probability of a spike block $[\omega]_1^n$ obeys:

$$c_1 \leq \frac{\mu_\psi([\omega]_1^n)}{\exp\left[-nP(\psi) + \sum_{k=1}^n \psi_{w_k w_{k+1}}\right]} \leq c_2. \tag{15}$$

Since $b_w\rangle > 0$ from PF theorem and assuming $\mu_{w_1} > 0$ (otherwise the probability (14) is zero) one may take $0 < c_1 < \mu_{w_1} \frac{b_{w_n}\rangle}{b_{w_1}\rangle} < c_2$.

since $\sum_{w\in\mathscr{A}} e^{H_{w'w}} = \sum_{w\in\mathscr{A}} L_{w'w} b_w\rangle = s b_{w'}\rangle$ 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:

$$\mathscr{M}_{w'w} = Prob(w\,|w') = \frac{1}{Z(w')} e^{H_{w'w}}, \tag{17}$$

which has the classical form "$\frac{1}{Z}e^{-\beta H}$" but where Z depends[6] on w' explaining the terminology "conditional partition function" (compare with eq. (1) in ref [43]). 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 (16) since H depends on $b\rangle$.

2.3 From parametric Gibbs potential to Markov chains

In the previous section we have seen how starting from conditional intensities $Prob\left(\omega(t)\,|\,[\omega]_{t-R}^{t-1}\right)$ one can construct a Markov chain whose invariant probability is a Gibbs distribution μ_ψ. However, in real situations neither $Prob\left(\omega(t)\,|\,[\omega]_{t-R}^{t-1}\right)$ 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[7] 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:

$$\psi_{w_k,w_{k+1}} = \sum_{n=1}^{R} \sum_{(i_1,t_1),\dots,(i_n,t_n)\in\mathscr{P}(N,R),} \lambda^{(n)}_{i_1,t_{i_1},\dots,i_n,t_{i_n}} \omega_{i_1}(k+t_{i_1})\dots\omega_{i_n}(k+t_{i_n}), \tag{18}$$

where we used the encoding introduced in section 2.1.2 (eq. (1)) and where $\mathscr{P}(N,R)$ is the set of non repeated pairs of integers with $i \in \{0,\dots,N-1\}$ and $t_{i_l} \in \{-R,\dots,0\}$.

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

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

Since we are dealing with stationary dynamics ψ is translation invariant (the $\lambda^{(n)}_{i_1,t_{i_1},\ldots,i_n,t_{i_n}}$'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 [13]) and is nothing but a Taylor expansion of $\log(P\left[\omega(0)\,|\,[\omega]_{-R}^{-1}\right])$, where one collects all terms of the form $\omega_{i_1}^{k_1}(t_{i_1})\ldots\omega_{i_n}^{k_n}(t_{i_n})$, for integer $k_1,\ldots k_n$'s, using that $\omega_i^k(t)=\omega_i(t)$, for any $k>0$ 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 (18) remains quite cumbersome since the number of terms in (19) 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:

$$\psi_{w'w}=\sum_{l=1}^{L}\lambda_l\phi_l(w',w), \tag{19}$$

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 $w'\sim[\omega]_{-R}^{-1}$, $w\sim[\omega]_{-(R+1)}^{0}$ 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, ..., neuron i_n fires at time t_n". The choice of the parametric form (19) defines what we call a "statistical model", namely a Gibbs distribution.

2.3.3 The normalization problem

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 (18), which is the polynomial expansion of the log of the conditional intensity, to a generic parametric form (19), one introduces several biases. First, one may add terms which are not in the original potential. Second, since (18) 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 (19), 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[8]. 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 (Krylov subspace iterations methods are also available [23]).

[8]Note that this is the key that conduces to a the change of paradigm for the parametric estimation from Markov chain sampling to matrix computations (See section (3.3) for the numerical development).

2.3.4 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 $w \sim \omega(0)$ is independent of $w' \sim \omega(-1)$. In this case, $\psi_{w'w} \equiv \psi_w$ and $\mathscr{L}_{w'w} = e^{\psi_w}$ does not depend on w'. As a consequence, all columns are linearly dependent which implies that there are $N-1$ 0-eigenvalues while the largest eigenvalue is $s = \sum_{w \in \mathscr{A}} e^{\psi_w} \stackrel{\text{def}}{=} Z$. 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 (9), $\mu_{\psi_w} = \frac{e^{\psi_w}}{Z}$. Here, we have the classical form of a Gibbs distribution with a constant partition function Z and the normalization only consists of subtracting $\log(Z)$ to ψ.

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

$$\psi_w = \sum_{i=0}^{N-1} \lambda_i \omega_i(0). \tag{20}$$

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 $Prob\,(\omega_i(0)=1) = \frac{e^{\lambda_i}}{1+e^{\lambda_i}}$. Hence, the parameter λ_i is directly related to the so-called firing rate, $r_i = Prob\,(\omega_i(0)=1)$.

"Ising" like potentials. This type of range-1 potential has been used by Schneidman and collaborators in [70]. It reads, in our notations,

$$\psi_w = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1}\sum_{j=0}^{i-1} \lambda_{ij}\omega_i(0)\omega_j(0). \tag{21}$$

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 memory-less.

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 ([63, 65, 19]).

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

$$\psi(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1}\sum_{j=0}^{i-1}\sum_{\tau=-1}^{0} \lambda_{ij\tau}\omega_i(0)\omega_j(\tau). \tag{22}$$

This case has been investigated in [43] using the computation of the conditional partition function (16) with a detailed balance approximation.

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

$$\psi(\omega) = \sum_{i=0}^{N-1} \lambda_i \omega_i(0) + \sum_{i=0}^{N-1}\sum_{j=0}^{i-1}\sum_{\tau=-k}^{k} \lambda_{ij\tau}\omega_i(0)\omega_j(\tau), \tag{23}$$

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

Pairwise Time-Dependent k (PTD-k) potentials.
A variation of (23) is to avoid the explicit constraints associated to firing rates :

$$\psi(\omega) = \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \sum_{\tau=-k}^{k} \lambda_{ij\tau} \omega_i(0) \omega_j(\tau), \tag{24}$$

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 $\mu_\psi[\phi_l]$ the average of ϕ_l with respect to μ_ψ. Since $\mu_\psi[\phi_l] = \sum_{w', w \in \mathscr{A}} \mu_\psi(w', w)\, \phi_l(w', w)$ one obtains:

$$\mu_\psi[\phi_l] = \sum_{w', w \in \mathscr{A}} \mu_{\psi_{w'}} M_{w'w}\, \phi_l(w', w), \tag{25}$$

where $\mu_{\psi_{w'}}$ is given by (9) and $M_{w'w}$ by (10). This provides a fast way to compute $\mu_\psi[\phi_l]$.

2.4.2 The topological pressure.

The PF theorem gives a direct access to the topological pressure $P(\psi)$ which is the logarithm of the leading eigenvalue s, easily obtained by a power method (see eq. (8)). In the case of range-$R+1$ potentials (19) 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:

$$\frac{\partial P(\lambda)}{\partial \lambda_l} = \mu_\psi[\phi_l]. \tag{26}$$

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.

2.4.3 The Maximal Entropy (MaxEnt) principle.

Gibbs distributions obeys the following variational principle:

$$P(\psi) \stackrel{\text{def}}{=} h(\mu_\psi) + \mu_\psi(\psi) = \sup_{\mu \in m^{(inv)}(\Sigma)} h(\mu) + \mu(\psi), \tag{27}$$

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[9]. When $\mu(\psi)$ is imposed (e.g. by experimental measurement), this corresponds to finding a probability that maximizes the statistical entropy under constraints [34]. The value of the maximum is the topological pressure.

2.4.4 Computing the entropy

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

$$h\left[\mu_\psi\right] = P(\lambda) - \sum_l \lambda_l \mu_\psi\left[\phi_l\right]. \tag{30}$$

2.4.5 Comparing several Gibbs statistical models.

The choice of a potential (19), 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,

$$d(\nu,\mu_\psi) = \limsup_{n\to\infty} \frac{1}{n} \sum_{[\omega]_0^{n-1}} \nu\left([\omega]_0^{n-1}\right) \log\left[\frac{\nu\left([\omega]_0^{n-1}\right)}{\mu_\psi\left([\omega]_0^{n-1}\right)}\right], \tag{31}$$

where ν and μ_ψ are two invariant probability measures, provides some notion of asymmetric "distance" between μ_ψ 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, 66, 38, 16]:

$$d\left(\nu,\mu_\psi\right) = P(\psi) - \nu(\psi) - h(\nu). \tag{32}$$

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

[9] This is:

$$h[\mu] = \lim_{n\to+\infty} \frac{h^{(n)}[\mu]}{n}, \tag{28}$$

where

$$h^{(n)}[\mu] = -\sum_{\omega\in\Sigma^{(n)}} \mu\left([\omega]_0^{n-1}\right) \log\mu\left([\omega]_0^{n-1}\right), \tag{29}$$

$\Sigma^{(n)}$ 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.

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:

$$\bar{f}^{(\mathcal{N},T)} = \frac{1}{\mathcal{N}T} \sum_{m=1}^{\mathcal{N}} \sum_{t=1}^{T} f(\sigma^t \omega^{(m)}), \tag{33}$$

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[10] 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[11] $f : \Sigma \to \mathbb{R}$,

$$\pi^{(T)}(f) = \bar{f}^{(\mathcal{N},T)}. \tag{34}$$

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

$$\pi^{(T)}\left([\omega]_{t_1}^{t_2}\right) = \frac{1}{\mathcal{N}T} \sum_{m=1}^{\mathcal{N}} \sum_{t=1}^{T} \chi_{[\omega]_{t_1}^{t_2}}(\sigma^t \omega^{(m)}), \tag{35}$$

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 an 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 $T \to \infty$.

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[12] eq. (32) :

$$d(\pi^{(T)}, \mu_\psi) = P(\psi) - \pi^{(T)}(\psi) - h(\pi^{(T)}), \tag{36}$$

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

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

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

[12] This is an approximation because $\pi^{(T)}$ is not invariant [38]. It becomes exact as $T \to +\infty$.

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:

$$\tilde{h}[\psi] = P[\psi] - \pi^{(T)}(\psi), \tag{37}$$

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

This relation holds for any potential. In the case of a parametric potential of the form (19) we have to minimize

$$\tilde{h}[\lambda] = P[\lambda] - \sum_{l=1}^{L} \lambda_l \pi^{(T)}(\phi_l). \tag{38}$$

Thus, from (26) and (34),(33), given the parametric form, the set of λ_l's minimizing the KL divergence are given by:

$$\mu_\psi[\phi_l] = \pi^{(T)}(\phi_l), \qquad l = 1 \ldots L. \tag{39}$$

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 (39) impose constraints on the sought Gibbs distribution. In view of the variational principle (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 (39)*, 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 funding paper of Schneidman et al. [70]. There 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* [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 (38), i.e. knowing if the " model"ψ_2 is significantly "better" than ψ_1, reduces to verifying:

$$\tilde{h}[\psi_2] \ll \tilde{h}[\psi_1], \tag{40}$$

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 Convexity and estimation well-definability.

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 (38), 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 (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 (38), 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 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[13]. Examples of these cases are provided in section 4. 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 consists of starting from the largest possible potential with range $R+1$ [14] . 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 [15], $L_{eff} < 2^{N(R+1)}$) which reduces the number of situations one can numerically handle. These two approaches are used in section 4.

[13] 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 [48]. Some of the current available tests might offer additional algorithmic tools that would be explored in a forthcoming paper

[14] ibid.

[15] 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.

2.5.5 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. (33) 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]:

$$\mu_\psi \left\{ |\pi^{(T)}(\phi_l) - \mu_\psi(\phi_l)| \geq \varepsilon \right\} \sim \exp(-T I_l(\varepsilon)), \tag{41}$$

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 [49].

Another important large deviations property also results from the present formalism [38, 14, 22]. 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 $\mu_{\psi^*} \left\{ \|\pi^{(T)} - \mu_{\psi^{(test)}}\| < \varepsilon \right\}$ that $\pi^{(T)}$ is ε-close to the "wrong" probability $\mu_{\psi^{(test)}}$ decays exponentially fast as:

$$\mu_{\psi^*} \left\{ \|\pi^{(T)} - \mu_{\psi^{(test)}}\| < \varepsilon \right\} \sim \exp(-T \inf_{\mu, \|\mu - \mu_{\psi^{(test)}}\| < \varepsilon} d(\mu, \mu_{\psi^*}))). \tag{42}$$

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 $\frac{\mu_{\psi^*}\left\{\|\pi^{(T)} - \mu_{\psi^*}\| < \varepsilon\right\}}{\mu_{\psi^*}\left\{\|\pi^{(T)} - \mu_{\psi^{(test)}}\| < \varepsilon\right\}}$ 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.

2.5.6 Other statistics criteria for Gibbs distributions and 'test Statistics'.

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 $P(\lambda)$. 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 [49]. 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 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[16] procedure proposed here decomposes in three steps:

1. *Choosing a statistical model*, i.e. fixing a guess potential (19) (equivalently, the set of monomials).
2. *Computing the empirical average of monomials*, i.e. determining them from the empirical raster, using eq. (33).
3. *Performing the parametric estimation*, i.e. use a variational approach to determine the value of the λ_l's.

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

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, [29] and developed by [70];
- 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[17] 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.

[16] The code is available at `http://enas.gforge.inria.fr/classGibbsPotential.html`
[17] Subtracting the firing rates of i and j.

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

Rate potential are range-1 potentials, as defined in eq. (20). 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 ([70]) and Shlens and collaborators ([77]) but previously introduced in neuroscience by other authors(see [44, 32] for historical references). These potentials take in account rate and synchronization of neurons pairs, as studied in, e.g. [29]. 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.

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

These potentials defined previously by eq. (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 ([43, 32, 64]).

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. [24, 29, 6] for important applications).

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

3.1.4 Taking plasticity into account: "STDP" potentials

In [10] we considered Integrate-and-Fire neural networks with Spike-Time Dependent Plasticity of type:

$$W'_{ij} = \varepsilon \left[r_d W_{ij} + \frac{1}{T} \sum_{t=T_s}^{T+T_s} \omega_j(t) \sum_{u=-T_s}^{T_s} f(u)\, \omega_i(t+u) \right], \tag{43}$$

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]. $T_s \stackrel{\text{def}}{=} 2\max(\tau_+, \tau_-)$ is a characteristic time scale. We argued that this synaptic weights adap-

tation rule produces, when it has converged, spike trains distributed according to a Gibbs distribution with potential:

$$\psi(\omega) = \sum_{i=0}^{N} \lambda_i^{(1)} \omega_i(0) + \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \lambda_{ij}^{(2)} \sum_{u=-T_s}^{T_s} f(u)\omega_i(0)\omega_j(u). \quad (44)$$

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 $\mathscr{P}_s$ 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:

$$\psi_{\mathscr{P}_s}(\omega) = \sum_{i\in\mathscr{P}_s} \lambda_i^{(1)} \omega_i(0) + \sum_{i,j\in\mathscr{P}_s 0} \sum_{j=0}^{N-1} \lambda_{ij}^{(2)} \sum_{u=-T_s}^{T_s} f(u)\omega_i(0)\omega_j(u). \quad (45)$$

Here, the potential is a function of both past and future. A simple way to relate this potential to a conditional intensity, 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 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 $N(R+1) \leq 18$. Such an implementation is now available[18]. The handling of larger neural ensembles and/or ranges require an extension of the current implementation, using parallel computing algorithms, sparse matrix storage techniques and/or distributed memory.

3.2 Computing the empirical measure: prefix-tree construction.

For one sample ($\mathscr{N} = 1$) the empirical probability (34) 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, $\pi^{(T)}([\omega]_{-D}^{t}) = \pi^{(T)}([\omega]_{0}^{t+D})$.

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

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 $D \leq R+1$[19]. It is known (see, e.g., [30] 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 depend 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. [30, 28], we detail the method here.

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 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.

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).

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

3.3 Performing the parametric estimation

In a nutshell, the parametric estimation reduces to minimizing (36), hence (37), by calculating the topological pressure $P(\psi) \equiv P(\lambda)$ using (8) 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.

3.3.1 Potential eigen-elements calculation.

It has been shown in the theoretical section that the Perron-Frobenius matrix eigenelements 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 $\mathscr{L}(\psi)$ matrix, associated to the highest eigenvalue, using a standard power-method series [20]:

$$\begin{array}{rcl} s^{(n)} & = & \|\mathscr{L}(\psi)\, v^{(n-1)}\| \\ v^{(n)} & = & \frac{1}{s^{(n)}}\mathscr{L}(\psi)\, v^{(n-1)} \end{array}$$

   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 (8). 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 $\mathscr{L}(\psi)$ is a sparse $2^{NR} \times 2^{NR}$ matrix, as outlined in the section 2.1.2. As a consequence calculating $\mathscr{L}(\psi)\, v$ is a $O(2^{N+NR}) \ll O(2^{2NR})$ 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).

   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 $\mathscr{M}$, which can be used to generate spike trains distributed according the Gibbs distribution μ_ψ and used as benchmarks in the section 4.

[20] This choice is not unique and several alternative numerical methods exists (e.g Krylov subspace methods) [23].

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 probability given by (9), from which probabilities of any block can be computed (eq.7, 11).

   (b) The theoretical value of the observables average $\mu_\psi[\phi_l]$, as given in (25).

   (c) The theoretical value of the distribution entropy $h\left[\mu_\psi\right]$, as given in (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 guess 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 guess Gibbs distribution fits at best with the empirical measure. We have discussed why minimizing (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. (38), where $\mu_\psi[\phi_l]$ is given by (25. 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[21] to delegate the minimization of this criterion to any standard powerful non-linear minimization routine.

We have implemented such a mechanism using the GSL[22] implementation of non-linear minimization methods. We have also made available the GSL implementation

[21] 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:

$$\tilde{h}^{k+1} = \tilde{h}^k - \varepsilon^k \left|\frac{\partial \tilde{h}}{\partial \lambda_l}\right|^2 + O\left((\varepsilon^k)^2\right).$$

[22] 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,

of the simplex algorithm of Nelder and Mead which does not require the explicit computation of a gradient like in eq. (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.

3.3.4 Measuring the precision of the estimation.

Once the quantity $\tilde{h}[\psi] = P[\psi] - \pi^{(T)}(\psi)$ (eq. (38)) 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 (28) and (29). In order to interpolate the limit (29), we have adapted an interpolation method from [30] 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 benefit from this part of the work.

3.4.1 Main properties of the algorithm.

Convexity. As indicated in the section 2.5.4 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 [21].

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.

Finite sample effects. As indicated in the section 2.5.5 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 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 limit the stability of the method, from a numerical point of view.

The first factor is that the matrix $\mathscr{L}(\psi)$ 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 $\mathscr{L}(\psi)$, and numerical instabilities may occur. Now, a strong negative value of the potential corresponds to events with small conditional probability and this instability can be removed by considering that these events have in fact a zero probability. In this case, the corresponding matrix-element is directly set to zero without computing the corresponding potential value. The criterion that we used is to consider that events appearing with an empirical probability less than x per cents (where x is a parameter of the method typically fixed to $\frac{10}{T}$) are artifacts and are attributed a zero probability.

On the other hand, large values of the potential corresponds to events having a high conditional probability. Now, adding to the potential some constant λ_0 does not change the corresponding normalised potential (see eq. (10)). Indeed, this corresponds to multiplying $\mathscr{L}(\psi)$ by e^{λ_0}; in this case s is multiplied by e^{λ_0} while the corresponding eigenvectors are unchanged. This allows to remove the instability due to high values of the potential. When this event occurs (typically for values higher than 10^4) a warning is generated in the code.

Moreover, several coherent tests regarding the calculation of the PF eigen-elements have been 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}). When computing the normalized potential (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 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 [30, 71]. See [28] 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 [71] is not directly tested in [28], 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 [71] is directly generalizable to non-binary (thus multi-neurons) spike trains, whereas the context-tree methods seems intrinsically 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 statistical entropy from eq. (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.

3.4.2 Key aspects of the numerical implementation.

Unobserved blocks.

We make here the (unavoidable) approximation that unobserved blocks or blocks observed with a too weak statistics correspond to forbidden words (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 [10] 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 $\mathbf{Q}_{l,w} = \psi_l([\omega]_{-R}^{0})$, where w is given by (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 (19) 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 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 [23], 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 $\frac{\partial \tilde{h}}{\partial \lambda_l} \simeq \frac{\tilde{h}(\lambda_l+\varepsilon)-\tilde{h}(\lambda_l-\varepsilon)}{2\varepsilon}$ 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 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 [80].

Using other approximations of the KL-divergence criterion. Let us now discuss another class of variants: the proposed KL-divergence criterion in (31) and its empirical instantiation in (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.

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

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 (10) 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.

Estimation in the case of a normalized potential. In the case where the potential is normalized, the criterion (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.

4 Results

4.1 Basic tests: validating the method

4.1.1 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 (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 (19) and a number of neurons N we apply the following method:

1. Randomly choosing the parameter's values λ_l, $l = 1 \ldots L$ 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 (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 has 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.

4.1.2 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:

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

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. (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 (19) 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. 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.

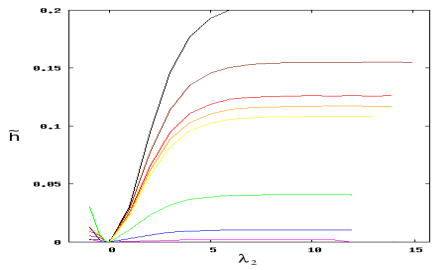


Figure 1: Entropy difference, using $\tilde{h}$, defined in (38), between the estimations of a range-1 and a range-2 model. The range-2 model reads $\psi_2 = \lambda_1\,\omega_0(0) + \lambda_2\omega_0(0)\,\omega_0(1)$ for $\lambda_1 = \{-1\ (black), -0.5\ (brown), -0.2\ (red), -0.1\ (orange), 0\ (green), 1\ (blue), 2\ (Magenta)\}$. 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 $\mathscr{L}(\psi)$ 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 deter-

mination of the largest eigenvalue of the $\mathscr{L}(\psi)$ 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 $T \simeq 10^5$.

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 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 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			

Table 2: Numerical precision of the method in the same conditions as table 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			

4.2 More general tests

4.2.1 Test framework.

In order to test more general potentials for $N = 2$ neurons we have studied the forms (21), (23), (24), that we recall here:

$$\begin{aligned} \text{Ising} : \psi(\omega) &= \lambda_1\, \omega_1(0) + \lambda_2\, \omega_2(0) + \lambda_3\, \omega_1(0)\, \omega_2(0). \\ \text{RPTD} - k : \psi(\omega) &= \lambda_1\, \omega_1(0) + \lambda_2\, \omega_2(0) + \sum_{i=-k}^{i=k} \hat{\lambda}_i\, \omega_1(0)\omega_2(i). \\ \text{PTD} - k : \psi(\omega) &= \sum_{i=-k}^{i=k} \hat{\lambda}_i\, \omega_1(0)\omega_2(i). \end{aligned} \tag{46}$$

Test 1 (estimation precision). Given a generating potential ψ^* of the form (46) 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 occuring 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 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 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;

(ii) the error is independent of the initial bias percentage (see figs 4.2.1);

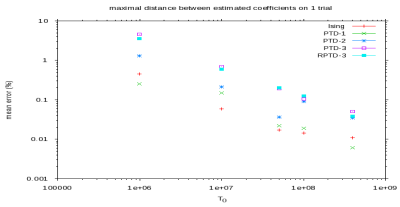


Figure 2: Mean error (in percentage) vs T_0 size.

Test 2 (models comparison). We select a potential ψ^* from (46); 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 (46). 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 :

(i) The statistical models with lowest mean value KL divergence have the same monomials as ψ^*, plus possibly additional monomials, in agreement with section 2.5.4.

(ii) For all these models, the criterion $\tilde{h}[\psi]$ (38) averaged over trials, is fairly equal up to a difference of order $\eta \approx 10^{-5}$, while the difference with respect to other types of statistical models is at least of 3 orders of magnitude larger. We recall that, according to section 2.5.5, the deviation probability is of order to $\exp(-\eta T)$. After estimation from a raster generated with an Ising model, the ratio δ of the deviation probabilities (42) between an **Ising** and a **RPTD-**1 model is $\sim \delta = \exp(-0.000051 \times 10^8)$, while between the **Ising** and the **PTD-**3 $\sim \delta = \exp(-0.0194 \times 10^8)$ 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 poten-

tial 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).

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 (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 [26] 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)[24]. Figure 5 shows, in log-scale, the box-plot for the distribution of the quantity defined as:

$$\varepsilon(w) = \left| \left(P_{est}(w) - P_{emp}(w)\right) / \sigma_{emp}(w) \right| \tag{47}$$

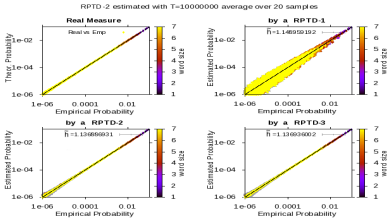
that is taken as a weighted measure of the deviations. We have considered this distribution when it takes into account, either all the words up to a given size R_{max}, or only the words of that given size. There is no visual difference for $R_{max} = 7$. The results shows that only models containing 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 (47). 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 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 (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 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 explicitly and rigorously known [12] while effects of synaptic plasticity on statistics have been studied in [10].

[24] 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.



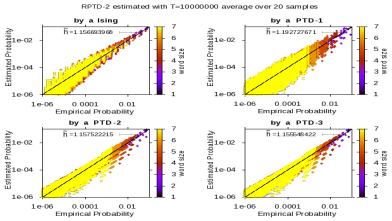
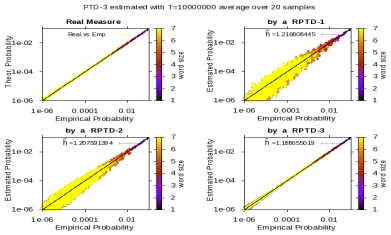


Figure 3: Figure 3a (top left) Expected probability μ_ψ versus empirical probability $\pi^{(T)}(w)$; Figure 3b (top right) to 8 (bottom right) Predicted probability versus empirical probability $\pi^{(T)}(w)$ for several guess potentials The generating potential is a **RPTD-2**.



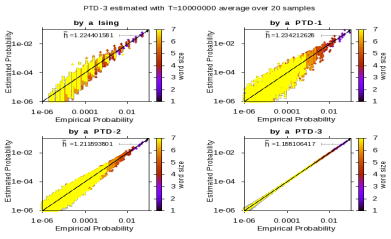


Figure 4: Same as previous figure where the generating potential is a **PTD-3**.

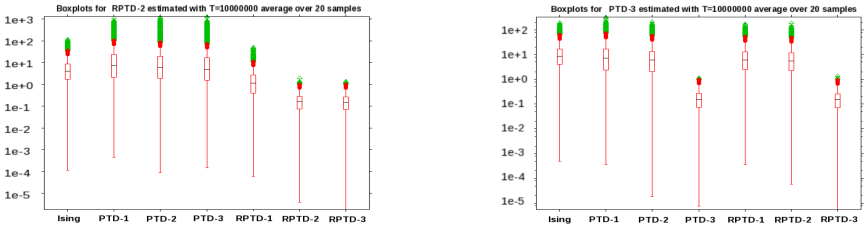


Figure 5: The box-plot (in log-scale) of the distributions of weighted deviations of word's probability versus their empirical probability, for several statistical models, using a generating potential of the form **(left) RPTD-**2 and **(right) PTD-**3. Midliers Outliers (see footnote 24) are shown by red dots and extreme outliers by green dots.

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

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 variable with variance 1. The network dynamics is given by:

$$V_i(t+1) = \gamma V_i(1 - Z[V_i(t)]) + \sum_{j=1}^{N} W_{ij} Z[V_j(t)] + I_i^{ext} + \sigma_B B_i(t); \qquad i = 1 \ldots N, \quad (48)$$

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 (48) 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. [33], provides a rough but realistic approximation of biological neurons behaviors. Its dynamics has been fully characterized for $\sigma_B = 0$ in [9] while the dynamics with noise is investigated in [12]. Its links to more elaborated models closer to biology is discussed in [11].

4.3.2 Exact spike trains statistics.

For $\sigma_B > 0$ in model (48), there is a unique Gibbs distribution in this model, whose potential is explicitly known. It is given by:

$$\psi(\omega_{-\infty}^0) = \sum_{i=1}^{N} \left[\omega_i(0) \log\left(\pi\left(\frac{\theta - C_i(\underline{\omega})}{\sigma_i(\underline{\omega})} \right) \right) + (1 - \omega_i(0)) \log\left(1 - \pi\left(\frac{\theta - C_i(\underline{\omega})}{\sigma_i(\underline{\omega})} \right) \right) \right], \quad (49)$$

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 in replacing $\underline{\omega} = \omega_{-\infty}^{-1}$ by ω_{-R}^{-1} in (49). 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 (18).

4.3.3 Numerical estimation of spike train statistics

Here we have considered only one example of model (48) (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:

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

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 (49), 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[25] (as proposed in [43]) ψ^{MEDF} ;
- (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 (4). Note that the estimated KL divergence of range-1 potentials slightly depend on R since the $\mathscr{L}(\psi)$ matrix, and thus the pressure, depend on R.

Table 4: Kullback-Leibler divergence between the empirical measure of a raster generated by (48) (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).

4.3.4 Synaptic plasticity.

Here the neural network with dynamics given by (48) has been submitted to the STDP rule (43). The goal is to check the validity of the statistical model given by (44), predicted in [10]. 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 [9, 11] 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

[25] or equivalently, a **RPTD-**1 from (46)

current $\mathbf{I}^{(ext)}$ in eq. (48) 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 (43). In this context we expect the potential for the whole population to be of the form (44) and for a subset of the population of the form (45). 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 (46), we estimate the coefficients that minimizes the Kullback-Leibler divergence.The probability of words of different sizes predicted by several statistical models from (46) 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 (6).
Results depicted on figure (6) show, on one hand, that the statistics is well fitted by (45). 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 (6)). 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 (46) with time-varying coefficients $\psi(t,\omega) = \lambda_1(t)\,\omega_1(0) + \lambda_2(t)\,\omega_2(0) + \lambda_3(t)\,\omega_1(0)\,\omega_2(0)$. 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 corresponding to the $\mathscr{L}(\psi)$ matrix. From this, we use equation (11) (with a time dependent matrix $\mathscr{L}$) 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 (7) 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.

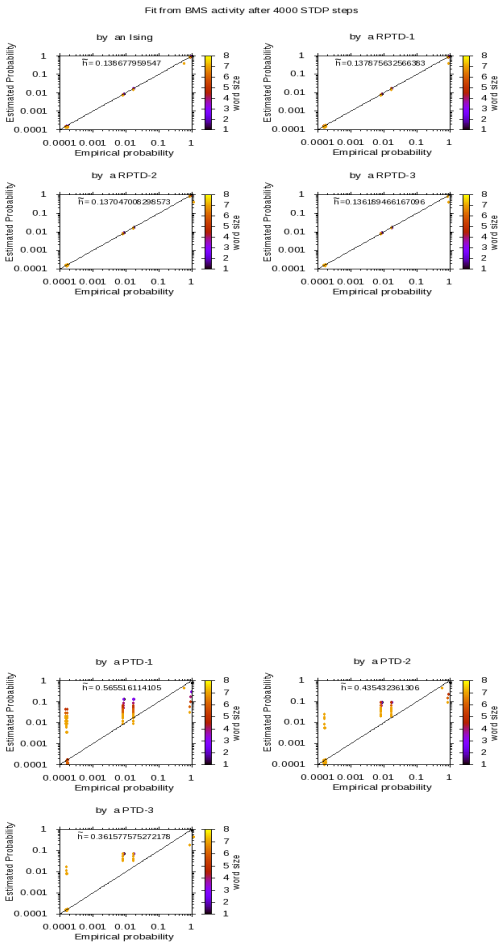


Figure 6: The probability of words of different sizes predicted by several statistical models from (46) versus empirical probability $\pi_\omega^{(T)}(w)$ obtained from a spike train generated by dynamics (48) after 4000 epochs of adaptation.The $\tilde{h}$ value (38) for each fitting model is shown inside the graphic. The potential is a pair potential of the form (45). Recall that **RPTD** Models include firing rates but **PTD** models do not.

Figure 7: **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).

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 the type of constraints. Actually, on mathematical grounds, it allows infinitely many constraints. Moreover, we do not need to compute the entropy.
- **Monte-Carlo methods.** Equation (11) enables us to generate spike trains Gibbs-distributed with and arbitrary potential (not normalized). The convergence is ensured by eq. (8). We do not need to assume detailed balance.
- **Boltzmann learning.** Our approach can be viewed as "Boltzmann learning" (as presented e.g. in [65]) 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 time-evolution of Gibbs Distributions induced by adaptation mechanisms as developed in [10]. Furthermore, the estimation method proposed in this paper (i.e, the learning scheme we implemented) remains well-behaved in the case of time-dependent 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 ([64]).

- **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 intrisic 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 [61, 88]) 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. [70, 47]). 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. An incoming re-implementation is going to overcome this barrier at the numerical level, for either relatively small raster or sparse distributions. However, this is not a problem intrinsic to our approach but to parametric estimations potentials of the form (19). We believe that other form of potential could be more efficient (see [12] 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 (19) seems rather delicate, as suggested by the nice work in [43] 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 [46] or in neural networks [78]. 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 [75]). Examples are the replica method [46], Thouless-Anderson-Palmer equations [82], the Plefka expansion [55], or more recently e.g. the Sessak-Monasson approximation [74] (for a recent review on mean-field methods see [52]). Since the seminal paper by Schneidman and collaborators [70] 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. [79, 78, 3, 67, 25] 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 (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 [3] and neural networks [67, 25]). This is what the present formalism offers.

5.2 Conclusion and perspectives

The present formalism allows us to provide closed-form calculations of interesting parameters related to spectral properties of the $\mathscr{L}(\psi)$ matrix. 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 current state of the art, the method we propose 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 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 [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 syn-fire chains [31] or polychronism mechanisms [54]. This may be the limit of the present class of approaches, and things are to be thought 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.

Acknowledgments

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):189–200, 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. A discrete time neural network model with spiking neurons. rigorous results on the spontaneous dynamics. *J. Math. Biol.*, 56(3):311–345, 2008.

[10] 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.

[11] B. Cessac and T. Viéville. On dynamics of integrate-and-fire neural networks with adaptive conductances. *Frontiers in neuroscience*, 2(2), July 2008.

[12] Bruno Cessac. A discrete time neural network model with spiking neurons ii. dynamics with noise. *J.Math. Biol., accepted*, 2010.

[13] Bruno Cessac. A discrete time neural network model with spiking neurons ii. dynamics with noise. *J.Math. Biol., accepted*, 2010.

[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):265–275, 1988.

[19] Simona Cocco, Stanislas Leibler, and Rémi Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *PNAS*, 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] James W. Demmel. *Applied Numerical Linear Algebra*. SIAM, 1 edition, August 1997.

[24] M. Diesmann, M-O. M-O Gewaltig, and A. Aertsen. Stable propagation of synchronous spiking in cortical neural networks. *Nature*, 402:529–533, 1999.

[25] 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.

[26] Michael Frigge, David C Hoaglin, and Boris Iglewicz. Implementation of the boxplot. *The American Statistician*, 43(1):50–54, February 1989.

[27] F. R. Gantmacher. *the theory of matrices*. AMS Chelsea Publishing, Providence, RI, 1998.

[28] 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.

[29] F. Grammont and A. Riehle. Precise spike synchronization in monkey motor cortex involved in preparation for movement. *Exp. Brain Res.*, 128:118–122, 1999.

[30] Peter Grassberger. Estimating the information content of symbol sequences and efficient codes. *IEEE Transactions on Information Theory*, 35, 1989.

[31] 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.

[32] Shun ichi Amari. Information geometry of multiple spike trains. In Sonja Grün and Stefan Rotter, editors, *Analysis of Parallel Spike trains*, volume 7 of *Springer Series in Computational Neuroscience*, part 11, pages 221–253. Springer, 2010. DOI: 10.1007/978-1-4419-5675.

[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):493–501, August 1983.

[36] R. E. Kass and V. Ventura. A spike-train probability model. *Neural Comput.*, 13(8):1713–1720, 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] 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.

[44] 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.

[45] 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.

[46] M. Mézard, G. Parisi, and M.A. Virasoro. *Spin-glass theory and beyond*. World scientific Singapore, 1987.

[47] 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.

[48] Gusztáv Morvai and Benjamin Weiss. Estimating the lengths of memory words. *IEEE Transactions on Information Theory*, 54(8):3804–3807, august 2008.

[49] 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.

[50] Ifije E. Ohiorhenuan, Ferenc Mechler, Keith P. Purpura, Anita M. Schmid, Qin Hu, and Jonathan D. Victor. Sparse coding and high-order correlations in fine-scale cortical networks. *Nature*, 466(7):617–621, 2010.

[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):1927–1961, 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] T. Plefka. Convergence condition of the tap equations for the infinite-ranged ising spin glass model. *J. Phys. A*, 15:1971, 1982.

[56] Alexandre Pouget and Gregory C DeAngelis. Paying attention to correlated neural activity. *Nature Neuroscience*, 11(12):1371–1372, December 2008.

[57] 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.

[58] 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.

[59] Christophe Pouzat and Antoine Chaffiol. On goodness of fit tests for models of neuronal spike trains considered as counting processes, 2009.

[60] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. Bradford Books, 1997.

[61] J.S. Rosenthal. Parallel computing and monte carlo algorithms. *Far East J. Theor. Stat.*, 4:207–236, 2000.

[62] Y. Roudi, S. Nirenberg, and P.E. Latham. Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't. *PLOS Computational Biology*, 5(5), 2009.

[63] Yasser Roudi, Erik Aurell, and John A Hertz. Statistical physics of pairwise probability models. *Frontiers in Computational Neuroscience*, page 15, 2009.

[64] Yasser Roudi and John Hertz. Mean field theory for non-equilibrium network reconstruction. *arXiv*, page 11, Sept 2010.

[65] 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.

[66] D. Ruelle. *Statistical Mechanics: Rigorous results*. Benjamin, New York, 1969.

[67] M. Samuelides and B. Cessac. Random recurrent neural networks. *European Physical Journal - Special Topics*, 142:7–88, 2007.

[68] O. Sarig. Thermodynamic formalism for countable markov shifts. http://www.wisdom.weizmann.ac.il/ sarigo/TDFnotes.pdf, 2010.

[69] Michael T. Schaub and Simon R. Schultz. The ising decoder: reading out the activity of large neural ensembles. *arXiv:1009.1828*, 2010.

[70] 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.

[71] Thomas Schürmann and Peter Grassberger. Entropy estimation of symbol sequences. *Chaos*, 6(3):414–427, 1996.

[72] R. Segev, I. Baruchi, E. Hulata, and E. Ben-Jacob. Hidden neuronal correlations in cultured networks. *Physical Review Letters*, 92:118102, 2004.

[73] E. Seneta. *Non-negative Matrices and Markov Chains*. Springer, 2006.

[74] V. Sessak and Remi Monasson. Small-correlation expansions for the inverse ising problem. *J. Phys. A*, 42:055001, 2009.

[75] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Physical Review Letters*, 35(26):1792+, December 1975.

[76] 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.

[77] 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.

[78] H. Sompolinsky, A. Crisanti, and H.J. Sommers. Chaos in random neural networks. *Phys. Rev. Lett.*, 61:259–262, 1988.

[79] 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.

[80] Danny C. Sorensen. Numerical methods for large eigenvalue problems. *Acta Numerica*, 11:519–584, 2002.

[81] 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.

[82] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of ‘ solvable model of a spin glass. *Philos. Mag.*, 35:593–601, 1977.

[83] Gašper Tkačik, Jason S. Prentice, Vijay Balasubramanian, and Elad Schneidman. Optimal population coding by noisy spiking neurons. *PNAS*, 107(32):14419–14424, August 2010.

[84] Gašper Tkačik, Elad Schneidman, Michael J. Berry II, and William Bialek. Spin glass models for a network of real neurons. *arXiv*, page 15, 2009.

[85] Wilson Truccolo and John P. Donoghue. Nonparametric modeling of neural point processes via stochastic gradient boosting regression. *Neural Computation*, 19(3):672–705, 2007.

[86] 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.

[87] 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.

[88] Jun Yan, Mary Cowles, Shaowen Wang, and Marc Armstrong. Parallelizing mcmc for bayesian spatiotemporal geostatistical models. *Statistics and Computing*, 17:323–335, 2007.

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399