Hearing the Maximum Entropy Potential of a spike-generating Markov process

Rodrigo Cofre, Bruno Cessac

Hearing the Maximum Entropy Potential of a spike-generating Markov process

R. Cofré[1] and B. Cessac[1]

[1]*NeuroMathComp team (INRIA, UNSA LJAD) 2004 Route des Lucioles, 06902 Sophia-Antipolis, France*

(Dated: August 22, 2013)

We consider a spike-generating stationary Markov process whose transition probabilities are known. We show that there is a canonical potential whose Gibbs distribution, obtained from the Maximum Entropy Principle (MaxEnt), is the equilibrium distribution of this process. We provide a method to compute explicitly and exactly this potential as a linear combination of *spatio-temporal* interactions. The method is based on the Hammersley Clifford decomposition and on periodic orbits sampling. As an application, we establish an explicit correspondence between the parameters of the Ising model and the parameters of Markovian models like the Generalized-Linear Model.

PACS numbers: 87.19.lo 05.10.-a 87.10.-e 87.85.dq

The understanding of how networks of neurons encode sensory information requires a detailed description of how the synaptic interactions and stimulus shape the collective statistics of spiking responses. Several approaches are currently proposed in the literature. One research trend is based on the assumption that spikes are generated by a Markov process whose transition probabilities have to be inferred from data observations. Prominent examples are the Linear-Non Linear model (LN) [4] and the Generalized Linear Model (GLM) [1], where phenomenological forms of Markov transition probabilities are proposed. Explicit computations of transition probabilities, having a similar structure as GLM, can also be analytically derived in neural networks models [2, 3, 5]. One of the main advantages of this approach is to handle non stationary dynamics where the parameters of the model can be interpreted in terms of "structural" parameters in the neural network (synaptic weights, stimulus). Another trend is based on the Maximum Entropy Principle (MaxEnt). It consists of fixing a set of constraints, determined as the empirical average of observables measured from the spiking activity. Maximizing the statistical entropy given those constraints provides a unique probability, called a Gibbs distribution. This method suffers unfortunately several caveats: (i) It assumes stationarity in the data; (ii) The choice of constraints is ad hoc; (iii) The interpretation of the coefficients shaping the Gibbs distribution (Lagrange multipliers) is controversial.

In this paper, we focus on the points (ii) and (iii) and address the following question. Assuming that a spike train has been generated by a stationary Markov process where transition probabilities are given, can we construct a MaxEnt model, with a minimum of constraints, reproducing exactly the (spatio-temporal) statistics of this process ? In this way, we would have a method to select the constraints from the observation of the process, resolving (ii). We show that there indeed exists a canonical MaxEnt potential, having the same equilibrium distribution as the Markov process, that can be constructed explicitly and whose coefficients are explicit functions of the parameters shaping the transition probabilities. In the case of Integrate-and-Fire neural networks [2, 3, 5] or phenomenological models such as GLM this establishes an exact correspondence between the structural parameters and the Lagrange multipliers of the MaxEnt, resolving (iii). We provide here an explicit example for Ising model, although our method extends to general potentials with spatio-temporal constraints. As we show, the Ising parameters are explicit non-linear functions of structural parameters.

We study a network composed of N neurons. Time has been discretized so that a neuron can at most fire a spike within one time bin. We define a spiking variable $\omega_k(n) = 1$ if neuron k fires at time n and $\omega_k(n) = 0$ otherwise. The *spiking pattern* at time n is the vector $\omega(n) = [\,\omega_k(n)\,]_{k=1}^{N}$. A *spike block* $\omega_{n_1}^{n_2}$ is an ordered list of spiking patterns where spike times range from n_1 to n_2. The *degree* d of a block is its number of non-zero bits. A *spike train* is a spike block $\omega_0^{+\infty}$. Although experimental spike trains have always a finite duration T, it is useful to us to consider infinite spike trains with $T \to +\infty$. To alleviate notations we simply write ω instead of $\omega_0^{+\infty}$. We note Ω the set of spike trains.

An *observable* is a function $\mathcal{O}$ which associates a real number $\mathcal{O}(\omega)$ to a spike train. The *time-range* of an observable is the minimal integer $R > 0$ such that, for any spike train ω, $\mathcal{O}(\omega) = \mathcal{O}\left(\omega_0^{R-1}\right)$. An observable of range R is *time-translation invariant* if, for any time $n > 0$ we have $\mathcal{O}\left(\omega_n^{n+R-1}\right) \equiv \mathcal{O}\left(\omega_0^{R-1}\right)$ whenever $\omega_n^{n+R-1} = \omega_0^{R-1}$. We focus on time-translation invariant observables with fixed range R from now on, and we set $D = R - 1$. Prominent examples of such observables are products of the form $\prod_{u=1}^{r} \omega_{k_u}(n_u)$ called *monomials*. A monomial takes therefore values in $\{\,0, 1\,\}$, and is 1 if and only if neuron k_1 spikes at time n_1, ..., neuron k_r spikes at time n_r in the raster ω.

Each spike block is associated to a unique integer $l = \sum_{k=1}^{N} \sum_{n=0}^{D} 2^{n\,N+k-1}\,\omega_k(n)$. We note $\omega^{(l)}$ the spike block corresponding to l. Likewise, and since a monomial is defined by a set of spike events (k_u, n_u), one can associate to each monomial a spike block or "mask" where

the only bits '1' are located at (k_u, n_u), $u = 1, \ldots, r$. To this block one can thus also associate an index l, and we may label each monomial m_l. The degree of a monomial m_l is the degree of the corresponding spike block $\omega^{(l)}$. There exists a natural hierarchy between spike blocks. We define the block inclusion $\sqsubseteq$ by $\omega_0^D \sqsubseteq {\omega'}_0^D$ if $\omega_k(n) = 1 \Rightarrow \omega'_k(n) = 1$, with the convention that the block of degree 0, $\omega^{(0)}$, is included in all blocks. For two integers l, l', $m_{l'}(\omega^{(l)}) = 1$ if and only if $\omega^{(l')} \sqsubseteq \omega^{(l)}$. It follows that any observable $\mathcal{O}$ of range R admits a unique decomposition, called the Hammersley-Clifford decomposition in reference to the paper [6], where:

$$\mathcal{O}(\omega^{(l)}) = \sum_{l'=0}^{L(N,R)} o_{l'}\, m_{l'}(\omega^{(l)}) = \sum_{\omega^{(l')} \sqsubseteq \omega^{(l)}} o_{l'}, \qquad (1)$$

where $L(N, R) = 2^{NR} - 1$. Monomials constitute therefore a basis in the space of observables.

In this context we consider an "energy" or a potential:

$$\mathcal{H} = \sum_{l=0}^{L(N,R)} h_l m_l, \qquad (2)$$

where some coefficients h_l in the expansion may vanish. All h_ls are assumed to be $> -\infty$ (no hard-core interaction). By analogy with spin systems, we see from (2) that monomials somewhat constitute *spatio-temporal* interactions between spikes: degree 1 monomials corresponds to "self-interactions", degree 2 to pairwise interactions, and so on. The relation $\sqsubseteq$ establishes a hierarchy between these interactions referred to as "Hammersley-Clifford hierarchy".

Let ν be a time-translation invariant (stationary) probability on Ω. The entropy of ν is $\mathcal{S}[\nu] = -\lim_{n\to\infty} \frac{1}{n+1} \sum_{\omega_0^n} \nu[\omega_0^n] \log \nu[\omega_0^n]$. Then, there is a unique stationary probability μ, called *Gibbs distribution with potential* $\mathcal{H}$ satisfying:

$$\mathcal{P}[\mathcal{H}] = \sup_{\nu \in \mathcal{M}} (\mathcal{S}[\nu] + \nu[\mathcal{H}]) = \mathcal{S}[\mu] + \mu[\mathcal{H}], \quad (3)$$

where $\mathcal{M}$ is the set of stationary probabilities on Ω, and $\nu[\mathcal{H}]$ is the average of $\mathcal{H}$ with respect to ν [7]. The quantity $\mathcal{P}[\mathcal{H}]$ is the *free energy*; this is a convex function of h_l's and $\frac{\partial \mathcal{P}[\mathcal{H}]}{\partial h_l} = \mu[m_l]$. The Kullback-Leibler divergence $d_{KL}(\nu, \mu)$ between a stationary probability ν and the Gibbs distribution μ is given by $d_{KL}(\nu, \mu) = \mathcal{P}[\mathcal{H}] - \nu[\mathcal{H}] - \mathcal{S}[\nu]$.

Two distinct potentials $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}$ of range $R = D + 1$ correspond to the same Gibbs distribution (are "equivalent"), if and only if there exists a range D function f such that [7]:

$$\mathcal{H}^{(2)}(\omega_0^D) = \mathcal{H}^{(1)}(\omega_0^D) - f(\omega_0^{D-1}) + f(\omega_1^D) + \Delta, \quad (4)$$

where $\Delta = \mathcal{P}[\mathcal{H}^{(2)}] - \mathcal{P}[\mathcal{H}^{(1)}]$. A prominent example associates a potential $\mathcal{H}$ to a unique equivalent *normalized* potential, the log of a transition probability of a Markov chain with memory depth D

$$\phi(\omega_0^D) = \log P[\omega(D) \,|\, \omega_0^{D-1}]. \qquad (5)$$

So, in this case the function $\mathcal{G}(\omega_0^D) = f(\omega_0^{D-1}) - f(\omega_1^D) - \Delta$ acts as a normalization function. For potentials of range $R = 1$ ($D = 0$), $\mathcal{G} = \log Z$, where Z is the partition function. This corresponds to a memoryless process where spikes occurring at successive times are independent. This is the only case where $\mathcal{G}$ reduces to a constant. For $R > 1$, $\mathcal{G}$ can be computed in terms of the largest eigenvalue and related right eigenvector of a transfer matrix [8, 9]. The free energy of a normalized potential is always 0. If all h_ls in (2) are $> -\infty$ the Markov chain defined by (5) has a *unique* equilibrium state which is precisely the Gibbs distribution μ associated with $\mathcal{H}$ [7]. On the opposite, any Markov chain with strictly positive time-translation invariant transition probabilities has a unique equilibrium state which is a Gibbs distribution [6] with a normalized potential ϕ given by the log of its transition probabilities. However, there are *infinitely many potentials* of the form (2) equivalent to ϕ due to the arbitrariness in the function f in (4).

Among all these potentials there is a canonical one defined as follows. Note that in the decomposition (2) there are monomials redundant with respect to the variational principle (3). Two monomials are redundant if one is mapped to the other by time-translation. Indeed, since we are considering stationary probabilities they have the same average value and correspond to the same constraint in (3). An easy example is given by the monomials $\omega_k(0)$ and $\omega_k(1)$ whose average is the firing rate of neuron k. More generally, these redundant monomials have a mask with 0's on the last column (time D). They can be eliminated in the decomposition (2) by a suitable choice of the function f [11]. A potential where such monomials has been eliminated (the corresponding h_l vanishes) is called *canonical*. A canonical potential contains thus, in general, $2^{NR} - 2^{N(R-1)}$ terms. Now, one can show that two canonical potential are equivalent if and only if their coefficients h_l, $l > 0$, are equal [11]. There is still an arbitrariness due to the term h_0 ("gauge" invariance). One can set it equal 0 without loss of generality. In this way, there is only one canonical potential, with a minimal number of monomials, corresponding to a given stationary Markov chain.

Assume that we are observing a spike train ω that has been generated by a stationary Markov chain with known transition probabilities $P[\omega(D) \,|\, \omega_0^{D-1}] > 0$. The invariant probability of this chain, μ^*, is a Gibbs distribution with canonical potential $\mathcal{H}^*$. Can we infer $\mathcal{H}^*$ from transition probabilities ? This means: (i) Determining the canonical monomials in $\mathcal{H}^*$; (ii) Determining the corresponding coefficients. The MaxEnt addresses (ii) by postulating a form for $\mathcal{H}^*$, i.e. by choosing (ad

hoc) a set χ of monomials. This determines a potential $\mathcal{H}^{(\chi)}$, corresponding to a Gibbs distribution $\mu^{(\chi)}$, which is in general not equal to μ^* (the KL divergence is positive). Contrarily to statistical physics where the shape of $\mathcal{H}$ is made from first principles, in spike train analysis χ is selected ad hoc and corresponds to working hypotheses. However, we now show how this set can be determined from the knowledge of transition probabilities (resp. the normalized potential ϕ) to recover $\mathcal{H}^*$.

The construction is based on periodic orbits of masks. We construct a set $\mathcal{C}$ of periodic orbits, with period τ dividing R, by taking each possible mask $\omega^{(l_1)}$ and shifting it circularly. We note σ the R-circular shift and use the notation $\sigma^n l_1$ to denote the index of the mask $\sigma^n\omega^{(l_1)}$. Remark that $\mathcal{C}$ depends only on N, R but not on ϕ. We note $\mathcal{C}^* = \mathcal{C} \setminus \omega^{(0)}$. Since the sought potential $\mathcal{H}^*$ is equivalent to ϕ, it obeys:

$$\sum_{n=1}^{R} \phi\left(\omega^{(\sigma^n l_1)}\right) = \sum_{n=1}^{R} \mathcal{H}^*\left(\omega^{(\sigma^n l_1)}\right) - R\mathcal{P}\left[\mathcal{H}^*\right], \quad (6)$$

for each periodic orbit $c \in \mathcal{C}$, because the sum of terms involving f in (4) cancel each other along the periodic orbit. This is a particular case of a more general result by Livsic [10], holding for infinite range potentials, stating that two potentials are equivalent if the sum of their value on each possible periodic orbits is equal. In the present setting where potentials have a finite range R, the sought potential $\mathcal{H}^*$ can be constructed step by step by a suitable choice of a finite set of periodic orbits, using the Hammersley-Clifford hierarchy.

Denote ϕ_l (resp. h^*_l) the coefficients of ϕ (resp. $\mathcal{H}^*$) in the decomposition (1). It results from (6) and Hammersley-Clifford decomposition that, for each periodic orbit in $\mathcal{C}$:

$$\sum_{l' \sqsubseteq l} \sum_{n=1}^{R} \phi_{\sigma^n l'} = \sum_{l' \sqsubseteq l} \sum_{n=1}^{R} h^*_{\sigma^n l'} - R\mathcal{P}\left[\mathcal{H}^*\right]. \quad (7)$$

This defines a hierarchy of relations with increasing degree d. For the block of degree $d = 0$ we obtain $\phi_0 \equiv \phi(\omega^{(0)}) = -\mathcal{P}\left[\mathcal{H}^*\right]$, since we have fixed the constant $h^*_0 = 0$. Then, by induction on block degree, for each periodic orbit $c \in \mathcal{C}^*$:

$$\sum_{n=1}^{R} \phi_{\sigma^n l} = \sum_{n=1}^{R} h^*_{\sigma^n l}. \quad (8)$$

The sums $\sum_{n=1}^{R} \phi_{\sigma^n l}$ can be computed by inverting the Hammersley-Clifford relation (1). For a periodic orbit $c \in \mathcal{C}$ denote K_c the number of canonical interactions. c contains $R - K_c$ non canonical interactions whose coefficients can be set to zero in the right hand side of (8). The remaining canonical coefficients are constrained by (8).

Relations (8) are however not sufficient to determine completely $\mathcal{H}^*$, since they provide less equations than sought coefficients. However, they can be completed by additional relations obtained using (6) with periodic orbits $\notin \mathcal{C}$. The idea is then to proceed iteratively. Eq. (8) allows to compute first the coefficients of degree 1. Assume that we want to compute the coefficient h^*_l for a monomial of degree $d > 1$. One then construct a periodic orbit with period $2d$ such that (6) holds, with only one unknown, the coefficient h^*_l, whereas all other terms in the orbit corresponds to already computed terms of degree $< d$ (see [11] for details).

When getting to larger and larger interactions degree the computation might become expensive. However, since $\mathcal{P}\left[\phi\right] = 0$, we have, from (3), $\mathcal{S}\left[\mu\right] = -\sum_{l=0}^{L} \phi_l \mu\left[m_l\right]$. Now, $\omega^{(l_1)} \sqsubseteq \omega^{(l_2)} \Rightarrow \mu\left[m_{l_1}\right] \geq \mu\left[m_{l_2}\right]$. Therefore, the contribution of high degree interactions to entropy becomes rapidly negligible. Thus, one can iterate the computation described above until the entropy contribution of the last step is below some threshold. Using this criterion, one replaces the exact potential $\mathcal{H}^*$ by a truncated potential $\mathcal{H}^{(co)}$ ("cut-off"). We have $d_{KL}(\mu_{\mathcal{H}^*}, \mu_{\mathcal{H}^{(co)}}) = \sum_{l=l_{co}+1}^{L(N,R)} h_l \mu_{\mathcal{H}^*}(m_l)$, which tends to 0 as l_{co} growths.

To illustrate our method we show here how the coefficients shaping the Ising model, used e.g. in [12], are written in terms of transition probabilities of a Markov process with range $R > 1$. The Ising potential only provides an approximation of the equilibrium distribution of the process, corresponding to degree 2 expansion in the Hammersley-Clifford hierarchy, where, besides, memory effects are neglected. Recall that the potential in the Ising model has the form $\sum_{i=1}^{N} \mathtt{h}_i\omega_i(0) + \sum_{i<j} \mathtt{J}_{ij}\omega_i(0)\omega_j(0)$. We keep here the spike description with 0's and 1's but the relation with the classical Ising Hamiltonian with spins $\in \{-1, 1\}$ is straightforward. These coefficients can be directly computed using the relation (7), (8) since $\mathtt{h}_i$'s corresponds to degree 1 interactions, whereas the $\mathtt{J}_{ij}$'s are symmetric. Computing, e.g. the pairwise coefficients of an Ising model with one step of memory (as in [13]) would require to add additional periodic orbits $\notin \mathcal{C}$ (see [11] for an explicit computation).

The coefficient $\mathtt{h}_i$ corresponds to the monomial with mask $\omega^{(l_i)}$ having 0's everywhere but on row i and column D [16]. Taking the periodic orbit $\in \mathcal{C}$ starting from $\omega^{(l_i)}$ and using (7) we obtain [17]:

$$\mathtt{h}_i = \sum_{n=1}^{R} \phi\left(\omega^{(\sigma^n l_i)}\right) - R\phi_0. \quad (9)$$

To compute the $\mathtt{J}_{ij}$'s one considers the mask $\omega^{(l_{ij})}$ with 0's everywhere but on rows i, j in column D, and take its periodic orbit in $\mathcal{C}$. We obtain:

$$\mathtt{J}_{ij} = \sum_{n=1}^{R} \phi\left(\omega^{(\sigma^n l_{ij})}\right) - \mathtt{h}_i - \mathtt{h}_j - R\phi_0. \quad (10)$$

To make these relations more explicit, let us consider transition probabilities of the form (5):

$$P\left[\,\omega(D)\,\middle|\,\omega_0^{D-1}\,\right] = f\left(b + K\cdot x + H(\omega_0^D)\right), \qquad (11)$$

where f is a non linear function (typically, a sigmoid); b is a vector fixing the baseline firing rate of neurons; K is a causal, time-translation invariant, linear convolution kernel that mimics a linear receptive field of neurons; x is a stimulus; H is a memory kernel that describes excitatory or inhibitory post spike effects of pre-synaptic on post-synaptic neurons. As such, it depends on the past spikes. This form is obtained analytically in discrete-time IF model [2] and can also be viewed as a discrete time version of GLM conditional intensities [1, 14].

Plugging this form of transition probabilities into (9), (10), it follows that the "local field" $\mathtt{h}_i$ depends non linearly on the complete stimulus x (not only the stimulus applied to neuron i); this is also a non linear function of the connectivity kernel H. Likewise, the "instantaneous pairwise" interaction depends on the whole connectivity kernel H (not only on the connection between i and j) as well as on the stimulus applied to all neurons in the network. This example clearly shows that there is no straightforward relation between the so-called "functional connectivity" in Ising model (represented by the $\mathtt{J}_{ij}$'s) and the real connectivity. Especially, the graph of $\mathtt{J}_{ij}$'s is in general complete, even if the graph of synaptic weights is not.

More generally, the h_l's of a canonical potential corresponding to transition probabilities of the form (11) are *generically* non zero. Therefore, while there are $O(N^2)$ parameters to constraint the transition probabilities (11), the number of MaxEnt parameters increases exponentially fast with N, the number of neurons. Thus, there is a great amount of redundant information in the h_l's.

As a conclusion, let us make a few remarks and perspectives for future works. First, in all the paper, we have assumed that transition probabilities were exactly known. This corresponds to considering infinite rasters, a not very realistic situation. Actually, our method works as well for finite size rasters, where transition probabilities can be estimated from empirical averages. Here, fluctuations on conditional probabilities estimations obviously affect the reliability of MaxEnt coefficients, depending on their rank in the Hammersley-Clifford hierarchy. This aspect will be discussed in a separate paper. Second, as mentioned in the introduction, the MaxEnt principle heavily relies on the highly questionable assumption of stationarity. Although Gibbs distributions can also be defined for non stationary processes [15], in the context of neural networks [14], they do not obey the MaxEnt principle. Although stationarity can be defended, e.g. in the case of movies presented to a retina [12], in many experiments the spike response to flashed stimuli is considered. In this case, there is no way to defend that spike train statistics is stationary. So, is there any use in studying MaxEnt Gibbs distributions ? An interesting hypothesis would be to assume, in the case of flashed images, that spontaneous dynamics is stationary, and has a MaxEnt Gibbs distributions. Then, the response to a stimuli could be analyzed in the realm of (linear) response theory. From this, analytic form of response kernels ("receptive fields") can be inferred. But the estimation of the response relies on a correct characterization of the spontaneous spike activity and related Gibbs distribution. Our method can be used for this.

Our method opens up new possibilities which allow a better understanding of the role of different neural network topologies and stimulus on spike responses. It is based on general results from probability theory and ergodic theory which are not limited to spike trains. Thus, it could also impact different areas of scientific knowledge where binary time series are considered.

This work was supported by the French ministry of Research and University of Nice (EDSTIC), INRIA, ERC-NERVI number 227747, KEOPS ANR-CONICYT and European Union Project # FP7-269921 (BrainScales), Renvision # 600847 and Mathemacs # FP7-ICT-2011.9.7.

[1] Y. Ahmadian, J. W. Pillow, and L. Paninski, Neural Computation **23**, 46 (2011).
[2] B. Cessac, J. Math. Biol. **62**, 863 (2011).
[3] B. Cessac, J. Math. Neuro. **1** (2011).
[4] Eero P Simoncelli, Liam Paninski and Jonathan Pillow, The Cog Neuro, **50**, (2004).
[5] R. Cofré and B. Cessac, Chaos, Solitons and Fractals **50**, 13 (2013).
[6] J. M. Hammersley and P. Clifford, unpublished (1971)
[7] J. Chazottes and G. Keller, Pressure and Equilibrium States in Ergodic Theory (Springer, 2009).
[8] J. C. Vasquez, O. Marre, A. G. Palacios, M. J. Berry, and B. Cessac, J. Physiol. Paris **106**, 120 (2012).
[9] H. Nasser, O. Marre, and B. Cessac, Journal Of Statistical Mechanics, P03006 (2013).
[10] A. Livšic, Math. USSR- Izvestia pp. 1278-1301 (1972).
[11] B. Cessac and R. Cofré, Research Report RR-8329, INRIA (2013).
[12] E. Schneidman, M. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).
[13] O. Marre, S. El Boustani, Y. Fregnac, and A. Destexhe, Phys. rev. Let. **102**, 138101 (2009).
[14] B. Cessac and R. Cofré, J. Physiol. Paris (2013), in Press.
[15] G. Maillard, Introduction to chains with complete connections, École Fédérale Polytechnique de Lausanne (2007).
[16] This position is arbitrary and does not affect the final result since $\mathtt{h}_i$ is computed using all blocks in the periodic orbit where the 1 goes from time 0 to time D, on the i-th row.
[17] By construction, this orbit contains only one canonical monomial of degree 1, corresponding to $\mathtt{h}_i$.