Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2020, Vol. 56, No. 4, 2377–2405
https://doi.org/10.1214/19-AIHP1042

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES

Sparse space–time models: Concentration inequalities and Lasso

G. Ost[a] and P. Reynaud-Bouret[b]

[a] *Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil. E-mail: guilhermeost@gmail.com*
[b] *Université Côte d'Azur, CNRS, LJAD, Nice, France. E-mail: reynaudb@unice.fr*

Received 5 October 2018; revised 12 August 2019; accepted 19 November 2019

Abstract. Inspired by Kalikow-type decompositions, we introduce a new stochastic model of infinite neuronal networks, for which we establish sharp oracle inequalities for Lasso methods and restricted eigenvalue properties for the associated Gram matrix with high probability. These results hold even if the network is only partially observed. The main argument rely on the fact that concentration inequalities can easily be derived whenever the transition probabilities of the underlying process admit a *sparse space–time representation*.

Résumé. En s'inspirant des décompositions de Kalikow, nous introduisons un nouveau modèle de réseaux neuronaux infinis, pour lesquels nous établissons des inégalités d'oracle précises pour des méthodes Lasso et des propriétés de valeur propre restreinte pour la matrice de Gram associée avec grande probabilité. Ces résultats sont vrais même si le réseau n'est que partiellement observé. L'argument principal est d'établir des inégalités de concentration quand les probabilités de transition sous-jacentes ont une représentation parcimonieuse en temps et espace.

MSC2020 subject classifications: Primary 60G10; secondary 60J99; 62M05

Keywords: Restricted eigenvalue; Chains of infinite order; Perfect simulation; Concentration inequalities; Oracle inequalities; Lasso estimator; Stochastic neuronal networks

1. Introduction

Lasso-type methods in classic regression settings assume that the corresponding Gram matrix G fulfills nice properties such as the Restricted Isometry Property (RIP), Restricted Eigenvalue (RE) conditions, etc. In many works (see for instance van de Geer and Bühlmann [32], Bühlmann and van de Geer [2], van de Geer [31] and references therein), the explanatory variables involved in the Gram matrix are given at first and it is natural to define the regression model conditionally to these variables. In this sense, it is also natural to assume such properties on the Gram matrix G without trying to show that they are fulfilled with high probability. In practice, it is computationally difficult to check whether G satisfies or not these assumptions and many works have shown how random matrices can fulfill such properties with high probability (see for instance Candès and Tao [3], Rudelson and Vershynin [27] or the references in Tropp [29]).

However, in several probabilistic frameworks it is difficult to separate the study of the Gram matrix from the study of the process itself (see for instance Kock and Callot [21] where the probabilistic framework is of auto-regressive kind or Gaïffas and Matulewicz [11] for Markovian continuous processes). Several works have therefore shown that with high probability, Lasso or other adaptive methods satisfy oracle inequalities or minimax results and that, on the same event, the corresponding Gram matrix (which is considered here also as a random variable) satisfies RIP or RE (see for instance Kock and Callot [21], Basu and Michailidis [1], Jiang, Raskutti and Willett [18], Hunt et al. [17]).

We are here interested in a particular type of stochastic process that can model the spiking events of a possibly infinite network of neurons. Many probabilistic models of neuronal activities in a network exist. As examples, let us cite continuous frameworks where both the voltage and the spiking activity of each neuron are modeled (see for instance Sacerdote and Giraudo [28]), or where the spike trains are directly modeled by point processes (see e.g. Chevallier [5]). There are also approaches, closer to the present one, where discrete time is used (see Cofré and Cessac [6], Galves and Löcherbach [14]).

Although many statistical methods have been developed to estimate the probability to spike given the past for various probabilistic models, most of the time one assumes that the network is fully observed (see for instance Pouzat and Chaffiol [25] on Wold processes, Chen et al. [4] on Hawkes processes or Mark, Raskutti and Willett [23] on Poisson counts). Let us underline the work by Hansen, Reynaud-Bouret and Rivoirard [15], which is the closest to ours from a statistical point of view, in which the authors applied a Lasso method on point processes and derived an oracle inequality on an event where the corresponding Gram matrix G is invertible. In a second step, the authors have shown that G is invertible with high probability when the observed process is a linear Hawkes process and the small fixed number of observed neurons correspond in fact to the totality of the network (see also Kelly et al. [20] for an application on real data of Lasso-type methods). However, in practice, data biologists record are much more scarce than a complete recordings of the whole network activity. Most of the time, just few tens of neurons are recorded and they correspond to neurons that are embedded in at least a network of thousands of neurons.

We are aware of only two articles which clearly deal more deeply with the problem of partial observation from a mathematical point of view: Lerasle and Takahashi [22] and Duarte et al. [8]. In Lerasle and Takahashi [22], the authors assume that the configuration describing the neural activity follows a Gibbsian distribution, which does not take dependency in time into account, fact which is of the utmost importance in neuroscience. In Duarte et al. [8], the interaction neighborhood of a given neuron is estimated by assuming that we observe more neurons than this neighborhood even if it is not the totality of network. In practice, the complexity of the algorithm makes it difficult to apply it on large data sets. In these two works, the purpose is clearly to deal with the fact that some neurons (or nodes of the network) are not recorded at all whereas the recordings are complete for the observed neurons. In a slightly different statistical setting, note also the work by Mark, Raskutti and Willett [24], which deals with another kind of missing data, the applied filter being i.i.d. in both time and nodes but without nodes that are completely unrecorded.

The aim of the present work is to show that with high probability, the Gram matrix G satisfies nice properties such as invertibility or RE condition with as few assumptions as possible on the underlying probabilistic models and this even if we observe only a small number of neurons embedded in an infinite neuronal network, for which most of the neurons are not observed at all.

Inspired by Galves and Löcherbach [13] model and Kalikow-type decompositions (see Kalikow [19], Galves et al. [12]), we consider discrete time models for which the probability of a neuron to spike at a given time unit given the past configuration may only depend on a few neurons (assimilated to space positions) and few time steps. Hence the dependencies in time and space should be very small but may be random and chosen at each step. We use this *probabilistic sparsity* in time and space to prove concentration inequalities for various functionals including Gram matrices. We employ these concentration inequalities to prove that RE properties are satisfied with large probability even on a partially observed infinite network. Therefore we show that Lasso methods have good theoretical properties even in this case.

The paper is organized as follows. In Section 2, we provide the main notation and present briefly the stochastic framework with its main assumptions. In Section 3, we prove an oracle inequality for the Lasso estimator of the transition probabilities of this model. The oracle inequality is derived on a certain event on which some properties of the Gram matrix are met. Examples of useful dictionaries are also presented in this section. We discuss, in Section 4, the definition of space–time decomposition and show examples of discrete time models where it applies. Besides, thanks to the definition of a simulation algorithm inspired by Galves and Löcherbach [13], we prove that stochastic models admitting a space–time decomposition have a stationary version even on infinite networks under some conditions of *probabilistic sparsity*. We prove that these conditions are usually much less stringent than the ones of the literature. Still in the same section, we obtain concentration inequalities for such processes, under some additional exponential constraints, by adapting arguments of Viennet [33]. This allows us to prove that the introduced Gram matrices based on a partial observation of the network are invertible or satisfy RE with high probability. This is done in Section 5. A brief conclusion is given in Section 6. All proofs are given in the Section 7.

2. Stochastic framework and notation

We write $\mathbb{N}$ to denote the set of natural numbers $\{0, 1, 2, \ldots\}$. The set of integers $\{\ldots, -1, 0, 1, \ldots\}$ is denoted by $\mathbb{Z}$. The set of strictly negative and positive integers are denoted by $\mathbb{Z}_-$ and $\mathbb{Z}_+$ respectively.

We consider a stationary stochastic chain $\mathbf{X} = (X_{i,t})_{i\in I, t\in\mathbb{Z}}$ taking values in $\{0,1\}^{I\times\mathbb{Z}}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where I is a countable (possibly infinite) set. The set I represents the set of neurons in the network. For each $i \in I$ and $t \in \mathbb{Z}$,

$$X_{i,t} = \begin{cases} 1 & \text{if neuron } i \text{ spikes at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

The configuration of $\mathbf{X}$ at time $t \in \mathbb{Z}$ is denoted by $X_t = (X_{i,t}, i \in I)$. For $s, t \in \mathbb{Z}$ with $s < t$, $X_{i,s:t}$ stands for the collection $(X_{i,s}, \dots, X_{i,t})$ and $X_{s:t}$ for the collection $(X_{i,r})_{i\in I, s\le r\le t}$. For each $t \in \mathbb{Z}$, $X_{-\infty:t}$ denotes the past history $(\dots, X_{t-1}, X_t)$ of $\mathbf{X}$ at time $t+1$. Note that the past histories have space–time components. For $F \subset I$ and $t \in \mathbb{Z}$, $X_{F,t} = (X_{i,t}, i \in F)$ denotes the configuration of $\mathbf{X}$ at time t restricted to set F. More generally, for any subset $v \subset I \times \mathbb{Z}$, X_v denotes the collection $(X_{i,t})_{(i,t)\in v}$.

Assumption 1. For each $t \in \mathbb{Z}$, given the past history $X_{-\infty:t}$, the neurons spike independently of each other at time $t+1$, i.e., for any finite set $J \subset I$ and $(a_i)_{i\in J} \in \{0,1\}^J$,

$$\mathbb{P}\Big(\bigcap_{i\in J}\{X_{i,t+1} = a_i\} | X_{-\infty:t} = x\Big) = \prod_{i\in J}\mathbb{P}(X_{i,t+1} = a_i | X_{-\infty:t} = x), \quad \mathbb{P}\text{-a.e. } x \in \{0,1\}^{I\times\mathbb{Z}_-}.$$

Since the stochastic chain $\mathbf{X}$ is stationary, Assumption 1 implies that the dynamics of $\mathbf{X}$ is fully characterized by the transition probabilities

$$p_i(x) = \mathbb{P}(X_{i,0} = 1 | X_{-\infty:-1} = x), \quad x \in \{0,1\}^{I\times\mathbb{Z}_-}, i \in I.$$

These transition probabilities are all assumed to be measurable functions of $x \in \{0,1\}^{I\times\mathbb{Z}_-}$.

Hereafter, we need the following notation. For any neighborhood $v \subset I \times \mathbb{Z}_-$ and $x, y \in \{0,1\}^{I\times\mathbb{Z}_-}$, we write $x \overset{v}{=} y$ to indicate $y_v = x_v$. For any real-valued function f on $\{0,1\}^{I\times\mathbb{Z}_-}$ and subset $v \subset I \times \mathbb{Z}_-$, we say f is *cylindrical* in v and write $f(x) = f(x_v)$, if $f(x) = f(y)$ for any $x, y \in \{0,1\}^{I\times\mathbb{Z}_-}$ such that $x \overset{v}{=} y$.

Let us now present briefly the three other main probabilistic assumptions that we may require in the article. They will be discussed in more details in the sequel.

We denote by $\mathcal{V}$ the collection of finite neighborhoods, i.e. finite subsets of $I \times \mathbb{Z}_-$ and we consider processes for which the following decomposition holds.

Assumption 2 (Space–time decomposition). For all v in $\mathcal{V}$ and i in I, there exists a $[0,1]$-valued measurable function $p_i^v(\cdot)$, cylindrical in v, and a non negative weight $\lambda_i(v)$, such that for all $x \in \{0,1\}^{I\times\mathbb{Z}_-}$ and $i \in I$,

$$\begin{cases} p_i(x) = \lambda_i(\varnothing) p_i^{\varnothing}(x) + \sum_{v\in\mathcal{V}, v\neq\varnothing} \lambda_i(v) p_i^v(x), \\ \sum_{v\in\mathcal{V}} \lambda_i(v) = 1. \end{cases}$$

As we will see in Section 4 where we discuss in details the implication of this assumption, this Kalikow-like assumption is actually met by a wide variety of stochastic processes modeling neuronal networks.

For any neighborhood $v \in \mathcal{V}$, let $T(v) = -\min(s \in \mathbb{Z}_-, (j,s) \in v)$ be the corresponding time length, with the convention that $T(\varnothing) = 0$. We also assume the following.

Assumption 3. There exists a strictly positive θ such that for all i,

$$\varphi_i(\theta) = \sum_{v\in\mathcal{V}} |v| e^{\theta T(v)} \lambda_i(v)$$

is finite and

$$\varphi(\theta) = \sup_{i\in I} \varphi_i(\theta) < 1. \tag{2.1}$$

This assumption, which can be seen as *probabilistic sparsity* (see Section 4), means that the neighborhoods used in the previous decomposition are in average nor too large neither too timely spread.

Finally, we will use also the following assumption to control the Gram matrices in Section 5.

Assumption 4. There exists some positive μ such that for all $i \in I$ and for all $x \in \{0,1\}^{I\times\mathbb{Z}_-}$,

$$\mu \le p_i(x) \le 1 - \mu.$$

This means that whatever the past, there is always enough randomness in the system.

Note that all these 4 assumptions are very general and are there to control the randomness of the underlying neuronal system without specifying any parametric model. In particular in what follows, the p_i's are just approximated by a finite combination of elements of dictionaries, without being one (see Section 3): the overall purpose is to approximate p_i by what happens in a finite subset of observed neurons, namely F, always keeping in mind that F is embedded in a much more complex and potentially infinite network I, the overall complex stochastic interactions inside I being governed by Assumptions 1, 2, 3 and/or 4 depending on the results.

3. Lasso method and statistical notation

For a finite $F \subset I$, subset of observed neurons, and integers $T > m \geq 1$ measuring the observation window, the aim is to estimate $x \mapsto p_i(x)$ for a fixed neuron $i \in F$, given the sample $X_{F,-(m-1):T}$. To that end, for each time $1 \leq t \leq T$, we compare the past $X_{F,(t-m):(t-1)}$ to the current observation $X_{i,t}$ to guess what can be a good approximation of $p_i(x)$. The intuition behind this strategy is that for a well-chosen space–time neighborhood $v \subset I \times \mathbb{Z}_-$, it might be sufficient to know x_v and not the whole past configuration x to well approximate $p_i(x)$.

Given the sample $X_{F,-(m-1):T}$, one might consider several candidates to approximate $p_i(x)$. Here, we shall approximate $p_i(x)$ by linear combinations of a given dictionary Φ, i.e. a finite set of real-valued functions on $\{0,1\}^{I\times\mathbb{Z}_-}$ which are cylindrical in $F \times \underline{m}$ with $\underline{m} = \{-m, \dots, -1\}$. More precisely, for each vector $a = (a_\varphi)_{\varphi\in\Phi} \in \mathbb{R}^\Phi$, we denote

$$x \mapsto f_a(x) = \sum_{\varphi\in\Phi} a_\varphi \varphi(x), \tag{3.1}$$

the candidate encoded by the vector a that should approximate $p_i(x)$. We assume that the functions in the dictionary are bounded in infinite norm by $\|\Phi\|_\infty$.

We use the least-squares contrast defined by

$$C(f_a) = -\frac{2}{T}\sum_{t=1}^{T} f_a(X_{F,(t-m):(t-1)})X_{i,t} + \frac{1}{T}\sum_{t=1}^{T} f_a^2(X_{F,(t-m):(t-1)}), \quad a = (a_\varphi)_{\varphi\in\Phi} \in \mathbb{R}^\Phi.$$

Observe that if for real-valued functions f and g on $\{0,1\}^{I\times\mathbb{Z}_-}$, we denote $\langle f, g\rangle_T = \frac{1}{T}\sum_{t=0}^{T-1} f(X_{-\infty:t})g(X_{-\infty:t})$ and $\|f\|_T$ the corresponding norm, then one has

$$C(f_a) = -2\langle f_a, X_{i,\cdot+1}\rangle_T + \|f_a\|_T^2 = \|f_a - X_{i,\cdot+1}\|_T^2 - \|X_{i,\cdot+1}\|_T^2,$$

and C is minimum when $\|f_a - X_{i,\cdot+1}\|_T^2$ is minimum. In this sense, minimizing C over functions that are only depending on the past might give a good estimator of $p_i(x)$.

Notice also that, if for $\varphi, \varphi' \in \Phi$ we write,

$$b_\varphi = \frac{1}{T}\sum_{t=1}^{T} \varphi(X_{F,(t-m):(t-1)})X_{i,t} \quad \text{and} \quad G_{\varphi,\varphi'} = \frac{1}{T}\sum_{t=1}^{T} \varphi(X_{F,(t-m):(t-1)})\varphi'(X_{F,(t-m):(t-1)}), \tag{3.2}$$

then $C(f_a)$ can be rewritten as

$$-2a^\intercal b + a^\intercal G a,$$

where $b = (b_\varphi, \varphi \in \Phi)$ is a vector of $\mathbb{R}^\Phi$, $G = (G_{\varphi,\varphi'})_{\varphi,\varphi'\in\Phi}$ is the Gram matrix and $a^\intercal$ is the transpose of vector a.

In the sequel, let $|a| = (|a_\varphi|, \varphi \in \Phi)$, $|a|_\infty = \max_{\varphi\in\Phi} |a_\varphi|$, $\|a\| = \sqrt{a^\intercal a}$ and $|a|_1 = \mathbf{1}^\intercal |a|$ where $\mathbf{1}$ is the vector with all coordinates equal to 1.

Following Hansen, Reynaud-Bouret and Rivoirard [15], we minimize $C(f_a)$ subject to a ℓ_1-penalization on the vector $a = (a_\varphi, \varphi \in \Phi)$. Precisely, we choose the function $\hat{f} = f_{\hat{a}}$ where

$$\hat{a} \in \arg\min_{a\in\mathbb{R}^\Phi} \big\{-2a^\intercal b + a^\intercal G a + \gamma d|a|\big\}, \tag{3.3}$$

for d a positive term controlling the random fluctuations and $\gamma > 0$, a tuning parameter.

The active set $S(a)$ of a vector $a \in \mathbb{R}^\Phi$ is the set $S(a) = \{\varphi : a_\varphi \neq 0\}$. We shall denote for any subset $J \subset \Phi$ and any $a \in \mathbb{R}^\Phi$, $a_J \in \mathbb{R}^\Phi$ the vector whose coordinates in J are equal to the ones of a and 0 anywhere else. We also denote by $|J|$ the cardinality of J.

For later use, let us write for each $\varphi \in \Phi$,

$$\bar{b}_\varphi = \frac{1}{T}\sum_{t=0}^{T-1} \varphi(X_{F,t-m:t})\, p_i(X_{-\infty:t}). \tag{3.4}$$

3.1. *Examples of dictionaries*

Let us present briefly some examples of dictionaries that might be useful.

Short memory effect. Let the dictionary Φ be defined by the set $\{\varphi_j : j \in F\}$ where

$$\varphi_j(x) = \begin{cases} 1 & \text{if } x_{j,s} = 1 \text{ for some } -m \le s \le -1, \\ 0 & \text{otherwise,} \end{cases} \qquad x \in \{0,1\}^{I\times\mathbb{Z}_-},$$

so that we are trying to explain the presence of a spike on neuron i at time t by a linear combination of the presence of a spike on neuron j in a small window just before time t.

Cumulative effect. We can also think that $m = \eta L$ is a much larger parameter and cut the past $\underline{m}$ into L small pieces of length η, where the effect of the spikes are different and cumulative. This leads to the dictionary Φ defined by the set $\{\varphi_{j,\ell} : j \in F \text{ and } 1 \le \ell \le L\}$ where

$$\varphi_{j,\ell}(x) = \sum_{s=-\eta\ell}^{-\eta(\ell-1)-1} x_{j,s}, \quad x \in \{0,1\}^{I\times\mathbb{Z}_-}.$$

Cumulative effect with spontaneous apparition. It can be important to take into account a background activity, especially to explain the apparition of spikes due to the unobserved part of the network. To do so, we may add to the previous dictionary an extra function

$$\varphi_0 = 1,$$

whose corresponding coefficient corresponds to a spontaneous activity.

Hawkes dictionary. In both the cumulative effect and the cumulative effect with spontaneous part, one might be interested in a particular example where $\eta = 1$ and $L = m$. In particular, in the case with spontaneous part, we are therefore interested in approximating $p_i(x)$ by

$$f_a(x) = a_0 + \sum_{j\in F}\sum_{-m\le s\le -1} a_{j,-s} x_{j,s}, \quad x \in \{0,1\}^{I\times\mathbb{Z}_-},$$

which is the exact form of a discrete Hawkes process restricted to $F \times \underline{m}$ (see Section 4) and this even if p_i is not of this shape.

Note that whatever the dictionary, m represents the maximal delay in the dictionary and $|F|$ the number of observed neurons. As we will see in Section 5, both these quantities have to be usually less than a certain increasing function of T, which depends on the dictionary (typically $\log(T)$), to derive an RE property on the Gram matrix. In particular $|m|$ might grow slightly with T to ensure a good asymptotic approximation of the dependency in time. Mathematically speaking, the same holds for $|F|$, although the size of F is dictated by the neurophysiological experiment and, for practical prupose, it is always thought to be a constant with respect to T.

3.2. *Oracle inequality*

It is classical, by now, to derive oracle inequalities for Lasso procedures if G satisfies some properties. We use two of them.

Definition 1. Let $\kappa > 0$. The matrix G satisfies Property **Inv**(κ) if

$$\forall a \in \mathbb{R}^\Phi, \quad a^\intercal G a \ge \kappa \|a\|^2.$$

A weaker version is the restricted eigenvalue condition.

Definition 2. Let $c > 0$, $\kappa > 0$ and $s \in \mathbb{N}$. The matrix G satisfies Property **RE**(κ, c, s) if for all subset J such that $|J| \leq s$ and for all $a \in \mathbb{R}^{\Phi}$ such that

$$|a_{J^c}|_1 \leq c|a_J|_1,$$

the following holds

$$a^{\intercal} G a \geq \kappa \|a_J\|^2.$$

Our first result establishes an oracle inequality for the estimator $\hat{f} = f_{\hat{a}}$ where $\hat{a}$ is defined by (3.3).

Theorem 1. *Let $\gamma \geq 2$, $\kappa > 0$ and $s \in \mathbb{N}$. On the event on which*

(i) *for all $\varphi \in \Phi$, $|b_\varphi - \bar{b}_\varphi| \leq d$,*
(ii) *and G satisfies* **RE**$(\kappa, c(\gamma), s)$ *with* $c(\gamma) = \frac{\gamma+2}{\gamma-2}$,

the following inequality is satisfied

$$\left\| \hat{f} - p_i(\cdot) \right\|_T^2 \leq \inf_{a \in \mathbb{R}^{\Phi} : |S(a)| \leq s} \left\{ \left\| f_a - p_i(\cdot) \right\|_T^2 + \kappa^{-1} |S(a)| d^2 \frac{(\gamma+2)^2}{4} \right\}. \tag{3.5}$$

Moreover for any $0 < \delta < 1$, if $d = d_\delta$ with

$$d_\delta = \sqrt{\|\Phi\|_\infty^2 \frac{\log|\Phi| + \log(2\delta^{-1})}{2T}},$$

then $\mathbb{P}(\exists \varphi \in \Phi : |b_\varphi - \bar{b}_\varphi| > d) \leq \delta$.

Equation (3.5) is a classical oracle inequality (see for instance Hansen, Reynaud-Bouret and Rivoirard [15] or Hunt et al. [17] for close set-ups). This result means that the Lasso estimator gives the best s-sparse approximation of p_i based on the dictionary Φ and that the price to pay is of the order of $\kappa^{-1} s d^2$, if we assume that $\|\Phi\|_\infty \leq 1$. With the choice $d = d_\delta$, we have therefore a price of the order of $\kappa^{-1} s \frac{\log|\Phi| + \log(2\delta^{-1})}{T}$. Note that if we knew that p_i can be indeed decomposed on Φ, meaning that the model is true and that in particular p_i only depends on s elements of the dictionary Φ, the price to pay anyway to estimate p_i would be roughly of the order of s/T. Therefore if the logarithmic factor is a classical loss for adaptation in (3.5), it remains to see the order of κ, to see if (3.5) gives roughly the best possible rate.

Note that if G satisfies **Inv**(κ) then one can choose $\gamma = 2$ and $s = |\Phi|$ in Theorem 1 and (3.5) can be rewritten as

$$\left\| \hat{f} - p_i(\cdot) \right\|_T^2 \leq \inf_{a \in \mathbb{R}^{\Phi}} \left\{ \left\| f_a - p_i(\cdot) \right\|_T^2 + 4\kappa^{-1} |S(a)| d^2 \right\},$$

which is a sharper version of the result proved in Hansen, Reynaud-Bouret and Rivoirard [15] in continuous time, up to the fact that they used more general weights which leads to a weighted ℓ_1 norm in the criterion. The same refinement would have been possible but since the focus is here on the Gram matrix, we have decided to use a classical ℓ_1 norm for sake of simplicity.

Note also that another (very easy) refinement consists in clipping $\hat{f}$ to ensure that it remains between 0 and 1. The same result holds for this clipped version. Another way to find similar results for estimators that are forced to be in $[0, 1]$ is to use penalized maximum likelihood. Many works have used it (see for instance Mark, Raskutti and Willett [23] for Poisson counts or Basu and Michailidis [1], Gaïffas and Matulewicz [11] in Gaussian Markovian set-ups). This comes with additional technicalities, in particular if the likelihood of the statistical model is not easy to compute, because the model is not Gaussian. In particular, Mark, Raskutti and Willett [24] use a setting very close to ours, but simpler (see also the examples in Section 4) and make use of Taylor expansion to approximate the criteria. Translated here, the approximation would depend on the dictionary we use and would be more complex for each dictionary. Once again, because the focus is here on the Gram matrix, we have decided to stick with the simplest Lasso result made for least-squares contrast.

Results for controlling Gram matrices are numerous (see for instance Basu and Michailidis [1], Gaïffas and Matulewicz [11], Hunt et al. [17] for simpler settings than the present one) but always assume that the whole network is observed

and that *the target can be written on the dictionary*. In Hansen, Reynaud-Bouret and Rivoirard [15], which is the closest framework to the present one, it has been proved for instance that, if one observes the whole finite network and if the spike trains are linear Hawkes processes, then G is invertible with large probability for well chosen dictionaries. In this case, the corresponding κ is roughly lower bounded by a quantity which is exponentially small in the total number of neurons in the network. Here we would like to go beyond these assumptions and prove that even if

- the model is wrong (i.e. p_i is not Hawkes for instance) and p_i cannot be written on the dictionary,
- the network is infinite,
- we only observe a very partial subnetwork,

it is still possible to find good κ with high probability and that the dependency in the number of neurons can be much better than these previous results.

The main idea consists in using very general Kalikow-type decomposition of the transition probabilty $p_i(x)$, that are available in discrete time (see Galves et al. [12]) and that do not exist with such generality in continuous time (see however Hodara and Löcherbach [16] for promising results in this direction).

4. Space–time decomposition and concentration

4.1. *Definition*

Let us first recall Assumption 2: For all v in $\mathcal{V}$ and i in I, there exists a $[0,1]$-valued measurable function $p_i^v(\cdot)$, cylindrical in v, and a non negative weight $\lambda_i(v)$, such that for all $x \in \{0,1\}^{I\times\mathbb{Z}_-}$ and $i \in I$,

$$\begin{cases} p_i(x) = \lambda_i(\varnothing)p_i^{\varnothing}(x) + \sum_{v\in\mathcal{V}, v\neq\varnothing} \lambda_i(v)p_i^v(x), \\ \sum_{v\in\mathcal{V}} \lambda_i(v) = 1. \end{cases}$$

The aforementioned decomposition can be interpreted as follows. At each time step, to decide whether the neuron i spikes or not, we first choose a random space–time neighborhood $V \in \mathcal{V}$ according to the distribution λ_i. Once this neighborhood V is chosen, we decide if a spike of neuron i occurs with probability $p_i^V(x_V)$ that depends only on the past history restricted to V. Note that $p_i^{\varnothing}(x)$ does not depend on x at all, and we denote this value $p_i^{\varnothing}$.

Such a space–time decomposition of the transition probabilities $\{p_i(x), i \in I, x \in \{0,1\}^{I\times\mathbb{Z}_-}\}$ generalizes the classical Kalikow decomposition introduced in Kalikow [19] and further developed in Comets, Fernandez and Ferrari [7], Galves et al. [12] and Galves and Löcherbach [13]. The main difference consists in not forcing the nesting of the neighborhoods v that lie in the support of λ_i. This helps us to exploit the fact that in many cases the distributions λ_i charge very few neighborhoods and that the cardinality of this neighborhood is usually very small, if the nesting is not forced. We speak in this case of *probabilistic sparsity*.

Remark 1. If we denote $q_i(x) = \mathbb{P}(X_{i,0} = 0 | X_{-\infty:-1} = x) = 1 - p_i(x)$ and $q_i^v(x) = 1 - p_i^v(x)$ for all $i \in I$, $x \in \{0,1\}^{I\times\mathbb{Z}_-}$, we can also write

$$q_i(x) = \lambda_i(\varnothing)q_i^{\varnothing}(x) + \sum_{v\in\mathcal{V}, v\neq\varnothing} \lambda_i(v)q_i^v(x),$$

where for each $v \in \mathcal{V}$, the function q_i^v is cylindrical in v.

Remark 2. For a given space–time decomposition, one can use Remark 1 to deduce that for all $i \in I$,

$$\inf_{x\in\{0,1\}^{I\times\mathbb{Z}_-}} p_i(x) + \inf_{x\in\{0,1\}^{I\times\mathbb{Z}_-}} q_i(x) \geq \lambda_i(\varnothing).$$

More generally, for any $v \in \mathcal{V}$, one can show that

$$\inf_{x\in\{0,1\}^{I\times\mathbb{Z}_-}} \left\{ \inf_{y\in\{0,1\}^{I\times\mathbb{Z}_-} : y\overset{v}{=}x} p_i(y) + \inf_{y\in\{0,1\}^{I\times\mathbb{Z}_-} : y\overset{v}{=}x} q_i(y) \right\} \geq \lambda_i(\varnothing) + \sum_{w\subseteq v, w\neq\varnothing} \lambda_i(w).$$

One can also show that the space–time decomposition is not unique. This fact raises the question of whether there is an "optimal" decomposition of a given transition probability. Such a question, however, is not discussed in this article.

4.2. *Main examples*

4.2.1. *Markov chains*

Suppose I is a singleton, say $I = \{1\}$, and denote X_t instead of $X_{1,t}$ for convenience. Let us assume also that for all $x \in \{0,1\}^{\mathbb{Z}_-}$,

$$p(x) = \mathbb{P}(X_0 = 1 | X_{-\infty:-1} = x) = \mathbb{P}(X_0 = 1 | X_{-1} = x_{-1}),$$

that is, the stochastic chain $\mathbf{X} = (X_t)_{t\in\mathbb{Z}}$ is a Markov chain of order 1 taking values in $\{0,1\}$. To shorten notation, let

$$p^1 = \mathbb{P}(X_t = 1 | X_{t-1} = 1), \qquad q^1 = 1 - p^1, \qquad p^0 = \mathbb{P}(X_t = 1 | X_{t-1} = 0) \quad \text{and} \quad q^0 = 1 - p^0.$$

We can always write the transition probability $p(x)$ as

$$p(x) = p^1 x_{-1} + p^0(1 - x_{-1}).$$

Let us denote $p = p^1 \wedge p^0$, $q = q^1 \wedge q^0$ and $\mu = p + q$. If $0 < \mu < 1$, then one can write

$$p(x) = \mu \frac{p}{\mu} + (1-\mu)\left[\frac{p^1 - p}{1-\mu} x_{-1} + \frac{p^0 - p}{1-\mu}(1 - x_{-1})\right].$$

So one can use as space–time decomposition

$$\begin{cases} \lambda(\varnothing) = \mu, \\ p^{\varnothing} = \frac{p}{\mu}, \\ \lambda(\{-1\}) = 1 - \mu, \\ p^{\{-1\}}(x) = \frac{p^1 - p}{1-\mu} x_{-1} + \frac{p^0 - p}{1-\mu}(1 - x_{-1}). \end{cases} \tag{4.1}$$

Note that the support of λ is then reduced to $\{\varnothing, \{-1\}\}$. Note also than it is quite straightforward to extend this to the multidimensional models considered in Mark, Raskutti and Willett [24], for instance by saying that when the neighborhood is not empty, one picks I which is finite in their case.

4.2.2. *Chains of infinite order*

Again suppose I is a singleton. In this case, the stochastic chain $\mathbf{X}$ is described by the transition probability $\{p(x), x \in \{0,1\}^{\mathbb{Z}_-}\}$. Denote for $\ell \in \mathbb{Z}_+$, $\underline{\ell}$ the set $\{-\ell, \dots, -1\}$ and

$$\beta_\ell = \sup_{x\in\{0,1\}^{\mathbb{Z}_-}} \sup_{\substack{y,z\in\{0,1\}^{\mathbb{Z}_-} \text{ s.t.} \\ y \overset{\ell}{=} z \overset{\ell}{=} x}} \big\{ \big| p(y) - p(z) \big| \big\}.$$

If there exist $\ell_0 \geq 1$ such that $\beta_\ell = 0$ for all $\ell \geq \ell_0$, then the stochastic chain $\mathbf{X}$ is called Markov Chain of Order ℓ_0. Otherwise, $\mathbf{X}$ is called Chain of Infinite Order. We refer the reader to Fernandéz, Ferrari and Galves [9] for a comprehensive introduction to Chains of Infinite Order.

If $\beta_\ell \to 0$ as $\ell \to \infty$, then $\{p(x), x \in \{0,1\}^{\mathbb{Z}_-}\}$ is said to be *continuous* and the sequence $(\beta_\ell)_{\ell\in\mathbb{Z}_+}$ is called the *continuity rate*. Recall that $q(x) = 1 - p(x)$ for all $x \in \{0,1\}^{\mathbb{Z}_-}$. One can then compute for $\ell \in \mathbb{Z}_+$,

$$\alpha(\ell) = \inf_{x\in\{0,1\}^{\mathbb{Z}_-}} \Big\{ \inf_{y\in\{0,1\}^{\mathbb{Z}_-} \text{ s.t. } y \overset{\ell}{=} x} p(y) + \inf_{y\in\{0,1\}^{\mathbb{Z}_-} \text{ s.t. } y \overset{\ell}{=} x} q(y) \Big\}.$$

This allows us to define the distribution λ which has support only on the $\underline{\ell}$'s and

$$\lambda(\underline{\ell}) = \alpha(\ell) - \alpha(\ell - 1), \tag{4.2}$$

where $\alpha(0) = \lambda(\varnothing) = \inf_{x\in\{0,1\}^{\mathbb{Z}_-}} p(x) + \inf_{x\in\{0,1\}^{\mathbb{Z}_-}} q(x)$. One can show that every continuous transition probability $\{p(x), x \in \{0,1\}^{\mathbb{Z}_-}\}$ admits a decomposition of the form:

$$\begin{cases} p(x) = \lambda(\varnothing) p^{\varnothing} + \sum_{\ell\in\mathbb{Z}_+} \lambda(\underline{\ell}) p^{\underline{\ell}}(x), \\ \lambda(\varnothing) + \sum_{\ell\in\mathbb{Z}_+} \lambda(\underline{\ell}) = 1. \end{cases} \tag{4.3}$$

Moreover (4.3) is a space–time decomposition since $p^{\varnothing} \in [0,1]$ and for each $\ell \in \mathbb{Z}_+$, $\{p^{\underline{\ell}}(x), x \in \{0,1\}^{\mathbb{Z}_-}\}$ is a transition probability of a Markov chain of order ℓ.

4.2.3. *Discrete-time linear Hawkes processes*

Multivariate Hawkes processes (also referred in the neuroscience literature as a particular case of generalized linear models) are often used to model interacting spike trains and especially the synaptic integration. In contrast to the classical framework where these processes are described in continuous time and are not linear, we focus here on a discrete-time and linear formulation, where for $i \in I$:

$$\psi_i(x) = \nu_i + \sum_{s\in\mathbb{Z}_-}\sum_{j\in I} h_{j\to i}(-s)x_{j,s} \quad \text{and} \quad \begin{cases} p_i(x) = \psi_i(x) & \text{if } \psi_i(x) \in [0,1], \\ p_i(x) = 1 & \text{if } \psi_i(x) > 1, \\ p_i(x) = 0 & \text{if } \psi_i(x) < 0. \end{cases} \tag{4.4}$$

In this formula, $\nu_i \geq 0$ represents the spontaneous activity of neuron i, that is its ability to produce spikes when there is no interaction. The interaction function $h_{j\to i}$ measures the amount of excitation (if positive) or inhibition (if negative) that a spike of neuron j has on neuron i after a delay $-s$ (a spike of neuron j with delay $-s$ corresponds to $x_{j,s} = 1$).

For a given neuron $i \in I$, we write

$$A_i^+ = \big\{(j,s) \in I \times \mathbb{Z}_- : h_{j\to i}(-s) > 0\big\} \quad \text{and} \quad A_i^- = \big\{(j,s) \in I \times \mathbb{Z}_- : h_{j\to i}(-s) < 0\big\},$$

and define the maximal excitatory (respectively inhibitory) strength by

$$\Sigma_i^+ = \sum_{(j,s)\in A_i^+} \big|h_{j\to i}(-s)\big| \quad \text{and} \quad \Sigma_i^- = \sum_{(j,s)\in A_i^-} \big|h_{j\to i}(-s)\big|.$$

Let us assume that

$$0 \leq \nu_i - \Sigma_i^- \quad \text{and} \quad \nu_i + \Sigma_i^+ \leq 1, \tag{4.5}$$

which implies in particular that whatever the past configuration $x \in \{0,1\}^{I\times\mathbb{Z}_-}$, the transition probability $p_i(x) \in [0,1]$ is always equal to $\psi_i(x)$. It also implies that $\Sigma_i^+ + \Sigma_i^- \in [0,1]$.

Then one can use for the space–time decomposition:

$$\begin{cases} \lambda_i(\varnothing) = 1 - (\Sigma_i^+ + \Sigma_i^-) & \text{which is } \geq 0 \text{ since } 0 \leq \Sigma_i^+ + \Sigma_i^- \leq 1, \\ p_i^{\varnothing} = \frac{\nu_i - \Sigma_i^-}{\lambda_i(\varnothing)} & \text{which is } \leq 1 \text{ since } \nu_i + \Sigma_i^+ \leq 1, \\ \lambda_i(\{(j,s)\}) = |h_{j\to i}(-s)| & \text{for all } (j,s) \in A_i^+ \cup A_i^-, \\ p_i^{\{(j,s)\}}(x) = x_{j,s} & \text{for all } (j,s) \in A_i^+, \\ p_i^{\{(j,s)\}}(x) = (1 - x_{j,s}) & \text{for all } (j,s) \in A_i^-. \end{cases} \tag{4.6}$$

It is moreover sufficient to assume that $\Sigma_i^+ + \Sigma_i^- < 1$ to have $\lambda_i(\varnothing) > 0$.

The discrete-time linear Hawkes model is an interesting example, because even if the true interaction graph, that is the set of edges $(j,i) \in I \times I$ for which $h_{j\to i}$ is non zero, is complete, the random neighborhoods $V \in \mathcal{V}$ of the space–time decomposition have cardinality at most 1 almost surely. This *probabilistic sparsity* is exploited in the sequel to obtain concentration inequalities.

Note that it is classically assumed for general Hawkes models that the spectral radius of the matrix $(\int |h_{j\to i}|)_{i,j}$ is smaller than 1 to ensure stationnarity of the whole multivariate process. Here this matrix can be reinterpreted as $H = (\sum_{\ell>1} |h_{j\to i}(\ell)|)_{i,j}$. Therefore (4.5) is different from the usual assumption: it implies in particular that

$$\Sigma^+ + \Sigma^- = H\mathbf{1} \leq \mathbf{1} \quad \text{coordinate per coordinate,}$$

where $\Sigma^+ = (\Sigma_i^+)_i$, $\Sigma^+ = (\Sigma_i^+)_i$ and $\mathbf{1}$ is the vector of 1's.

4.2.4. *GL neuron model*

Let $W_{j\to i} \in \mathbb{R}$ with $i, j \in I$, be a collection of real numbers such that $W_{j\to j} = 0$ for all j. For each $i \in I$, let $\varphi_i : \mathbb{R} \to [0,1]$ be a non-decreasing measurable function and $g_i = (g_i(\ell))_{\ell\in\mathbb{Z}_+}$ be a sequence of strictly positive real numbers. Here, $W_{j\to i}$ is interpreted as the *synaptic weight of neuron j on neuron i*, φ_i as the *spike rate function* of neuron i and g_i as the *postsynaptic current pulse* of neuron i.

For each $x \in \{0,1\}^{I\times\mathbb{Z}_-}$ and $i \in I$, we define $L^i(x) = \sup\{s \in \mathbb{Z}_- : x_{i,s} = 1\}$. The stochastic chain $\mathbf{X}$ satisfies a GL neuron model if the transition probabilities $\{p_i(x), i \in I, x \in \{0,1\}^{I\times\mathbb{Z}_-}\}$ are given by (see Galves and Löcherbach [13])

$$p_i(x) = \begin{cases} \varphi_i(0) & \text{if } L^i(x) = -1, \\ \varphi_i\big(\sum_{j\in I} W_{j\to i} \sum_{s=L^i(x)+1}^{-1} g_j(-s)x_{j,s}\big) & \text{otherwise.} \end{cases} \tag{4.7}$$

Remark 3. Notice that the functions L^i's introduce a structure of variable-length memory in the model. For this reason the GL neuron model was introduced in Galves and Löcherbach [13] under the name of *Systems of Interacting Chains with Memory of Variable Length*.

Linear spike rate functions. Let us consider the particular case where the parameters of the model are such that $\varphi_i(u) = \nu_i + u$ with $\nu_i \geq 0$ for each $i \in I$. Similarly to Section 4.2.3, let us denote for each $i \in I$,

$$A_i^+ = \big\{(j,s) \in I \times \mathbb{Z}_- : W_{j\to i} g_j(-s) > 0\big\} \quad \text{and} \quad A_i^- = \big\{(j,s) \in I \times \mathbb{Z}_- : W_{j\to i} g_j(-s) < 0\big\},$$

and define the maximal excitatory (respectively inhibitory) strength by

$$\Sigma_i^+ = \sum_{(j,s)\in A_i^+} |W_{j\to i}| g_j(-s) \quad \text{and} \quad \Sigma_i^- = \sum_{(j,s)\in A_i^-} |W_{j\to i}| g_j(-s).$$

We also assume that

$$0 \leq \nu_i - \Sigma_i^- \quad \text{and} \quad \nu_i + \Sigma_i^+ \leq 1. \tag{4.8}$$

Under these assumptions, one can check that the transition probabilities (4.7) also satisfy Assumption 2. Specifically, we can use

$$\begin{cases} \lambda_i(\varnothing) = 1 - (\Sigma_i^+ + \Sigma_i^-) & \text{which is } \geq 0 \text{ since } 0 \leq \Sigma_i^+ + \Sigma_i^- \leq 1, \\ p_i^{\varnothing} = \frac{\nu_i - \Sigma_i^-}{\lambda_i(\varnothing)} & \text{which is } \leq 1 \text{ since } \nu_i + \Sigma_i^+ \leq 1, \\ \lambda_i(\{(j,s)\}^{\downarrow i}) = |W_{j\to i}| g_j(-s) & \text{for all } (j,s) \in A_i^+ \cup A_i^-, \\ p_i^{\{(j,s)\}^{\downarrow i}}(x) = x_{j,s} 1_{x_{i,s:-1}=0} & \text{for all } (j,s) \in A_i^+, \\ p_i^{\{(j,s)\}^{\downarrow i}}(x) = (1 - x_{j,s}) 1_{x_{i,s:-1}=0} & \text{for all } (j,s) \in A_i^-, \end{cases} \tag{4.9}$$

where $\{(j,s)\}^{\downarrow i} = \{(j,s), (i,s), \ldots, (i,-1)\}$ is the augmentation of the set $\{(j,s)\}$ on the coordinate i for each $(j,s) \in A_i^+ \cup A_i^-$. Hence, the random neighborhoods $V \in \mathcal{V}$ have cardinality either 0 (when $V = \varnothing$) or $s+1$ (when $V = \{(j,s)\}^{\downarrow i}$ with $j \neq i$) or s (when $V = \{(i,s)\}^{\downarrow i}$).

Non-linear spike rate functions. In the previous work of Galves and Löcherbach [13], the space–time decomposition is restricted to growing sequences of neighborhoods v that are indexed by their range in time. For each $i \in I$, one assumes that there exists a growing sequence $J_i(1) = \{i\}$, $J_i(\ell) \subset J_i(\ell+1)$ of subsets of I that corresponds to the space positions that are needed when looking at a past of length ℓ, so that we can form $v_i(\ell) = J_i(\ell) \times \underline{\ell}$, defining a growing sequence of subsets of $I \times \mathbb{Z}_-$.

Next let us introduce the following quantities:

$$\alpha_i(\ell) = \inf_{x\in\{0,1\}^{I\times\mathbb{Z}_-}} \Big\{ \inf_{y\in\{0,1\}^{I\times\mathbb{Z}_-} : y \overset{v_i(\ell)}{=} x} p_i(y) + \inf_{y\in\{0,1\}^{I\times\mathbb{Z}_-} : y \overset{v_i(\ell)}{=} x} q_i(y) \Big\}$$

and $\lambda_i(v_i(\ell)) = \alpha_i(\ell) - \alpha_i(\ell-1)$, where for each $i \in I$, $q_i(y) = 1 - p_i(y)$ and $\lambda_i(\varnothing) = \alpha_i(0) = \inf_{x\in\{0,1\}^{I\times\mathbb{Z}_-}} p_i(x) + \inf_{x\in\{0,1\}^{I\times\mathbb{Z}_-}} q_i(x)$.

Let us assume that

$$\sup_{i\in I} \sum_{j\in I} |W_{j\to i}| < \infty, \qquad \sum_{\ell\in\mathbb{Z}_+} \sup_{i\in I} g_i(\ell) < \infty \quad \text{and} \quad \sup_{i\in I} \big|\varphi_i(u) - \varphi_i(v)\big| \leq \gamma |u - v|, \tag{4.10}$$

where γ is a positive constant.

It has been proved in Galves and Löcherbach [13] (see Proposition 2) that the transition probabilities $\{p_i(x), x \in \{0,1\}^{I\times\mathbb{Z}_-}\}$ admit the following space–time decomposition:

$$\begin{cases} p_i(x) = \lambda_i(\varnothing)p_i^{\varnothing} + \sum_{\ell\in\mathbb{Z}_+}\lambda_i(v_i(\ell))p_i^{v_i(\ell)}(x), \\ \lambda_i(\varnothing) + \sum_{\ell=1}^{+\infty}\lambda_i(v_i(\ell)) = 1, \end{cases} \tag{4.11}$$

with, $p_i^{\varnothing} \in [0,1]$ and for $\ell \geq 1$, $p_i^{v_i(\ell)}(x)$ is a $[0,1]$-valued measurable function which is cylindrical in $v_i(\ell)$.

Hence, the transition probabilities p_i's also satisfy Assumption 2 in the nonlinear case. The random neighborhoods $V \in \mathcal{V}$ have cardinality either 0 (when $V = \varnothing$) or $\ell|J_i(\ell)|$ (when $V = v_i(\ell)$). Note that in the non-linear case the neighborhoods $v_i(\ell)$ are dense in time by construction, whereas in the linear case one can obtain a stronger *probabilistic sparsity.*

4.3. *Main properties*

Before being able to assess a value to $X_{i,t}$ at site (i,t) for fixed neuron i and time t, we need to understand on which previous sites this value depends. To do so, we use the distribution λ_i to obtain a space–time neighborhood of (i,t). More precisely, because the distribution λ_i gives a neighborhood for neuron i at time 0, we need to shift it at time t to obtain a realization of the random neighborhood for neuron i at time t by stationarity. Hence if for every $t \in \mathbb{Z}$ and subset A of $I \times \mathbb{Z}$,

$$A^{\to t} = \big\{(j, s+t) \text{ for } (j,s) \in A\big\},$$

with the convention that $\varnothing^{\to t} = \varnothing$, we can define the random neighborhood $K_{i,t}$ of site (i,t) as

$$K_{i,t} = V_{i,t}^{\to t},$$

where $V_{i,t}$ is drawn independently of anything else according to λ_i. We can proceed independently for all sites (j,s) and obtain $K_{j,s} = V_{j,s}^{\to s}$.

By looking recursively at the neighborhoods of the neighborhoods, we are building a whole genealogy in space and time of the site (i,t), that is the list of sites that are really impacting the value $X_{i,t}$. This genealogy is random and depends only on the realizations of the neighborhoods, i.e. only on the distributions λ_i's.

The study of this space–time genealogy is of utmost importance. Indeed if the genealogy is almost surely finite then we are able to follow classical constructions as done by Galves and Löcherbach [13] to write a perfect simulation algorithm. Moreover the study of the length of the genealogy enables us to cut time into almost independent blocks and therefore to have access to concentration inequalities, this second construction being inspired by Viennet [33], Reynaud-Bouret and Roy [26] or Hansen, Reynaud-Bouret and Rivoirard [15].

4.3.1. *Sufficient condition for finite genealogies*

For all sites (i,t), let us define recursively $A_{i,t}^1 = K_{i,t}$ and for $n \geq 1$,

$$A_{i,t}^{n+1} = \Bigg(\bigcup_{(j,s)\in A_{i,t}^n} K_{j,s}\Bigg)\Big\backslash\big\{A_{i,t}^1 \cup \cdots \cup A_{i,t}^n\big\},$$

the genealogy stopped after $n+1$ generations.

The complete genealogy is $G_{i,t} = \bigcup_{n=1}^{\infty} A_{i,t}^n$. It is finite if and only if

$$N_{i,t} = \inf\big\{n \geq 1 : A_{i,t}^n = \varnothing\big\},$$

is finite.

This is a consequence of the following property.

Assumption 5. For each $i \in I$, we assume that the mean size of the random neighborhood on neuron i

$$\bar{m}_i = \sum_{v\in\mathcal{V}} |v|\lambda_i(v), \tag{4.12}$$

is finite and that the maximal mean size satisfies

$$\bar{m} = \sup_{i\in I}\bar{m}_i < 1. \tag{4.13}$$

Probabilistic sparsity corresponds here to the fact that the mean size of the random neighborhoods are strictly less than 1.

Thanks to this assumption, we can prove the following result.

Proposition 1. *For each $i \in I$, $t \in \mathbb{Z}$ and $\ell \in \mathbb{Z}_+$,*

$$\mathbb{P}(N_{i,t} > \ell) \leq (\bar{m})^{\ell}.$$

In particular, under Assumption 5, for all $i \in I$ and $t \in \mathbb{Z}$,

$$\mathbb{P}(N_{i,t} < \infty) = 1, \tag{4.14}$$

that is all genealogies are finite almost surely.

4.3.2. *Perfect simulation algorithm*

Fix a site (i,t) and suppose we want to simulate $X_{i,t}$.

Under Assumption 5, we know the genealogy is finite almost surely and it is possible to build this genealogy recursively without having to generate all the $V_{j,s}$. Once the genealogy is obtained by going backward in time, it is then sufficient to go forward and simulate the $X_{j,s}$'s in the genealogy according to the transitions $p^{V_{j,s}}(X_{K_{j,s}})$.

More formally, we can use two independent fields of independent uniform random variables on $[0,1]$, $\mathbf{U}^1 = (U^1_{i,t})_{i\in I, t\in\mathbb{Z}}$ and $\mathbf{U}^2 = (U^2_{i,t})_{i\in I, t\in\mathbb{Z}}$, such that the whole randomness of the construction is encompassed in the field $\mathbf{U}^1$ for the genealogies and in the field $\mathbf{U}^2$ for the forward transitions and such that conditionally on these two fields, the whole simulation algorithm is deterministic. But in practice, we generate $U^1_{j,s}$ and $U^2_{j,s}$ only if we need it. This leads to the following algorithm

Step 1. Generate $U^1_{i,t}$ random uniform variable on $[0,1]$. Since $\mathcal{V}$ is countable, one can order its elements such that $\mathcal{V} = \{v_1, \ldots, v_n, \ldots\}$. Define the c.d.f. of λ_i by $F_i(0) = \lambda_i(\varnothing)$ and for $n \geq 1$,

$$F_i(n) = \lambda_i(\varnothing) + \sum_{k=1}^{n} \lambda_i(V_k)$$

and pick the random neighborhood of (i,t) as

$$K_{i,t} = V_{i,t}^{\rightarrow t} \quad \text{with } V_{i,t} = \begin{cases} \varnothing & \text{if } U^1_{i,t} \leq F_i(0), \\ v_n & \text{if } F_i(n-1) < U^1_{i,t} \leq F_i(n) \text{ for some } n \geq 1. \end{cases}$$

Initialize $A^1_{i,t} \leftarrow K_{i,t}$.

Step 2. Generate recursively $U^1_{j,s}$ for $j,s \in A^n_{i,t}$, compute the corresponding $V_{j,s}$ and $K_{j,s}$ as in Step 1 and actualize $A^{n+1}_{i,t} \leftarrow (\bigcup_{j,s\in A^n_{i,t}} K_{j,s}) \setminus \{A^1_{i,t} \cup \cdots \cup A^n_{i,t}\}$. After a finite number of steps, $A^n_{i,t}$ is empty and [Step 2] stops. Let $N_{i,t}$ be the final n of this recursive procedure and the genealogy of (i,t) is given by $G_{i,t} = \bigcup_{n=1}^{N_{i,t}} A^n_{i,t}$.

Step 3. Note that the (j,s)'s in $A^{N_{i,t}-1}_{i,t}$ have therefore an empty neighborhood. Generate i.i.d. uniform variables $U^2_{j,s}$ for (j,s) in $A^{N_{i,t}-1}_{i,t}$ and define

$$X_{j,s} = 1\{U^2_{j,s} \leq p^{\varnothing}_j\}. \tag{4.15}$$

Step 4. Recursively generate $U^2_{j,s}$ for (j,s) in $A^{\ell}_{i,t}$ recursively from $\ell = N_{i,t} - 2$ to $\ell = 1$ and define

$$X_{j,s} = 1\{U^2_{j,s} \leq p^{V_{j,s}}_j(X_{K_{j,s}})\}, \tag{4.16}$$

In particular arrived at $\ell = 1$, one generates

$$X_{i,t} = 1\{U^2_{i,t} \leq p^{V_{i,t}}_j(X_{K_{i,t}})\}. \tag{4.17}$$

It is well-known that the algorithm above not only shows the existence but also the uniqueness of the stochastic chain $\mathbf{X}$ compatible with Assumptions 1, 2 and 5 (see for instance Galves and Löcherbach [13] for formal statement of this result in a close setup).

Note that when simulating the linear Hawkes process, the algorithm reduces to a random walk in the past to find the genealogy, a random decision on the state $X_{j,s}$ at the end of the random walk and a forward decision of the other states $X_{j,s}$ which is then completely deterministic and just depends on the sign of $h_{j\to i}(-s)$.

4.3.3. *Time length of a genealogy*
We are now interested by the time length of a genealogy. Let, for each non-empty subset A of $I\times\mathbb{Z}$,

$$\mathbb{T}(A)=\min\bigl\{s\in\mathbb{Z}:(j,s)\in A\bigr\}.$$

We are interested by the variable $T_{i,t}$ which is equal to $t-\mathbb{T}(A_{i,t})$ if the genealogy $G_{i,t}$ is non empty and equal to 0 if $G_{i,t}$ is empty. By stationarity its distribution does not depend on t and the behavior of this variable is of course linked to the one of the variables $T(V_j)=-\mathbb{T}(V_j)$ for V_j obeying the distribution λ_j, with the convention that $T(\varnothing)=0$. We are interested by conditions under which the variable $T_{i,t}$ has a Laplace transform, that is when

$$\theta\mapsto\Psi_i(\theta)=\mathbb{E}\bigl(e^{\theta T_{i,t}}\bigr)$$

is finite for some positive θ. To do so, we are going to assume Assumption 3 that we recall here: There exists a strictly positive θ such that for all i,

$$\varphi_i(\theta)=\sum_{v\in\mathcal{V}}|v|e^{\theta T(v)}\lambda_i(v)$$

is finite and

$$\varphi(\theta)=\sup_{i\in I}\varphi_i(\theta)<1. \tag{4.18}$$

Theorem 2. *Under Assumption 3, for all i in I, $\Psi_i(\theta)$ is finite and*

$$\Psi(\theta)=\sup_{i\in I}\Psi_i(\theta)\le\frac{\sup_{i\in I}\lambda_i(\varnothing)}{1-\varphi(\theta)}.$$

Note that if $\varphi_i(\theta)$ is finite for some positive θ, $\lim_{\theta\to0}\varphi_i(\theta)=\bar{m}_i$. Therefore if Assumption 5 is fulfilled, $\lim_{\theta\to0}\varphi_i(\theta)<1$ and it is possible to find $\theta>0$ such that $\varphi_i(\theta)<1$ as soon as λ_i has a Laplace transform. In this sense, and roughly speaking, Assumption 3 is a more stringent condition of *probabilistic sparsity* than Assumption 5.

4.3.4. *Application on the main examples*
Markov chains. In this case, $\bar{m}=1-\mu$ and the condition (4.13) is satisfied as soon as $\mu<1$. Moreover condition (4.18) reduces to $e^{\theta}(1-\mu)<1$ and it is always possible to find such a $\theta>0$ as soon as $\mu<1$.

Chains of infinite order. The space–time decomposition (4.3) implies that

$$\bar{m}=\sum_{\ell=1}^{\infty}\ell\lambda(\underline{\ell}).$$

Thus, the condition (4.13) is satisfied as soon as

$$\sum_{\ell=1}^{\infty}\ell\lambda(\underline{\ell})<1$$

and similarly the condition (4.18) is satisfied as soon as

$$\sum_{\ell=1}^{\infty}\ell e^{\theta\ell}\lambda(\underline{\ell})<1.$$

Hence both can be verified if λ is sufficiently exponentially decreasing. Typically one can have $\lambda(\underline{\ell})=e^{-\lambda}\lambda^{\ell}/\ell!$ with $0<\lambda<1$ (Poisson distribution on the range) or $\lambda(\underline{\ell})=(1-p)^{\ell}p$ with $1/2<p\le1$ (Geometric distribution on the range).

Discrete-time linear Hawkes processes. According to the space–time decomposition (4.6), it follows that for each $i \in I$,

$$m_i = \Sigma_i^+ + \Sigma_i^-.$$

Therefore, the condition (4.13) reduces to

$$\sup_{i\in I}\bigl(\Sigma_i^+ + \Sigma_i^-\bigr) = \sup_{i\in I}\sum_{j,s}\bigl|h_{j\to i}(-s)\bigr| < 1.$$

Moreover the condition (4.18) becomes

$$\sup_{i\in I}\sum_{j,s} e^{\theta s}\bigl|h_{j\to i}(-s)\bigr| < 1.$$

So if for instance we can rewrite $h_{j\to i}(-s) = w_{j\to i}g(-s)$ for a fixed function g of mean 1, the condition (4.13) reduces to

$$\sup_{i\in I}\sum_{j\in I}|w_{j\to i}| < 1,$$

and the additional condition (4.18) is fulfilled for a small enough θ as soon as g has finite exponential moment.

GL neuron model. In the nonlinear case, it has been proved in Galves and Löcherbach [13] (cf. inequalities (5.57) and (5.58)) that for each $i \in I$ the following estimates hold:

$$\lambda_i(\varnothing) \le \gamma\sum_{j\in I}|W_{j\to i}|\sum_{s\ge 1} g_j(s), \tag{4.19}$$

and for $\ell \ge 1$,

$$\lambda_i\bigl(v_i(\ell)\bigr) \le \gamma\Biggl(\sum_{j\notin v_i(\ell)}|W_{j\to i}|\sum_{s\ge 1} g_j(s) + \sum_{j\in v_i(\ell)}|W_{j\to i}|\sum_{s\ge \ell} g_j(s)\Biggr). \tag{4.20}$$

Therefore, a sufficient condition (cf. inequality (2.9) of Galves and Löcherbach [13]) for Assumption 5 to hold is

$$\sup_{i\in I}\sum_{\ell\ge 1}\ell\bigl|v_i(\ell)\bigr|\Biggl(\sum_{j\notin v_i(\ell)}|W_{j\to i}|\sum_{s\ge 1} g_j(s) + \sum_{j\in v_i(\ell)}|W_{j\to i}|\sum_{s\ge \ell} g_j(s)\Biggr) < \frac{1}{\gamma}.$$

In the linear case (i.e. when $\varphi_i(u) = \nu_i + u$), the condition above reduces to

$$\sup_{i\in I}\sum_{\ell\ge 1}\ell\bigl|v_i(\ell)\bigr|\Biggl(\sum_{j\notin v_i(\ell)}|W_{j\to i}|\sum_{s\ge 1} g_j(s) + \sum_{j\in v_i(\ell)}|W_{j\to i}|\sum_{s\ge \ell} g_j(s)\Biggr) < 1. \tag{4.21}$$

Using the decomposition (4.9), one can verify that the condition (4.13) is, in the linear case, equivalent to

$$\sup_{i\in I}\sum_{\ell\ge 1}\Biggl[\ell|W_{i\to i}|g_i(\ell) + \sum_{j\ne i,\,j\in I}(\ell+1)|W_{j\to i}|g_j(\ell)\Biggr] < 1. \tag{4.22}$$

Note that condition (4.21) is usually much stronger than condition (4.22) and that a sparse space–time decomposition of the process allows us to derive existence of the linear process on a larger set of possible choices for $w_{j\to i}$ and g_j. Once again condition (4.18) is fulfilled under a very similar expression

$$\sup_{i\in I}\sum_{\ell\ge 1} e^{\theta\ell}\Biggl[\ell|W_{i\to i}|g_i(\ell) + \sum_{j\ne i,\,j\in I}(\ell+1)|W_{j\to i}|g_j(\ell)\Biggr] < 1,$$

this can be easily fulfilled if $g_j(\ell) = g(\ell)$ is exponentially decreasing with $\sum_{\ell=1}^{\infty}(\ell+1)g(\ell) = 1$. Indeed (4.22) is implied as in the Hawkes case by

$$\sup_{i\in I}\sum_{j\in I}|W_{j\to i}| < 1$$

and it is easy to find by continuity a small $\theta > 0$ such that (4.18) is fulfilled too.

4.4. *Concentration*

4.4.1. *Block construction*

Thanks to the control of the time length genealogy it is possible to cut the observations $X_{F,-(m-1):T}$ into (overlapping) blocks that form with high probability two families of independent variables. This is a key tool to derive concentration inequalities. This construction is inspired by Viennet [33], who used as a central element, Berbee's lemma, which is replaced here by Theorem 2. Note that similar coupling arguments have been used in continuous and more restrictive settings (see Reynaud-Bouret and Roy [26], Hansen, Reynaud-Bouret and Rivoirard [15] for linear Hawkes processes, Chen et al. [4] for bounded Hawkes process and mixing arguments).

Lemma 1. *Let $m \in \mathbb{Z}_+$ and $F \subset I$ be a finite subset of the neurons, observed on $-(m-1):T$. Let B, the grid size, be an integer such that*

$$m \le B \le \lfloor T/2 \rfloor$$

and define $k = \lfloor \frac{T}{2B} \rfloor$. Let the $2k+1$ blocks be defined by, for $1 \le n \le 2k$,

$$I_n = \big\{(n-1)B + 1 - m, \dots nB\big\} \quad \text{and} \quad I_{2k+1} = \{2kB + 1 - m, \dots T\}.$$

There exist on a common probability space some stochastic chains $\mathbf{X}, \mathbf{X}^1, \dots, \mathbf{X}^{2k+1}$ satisfying the following properties:

1. *All the chains $\mathbf{X}^n = (X^n_{i,t})_{i\in I, t\in\mathbb{Z}}$ have the same distribution as $\mathbf{X}$ which satisfies Assumptions 1, 2 and 3 for a given θ, that is a sparse enough space–time decomposition with weights $(\lambda_i)_{i\in I}$ and transitions $(p^v_i)_{i\in I, v\in\mathcal{V}}$.*
2. *The odd chains $\mathbf{X}^1, \mathbf{X}^3, \dots, \mathbf{X}^{2k+1}$ are independent.*
3. *The even chains $\mathbf{X}^2, \dots, \mathbf{X}^{2k}$ are independent.*
4. *There exists an event, Ω_{good}, such that on Ω_{good}, $X_{F,I_n} = X^n_{F,I_n}$ for all $n = 1, \dots, 2k+1$ and such that the probability of Ω^c_{good}, under the notation of Theorem 2, is at most*

$$|F|(2k+1)\frac{\Psi(\theta)}{(1-e^{-\theta})}e^{-\theta(B+1-m)}. \tag{4.23}$$

In particular, by choosing $B = m + \theta^{-1}(2\log(T) + \log(|F|))$, we obtain that there exists a positive $c'(\theta)$ such that the probability of Ω^c_{good} is at most $c'(\theta)T^{-1}$.

4.4.2. *Applications*

As an application of Lemma 1, we can derive the following Hoeffding type concentration inequality.

Theorem 3. *Let $\mathbf{X} = (X_{i,t})_{i\in I, t\in\mathbb{Z}}$ be a stationary sparse space–time process satisfying Assumptions 1, 2 and 3 for a given θ. For $F \subset I$ finite, $m \in \mathbb{Z}_+$, let f be a real-valued function of $X_{F,t-m:t-1}$ bounded by M. Let $T \in \mathbb{Z}_+$ such that*

$$m + \theta^{-1}\big(2\log(T) + \log\big(|F|\big)\big) \le \lfloor T/2 \rfloor$$

and

$$Z(f) = \frac{1}{T}\sum_{t=1}^{T}\big(f(X_{F,t-m:t-1}) - \mathbb{E}\big[f(X_{F,t-m:t-1})\big]\big). \tag{4.24}$$

Then there exists nonnegative constant c', c'', which only depends on θ such that, for any $x > 0$,

$$\mathbb{P}\left(Z(f) > \sqrt{c''(\theta)M^2\frac{m + \log T + \log|F|}{T}x}\right) \le \frac{c'(\theta)}{T} + 2e^{-x}. \tag{4.25}$$

If there is a finite family $\mathcal{F}$ of such f, we also have that

$$\mathbb{P}\left(\exists f \in \mathcal{F}, Z(f) > \sqrt{c''(\theta)M^2\frac{m + \log T + \log|F|}{T}x}\right) \le \frac{c'(\theta)}{T} + 2|\mathcal{F}|e^{-x}.$$

There is a matrix counterpart to the previous inequality, which is an application of now classical results on random matrices (see Tropp [30] and the references therein).

Theorem 4. *Let $\mathbf{X} = (X_{i,t})_{i\in I, t\in\mathbb{Z}}$ be a stationary sparse space–time process satisfying Assumptions 1, 2 and 3 for a given θ. For $F \subset I$ finite, $m \in \mathbb{Z}_+$, let $\mathcal{F}$ be a finite family of bounded real-valued functions of $X_{F,t-m:t-1}$ and denote $M = \max\{\|fg\|_\infty : f, g \in \mathcal{F}\}$. Let $T \in \mathbb{Z}_+$ such that*

$$m + \theta^{-1}\big(2\log(T) + \log\big(|F|\big)\big) \le \lfloor T/2 \rfloor$$

and define the random matrix $Z = (Z(f,g))_{f,g\in\mathcal{F}}$ where for each $f, g \in \mathcal{F}$,

$$Z(f,g) = \frac{1}{T}\sum_{t=1}^{T}\big(f(X_{F,t-m:t-1})g(X_{F,t-m:t-1}) - \mathbb{E}\big[f(X_{F,t-m:t-1})g(X_{F,t-m:t-1})\big]\big). \tag{4.26}$$

Then there exists nonnegative constant c', c'', which only depends on θ such that, for any $x > 0$,

$$\mathbb{P}\left(\|Z\| > \sqrt{c''(\theta)M^4|\mathcal{F}|^2\frac{m + \log T + \log|F|}{T}x}\right) \le \frac{c'(\theta)}{T} + 4|\mathcal{F}|e^{-x}, \tag{4.27}$$

where $\|Z\|$ corresponds to the spectral norm, that is the largest eigenvalue of the self-adjoint matrix Z.

5. Back to the Gram matrices

To control the Gram matrix we need also Assumption 4 that we recall here: There exists some positive μ such that for all $i \in I$, for all x,

$$\mu \le p_i(x) \le 1 - \mu,$$

Note that in each of the examples (Markov chain, Hawkes, etc), this assumption is easily fulfilled. For instance in the Hawkes case, this adds the condition $\mu \le \nu_i - \Sigma_i^- \le \nu_i + \Sigma_i^+ \le (1-\mu)$.

This assumption is useful to bound expectation by changing the underlying measure.

Lemma 2. *Under Assumptions 1 and 4, for all non negative function f cylindrical on a fixed finite space–time neighborhood v,*

$$\big(2(1-\mu)\big)^{|v|}\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}\big[f(X_v)\big] \ge \mathbb{E}\big[f(X_v)\big] \ge (2\mu)^{|v|}\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}\big[f(X_v)\big],$$

where $\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}$ means that the expectation is taken with respect to the measure where all $X_{i,t}$'s are i.i.d. Bernoulli with parameter $1/2$.

5.1. *Inv(κ) property for general dictionaries*

In this section we prove that the Inv(κ) property holds on an event with high probability for the examples of dictionaries considered in Section 3.1. As a by product, we are able to derive oracle inequalities with high probability for these dictionaries. We start with the following result.

Theorem 5. *For a finite $F \subset I$ and integer $T > m \ge 1$, let $X_{F,-(m-1):T}$ be a sample produced by the stationary sparse space–time process $\mathbf{X} = (X_{i,t})_{i\in I,t\in\mathbb{Z}}$ satisfying Assumptions 1, 2 and 3. Let Φ denote a finite dictionary of bounded functions cylindrical in $F \times \underline{m}$ and G be the corresponding Gram matrix defined in (3.2). If the matrix $\mathbb{E}(G)$ satisfies property* Inv(κ') *for some positive constant κ', then for any $\delta > 0$ and T sufficiently large, the Gram matrix G satisfies the property* Inv(κ) *on an event of probability larger than $1 - \frac{c'}{T} - \delta$ with*

$$\kappa = \kappa' - c_1|\Phi|\|\Phi\|_\infty^2\sqrt{\frac{(m + \log(T) + \log|F|)(\log|\Phi| + \log\delta^{-1})}{T}},$$

where c' and c_1 are positive constants which only depends on the underlying distribution of $\mathbf{X}$.

To apply Theorem 5 to the dictionaries considered in Section 3.1 we must find the corresponding κ'. This is done below.

Short memory effect. To apply Theorem 5 we need first to find κ' for this class of models. This is done as follows. Let $Q = \mathcal{B}(1/2)^{\otimes\mathcal{V}}$ be the probability measure under which all $X_{i,t}$'s are i.i.d. Bernoulli with parameter $1/2$ and denote $p_j = Q(\varphi_j(X_{-\infty:-1}) = 1)$ for $j \in F$. Clearly, $p_j = 1 - (1/2)^m$ for all $j \in F$ and we write p to denote this common value. With this notation, one can check that,

$$\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(G) = \begin{pmatrix} p & p^2 & p^2 & \dots & p^2 \\ p^2 & p & p^2 & \dots & p^2 \\ & & & \dots & \\ p^2 & p^2 & p^2 & \dots & p \end{pmatrix}.$$

Such a matrix has only two eigenvalues, namely, $p + (|F| - 1)p^2$ of multiplicity 1 and $p - p^2 = (1/2)^m(1 - (1/2)^m)$ with multiplicity $|F| - 1$. Indeed, ξ is an eigenvalue $\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(G)$ if and only if there exists a non-null vector $u \in \mathbb{R}^F$ such that

$$\big(p - p^2\big)u + p^2 \sum_i u_i \mathbf{1} = \xi u.$$

On the one hand, by choosing the vector $u \neq 0$ such that $\sum_i u_i = 0$ gives that $\eta = p - p^2$ is an eigenvalue with multiplicity $|F| - 1$. On the other hand, the choice $\sum_i u_i = 1$ forces that $(p - p^2)u_i + p^2 = \xi u_i$ for all $i \in F$, ensuring that $\xi = p + p^2(|F| - 1)$ is the second eigenvalue. Its multiplicity is necessarily 1.

Note that if m is large, the smallest eigenvalue of $\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(G)$ is really small. This can be interpreted in the following way: when m is large, one will find a "1" on every observed neuron in the past, therefore all the φ_j's will be equal with high probability and one cannot infer a dependence graph with this dictionary anymore.

Thus, Lemma 2 implies that eigenvalue of $\mathbb{E}(G)$ can be lower bounded by

$$\kappa' = (2\mu)^{m|F|}(1/2)^m\big(1 - (1/2)^m\big). \tag{5.1}$$

Choosing for a fixed integer η

$$m = \eta \quad \text{and} \quad |F| \leq \log\log T, \tag{5.2}$$

gives κ' of the order $(\log(T))^{-c_3}$ for some constant $c_3 > 0$ depending on μ and η.

Cumulative effect. Let α denote the common value of $\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(\varphi^2_{j,\ell}(X_{-\infty:-1}))$ with $j \in F$ and $1 \leq \ell \leq L$, and β be the corresponding value of $\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(\varphi_{j,\ell}(X_{-\infty:-1})(\varphi_{k,n}(X_{-\infty:-1}))$ with $j, k \in F$ and $k \neq j$ and $1 \leq n, \ell \leq L$. With this notation, one can verify that

$$\alpha = \frac{\eta}{2} + \frac{\eta(\eta-1)}{4} = \frac{\eta}{4} + \frac{\eta^2}{4}, \qquad \beta = \frac{\eta^2}{4} \quad \text{and} \quad \mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(G) = \begin{pmatrix} \alpha & \beta & \beta & \dots & \beta \\ \beta & \alpha & \beta & \dots & \beta \\ & & & \dots & \\ \beta & \beta & \dots & \beta & \alpha \end{pmatrix}.$$

Hence, the smallest eigenvalue of $\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(G)$ is $\alpha - \beta = \frac{\eta}{4}$ which grows with $\eta = \frac{m}{K}$. This seems also reasonable since once looking for cumulative effects, the larger the bin size η, the more points you see in it and the more diverse the situations are (hence the dictionary has many different functions) whereas if η is small there is a large probability to see all $\varphi_{j,\ell}$'s null.

Thus, Lemma 2 implies that eigenvalue of $\mathbb{E}(G)$ can be lower bounded by

$$\kappa' = \frac{\eta}{4}(2\mu)^{\eta K|F|}.$$

Choosing for some fixed integer η

$$m = \eta K \quad \text{with } K \leq \sqrt{\log\log T} \quad \text{and} \quad |F| \leq \log\log T, \tag{5.3}$$

gives κ' of the order $(\log(T))^{-c_3}$ for some other constant $c_3 > 0$ depending on μ and η.

Cumulative effect with spontaneous apparition. With the same notation of the previous example, one can show that

$$\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(G) = \begin{pmatrix} 1 & \eta/2 & \eta/2 & \dots & \eta/2 \\ \eta/2 & \alpha & \beta & \dots & \beta \\ & & \dots & & \\ \eta/2 & \beta & \dots & \beta & \alpha \end{pmatrix}.$$

Reasoning by block with the vector (μ, a) with $\mu \in \mathbb{R}$ and $a \in \mathbb{R}^{K|F|}$, we end up with

$$(\mu,a)^\intercal \mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(G)(\mu,a) = \left(\mu + \frac{\eta}{2} \sum_{j\in F, k=1,\dots,K} a_{j,k}\right)^2 + \frac{\eta}{4}\|a\|_2^2.$$

But for all $0 < \theta < 1$,

$$\left(\mu + \frac{\eta}{2}\sum a_{j,k}\right)^2 + \frac{\eta}{4}\|a\|_2^2 \geq (1-\theta)\mu^2 + \left(1-\frac{1}{\theta}\right)\frac{\eta^2}{4}\left(\sum_{j\in F,k=1,\dots,K} a_{j,k}\right)^2 + \frac{\eta}{4}\|a\|_2^2$$

$$\geq (1-\theta)\mu^2 - \frac{1-\theta}{\theta}\frac{K|F|\eta^2}{4}\|a\|_2^2 + \frac{\eta}{4}\|a\|_2^2$$

By choosing $\theta = \frac{2\eta K|F|}{1+2\eta K|F|}$ we conclude, thanks to Lemma 2, that the smallest eigenvalue of $\mathbb{E}(G)$ can be lower bounded by

$$\kappa' = (2\mu)^{\eta K|F|} \min\left(\frac{1}{1+2\eta K|F|}, \frac{\eta}{8}\right).$$

Once again choosing for some fixed integer η

$$m = \eta K \quad \text{with } K \leq \sqrt{\log\log T} \quad \text{and} \quad |F| \leq \log\log T, \tag{5.4}$$

gives κ' roughly larger than $(\log(T))^{-c_3}$ for some other constant $c_3 > 0$ depending on μ and η.

Next, as a by product of Theorem 5 and Theorem 1, one can derive oracle inequalities for dictionaries above.

Corollary 1. *Let Φ be one of the dictionaries presented in Section 3.1, with the choices (5.2), (5.3) or (5.4). Assume one observes $X_{F,-(m-1):T}$, where the underlying process $\mathbf{X}$ satisfies Assumptions 1, 2, 3 and 4.*

With the notation of Theorem 1, for T large enough, on an event with probability $1 - c_1/T$, the following oracle inequality holds

$$\left\|\hat{f} - p_i(\cdot)\right\|_T^2 \leq \inf_{a\in\mathbb{R}^\Phi}\left\{\left\|f_a - p_i(\cdot)\right\|_T^2 + c_2|S(a)|\frac{(\log(T))^{c_3}}{T}\right\},$$

where the constant $c_1 > 0$ depends only on the underlying distribution of $\mathbf{X}$, $c_2 > 0$ depends on η and γ and constant $c_3 > 0$ depends on both the underlying distribution of $\mathbf{X}$ and η.

Note that the main improvement with respect to Hansen, Reynaud-Bouret and Rivoirard [15], is that in all the examples, the constant κ is roughly of order $(\log(T))^{-c_3}$, that is asymptotically decreasing in roughly speaking the number of neurons used in the dictionary and not the total number of neurons in the network. This number of neurons that are used, which is bounded by the number of observed neurons, can very slowly grow with T.

5.2. *Hawkes dictionary without spontaneous part*

In this case the $\varphi(X_{F,-m:-1})$'s are just the $X_{j,s}$ for $j \in F$ and $s \in \underline{m}$ and one can prove the following result.

Theorem 6. *For a finite $F \subset I$ and integer $T > m \geq 1$, let $X_{F,-(m-1):T}$ be a sample produced by the stationary sparse space–time process $\mathbf{X} = (X_{i,t})_{i\in I, t\in\mathbb{Z}}$ satisfying Assumptions 1, 2, 3 and 4. For the Hawkes dictionary without spontaneous part, i.e. $\varphi = \varphi_{j,s}$ with $\varphi_{j,s}(X_{F,-m:-1}) = X_{j,s}$ for $j \in F$ and $s \in \underline{m}$, the corresponding Gram matrix G defined by*

(3.2) *satisfies for all* $c > 0$, $s \leq m|F|$ *and* T *large enough, the property* $\mathbf{RE}(\kappa, c, s)$ *on an event of probability larger than* $1 - \frac{c'}{T} - \delta$ *with*

$$\kappa = \mu - \mu^2 - \big((1-2\mu) + R_T\big)(1+c)s,$$

where

$$R_T = \frac{c_1}{T^{1/2}}\big(m + \log T + \log|F|\big)^{1/2}\big(\log m + \log|F| + \log\delta^{-1}\big)^{1/2},$$

for some positive constant c' *and* c_1 *which only depends on the underlying distribution of* $\mathbf{X}$.

The major point to note is that asymptotically, for slowly growing m and $|F|$ as functions of T, the constant κ does not depend at all on the number of observed neurons and therefore the rate of convergence in Theorem 1 is not worsened by a huge number of observed neurons, $|F|$. This is a drastic improvement with respect to the previous result of Hansen, Reynaud-Bouret and Rivoirard [15] which depends on the total number of neurons in the network. For each fixed c and s, we only need here μ to be close enough to $1/2$ to have $\kappa > 0$.

It also means that the size of the dictionary might be growing with T, much more rapidly than before: typically m the delay might grow like $\log(T)$ and the number of observed neurons might grow like T or even more rapidly as long as $\log|F| = o(T^{1/2})$. Therefore if one can reasonably well approximate p_i by a sparse combination in space and time for which the precise location is unknown, one might by a growing set of observations find the correct set in space and time.

6. Conclusion

It is therefore possible to control the Gram matrix for various dictionaries and this even if the finite number of observed neurons is much smaller than the potentially infinite set of existing neurons. The main assumption on the underlying stochastic structure is the *probabilistic sparsity* (Assumption 3) which allows us to derive concentration inequalities via coupling.

As an open question, it remains to understand the complete link between a well chosen deterministic sparse approximation of p_i and the *probabilistic sparsity* of the λ_i's typically when both the approximation model and the true underlying model coincide, for instance for Hawkes processes. Another way to phrase this is "can we prove the variable selection property, that is typical of Lasso methods?" i.e. "can we find the set of neurons influencing i?". If the answer seems likely to be yes if they are all observed, it seems intuitive to think in general that a good set of sites (j, s) for the sparse approximation of p_i is a level set of the λ_i but the fact that the λ_i's are not unique makes this reasoning not straightforward.

Another open question is the minimax rate in this setting. This would involve speaking about regularity of the space–time decomposition, which is not done yet, since we do not even have for the moment uniqueness of the decomposition.

7. Proofs

7.1. *Proof of Theorem 1*

To prove Theorem 1 we use arguments from Gaïffas and Guilloux [10]. We will need the following Lemmas.

Lemma 3. *Let* $\hat{f} = f_{\hat{a}}$ *where* $\hat{a}$ *is defined by* (3.3). *For any vector* $a \in \mathbb{R}^{\Phi}$, *the following inequality holds*

$$2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T + \gamma d|\hat{a}_{S^c(a)}|_1 \leq \gamma d|\hat{a}_{S(a)} - a_{S(a)}|_1 + 2(b - \bar{b})^T(\hat{a} - a), \tag{7.1}$$

where $S(a) = \{\varphi : a_\varphi \neq 0\}$ *and the vectors* $b, \bar{b} \in \mathbb{R}^{\Phi}$ *are defined in* (3.2) *and* (3.4) *respectively.*

Proof. Throughout the proof we write $\partial g(p)$ to denote the subddiferential mapping of a convex function g at the point p. One can show that p is a global minimum of the convex function g if and only if $0 \in \partial g(p)$. Now since $\hat{a}$ is such that

$$\hat{a} \in \arg\min_{a \in \mathbb{R}^{\Phi}} \big\{a^T G a - 2a^T b + \gamma d|a|_1\big\},$$

it follows that

$$0 \in \partial\big(\hat{a}^T G\hat{a} - 2\hat{a}^T b + \gamma d|\hat{a}|_1\big) = 2G\hat{a} - 2b + \gamma d\partial|\hat{a}|_1.$$

Thus, it follows that for some $\hat{w} \in \partial|\hat{a}|_1$, the following equation holds

$$2G\hat{a} - 2b + \gamma d\hat{w} = 0,$$

which implies then

$$(2G\hat{a} - 2b + \gamma d\hat{w})^T(\hat{a} - a) = 0, \quad \text{for any } a \in \mathbb{R}^\Phi.$$

From the above equation we can deduce that for any vector $w \in \partial|a|_1$ and $a \in \mathbb{R}^\Phi$,

$$(2G\hat{a} - 2\bar{b})^T(\hat{a} - a) + \gamma d(\hat{w} - w)^T(\hat{a} - a) = -\gamma d w^T(\hat{a} - a) + 2(b - \bar{b})^T(\hat{a} - a). \tag{7.2}$$

One can easily show by the definition of subdifferentials that

$$(\hat{w} - w)^T(\hat{a} - a) \geq 0,$$

for all $\hat{w} \in |\hat{a}|_1$ and $w \in |a|_1$. Thus, using this fact in equation (7.2) together with the fact that $(2G\hat{a} - 2\bar{b})^T(\hat{a} - a) = 2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T$, we derive the following inequality

$$2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T \leq -\gamma d w^T(\hat{a} - a) + 2(b - \bar{b})^T(\hat{a} - a). \tag{7.3}$$

It is well know that

$$\partial|a|_1 = \left\{v : |v|_\infty \leq 1 \text{ and } v^T a = |a|_1\right\}.$$

In other words, $v \in \partial|a|_1$ if and only if $v_\varphi = \operatorname{sign}(a_\varphi)$ for $\varphi \in S(a)$ and $v_\varphi \in [-1, 1]$ for all $\varphi \in S^c(a)$. Now, take $w = (w_\varphi)_{\varphi\in\Phi} \in \partial|a|_1$ of the following form

$$w_\varphi = \begin{cases} \operatorname{sign}(a_\varphi) & \text{if } \varphi \in S(a), \\ \operatorname{sign}(\hat{a}_\varphi) & \text{if } \varphi \in S^c(a), \end{cases}$$

and observe that $w^T(\hat{a} - a) = \sum_{\varphi\in S(a)} \operatorname{sign}(a_\varphi)(\hat{a}_\varphi - a_\varphi) + |\hat{a}_{S^c(a)}|_1$. Thus, by plugging this identify into inequality (7.3), we obtain that

$$2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T + \gamma d|\hat{a}_{S^c(a)}|_1 \leq -\gamma d \sum_{\varphi\in S(a)} \operatorname{sign}(a_\varphi)(\hat{a}_\varphi - a_\varphi) + 2(b - \bar{b})^T(\hat{a} - a),$$

and the result follows, because $|-\sum_{\varphi\in S(a)} \operatorname{sign}(a_\varphi)(\hat{a}_\varphi - a_\varphi)| \leq |\hat{a}_{S(a)} - a_{S(a)}|_1$. □

Lemma 4. *Let $\hat{f} = f_{\hat{a}}$ where $\hat{a}$ defined by (3.3) with $\gamma \geq 2$ and $a \in \mathbb{R}^\Phi$. On an event on which*

(i) $\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T \geq 0$,
(ii) $|b_\varphi - \bar{b}_\varphi| \leq d$ *for all* $\varphi \in \Phi$,

the following inequality is satisfied,

$$|\hat{a}_{S^c(a)}|_1 \leq \frac{\gamma + 2}{\gamma - 2}|\hat{a}_{S(a)} - a_{S(a)}|_1, \tag{7.4}$$

where $S(a) = \{\varphi : a_\varphi \neq 0\}$.

Proof. Suppose that $\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T \geq 0$. In this case, Lemma 3 implies that

$$\gamma d|\hat{a}_{S^c(a)}|_1 \leq \gamma d|\hat{a}_{S(a)} - a_{S(a)}|_1 + 2\sum_{\varphi\in S(a)} (b_\varphi - \bar{b}_\varphi)(\hat{a}_\varphi - a_\varphi) + 2\sum_{\varphi\in S^c(a)} (b_\varphi - \bar{b}_\varphi)\hat{a}_\varphi.$$

On an event on which $|b_\varphi - \bar{b}_\varphi| \leq d$ for all $\varphi \in \Phi$, we then have that

$$\gamma d|\hat{a}_{S^c(a)}|_1 \leq (\gamma + 2)d|\hat{a}_{S(a)} - a_{S(a)}|_1 + 2d|\hat{a}_{S^c(a)}|_1,$$

and the result follows. □

To prove the first part of Theorem 1 we proceed as follows. First of all, on the event on which $\langle \hat{f} - f_a, \hat{f} - p_i \rangle_T < 0$, there is nothing to be proved, since in this case

$$\|\hat{f} - p_i\|_T^2 + \|\hat{f} - f_a\|_T^2 - \|f_a - p_i\|_T^2 = \langle \hat{f} - f_a, \hat{f} - p_i \rangle_T < 0.$$

Hence, in what follows, take $a = (a_\varphi)_{\varphi \in \Phi}$ such that $|S(a)| \leq s$ and $\langle \hat{f} - f_a, \hat{f} - p_i \rangle_T \geq 0$. In this case, thanks to Lemma 4, we can use Property $\mathbf{RE}(\kappa, c(\gamma), s)$ to the vector $\hat{a} - a$:

$$\|\hat{a}_{S(a)} - a_{S(a)}\|^2 \leq \kappa^{-1} (\hat{a} - a)^T G (\hat{a} - a).$$

Now, as in the proof of Lemma 4, we know that on an event on which $|b_\varphi - \bar{b}_\varphi| \leq d$ for all $\varphi \in \Phi$, the following bound holds:

$$2\big|(b - \bar{b})^T (\hat{a} - a)\big| \leq 2d\big|(\hat{a}_{S(a)} - a_{S(a)})\big|_1 + 2d|\hat{a}_{S^c(a)}|_1$$

By using this inequality together with Lemma 3, we conclude that

$$2\langle \hat{f} - f_a, \hat{f} - p_i \rangle_T + (\gamma - 2)d|\hat{a}_{S^c(a)}|_1 \leq (\gamma + 2)d|\hat{a}_{S(a)} - a_{S(a)}|_1. \tag{7.5}$$

Finally, by Cauchy–Schwarz inequality, we know that

$$|\hat{a}_{S(a)} - a_{S(a)}|_1 \leq \sqrt{S(a)} \|\hat{a}_{S(a)} - a_{S(a)}\| \leq \sqrt{S(a) \kappa^{-1} (\hat{a} - a)^T G (\hat{a} - a)}.$$

Plugging this last inequality into (7.5), we deduce that

$$2\langle \hat{f} - f_a, \hat{f} - p_i \rangle_T + (\gamma - 2)d|\hat{a}_{S^c(a)}|_1 \leq (\gamma + 2)d\sqrt{S(a) \kappa^{-1} (\hat{a} - a)^T G (\hat{a} - a)}.$$

To conclude the proof of the first part, note that

$$\begin{cases} 2\langle \hat{f} - f_a, \hat{f} - p_i \rangle_T = \|\hat{f} - p_i\|_T^2 + \|\hat{f} - f_a\|_T^2 - \|f_a - p_i\|_T^2, \\ (\hat{a} - a)^T G (\hat{a} - a) = \|\hat{f} - f_a\|_T^2, \end{cases}$$

and use the inequality $qy - y^2 \leq q^2/4$, which is valid for any $q, y > 0$.

For the second part of the result, to control the fluctuations of $b_\varphi - \bar{b}_\varphi$, let us note that $b_\varphi - \bar{b}_\varphi = M_T$, where $(M_t)_{1 \leq t \leq T}$ is the martingale defined by

$$M_t = \sum_{i=1}^{t} \frac{\varphi(X_{-\infty:t-1})}{T} \big[X_{i,t} - p_i(X_{-\infty:t-1})\big].$$

We can apply the classical bound of Hoeffding's inequality on each increment of the martingale ΔM_t. Note that if $\varphi(X_{-\infty:t-1})$ is positive,

$$-\frac{\varphi(X_{-\infty:t-1})}{T} p_i(X_{-\infty:t-1}) \leq \Delta M_t \leq \frac{\varphi(X_{-\infty:t-1})}{T} \big[1 - p_i(X_{-\infty:t-1})\big],$$

and if $\varphi(X_{-\infty:t-1})$ is negative,

$$\frac{\varphi(X_{-\infty:t-1})}{T} \big[1 - p_i(X_{-\infty:t-1})\big] \leq \Delta M_t \leq -\frac{\varphi(X_{-\infty:t-1})}{T} p_i(X_{-\infty:t-1}).$$

This leads for every $\theta > 0$ to

$$\mathbb{E}\big(e^{\theta \Delta M_t} | X_{-\infty:t-1}\big) \leq \exp\left(\frac{\theta^2 \varphi(X_{-\infty:t-1})^2}{8T^2}\right) \leq \exp\left(\frac{\theta^2 \|\Phi\|^2}{8T^2}\right).$$

Therefore

$$\mathbb{E}\big(e^{\theta M_T}\big) \leq \exp\left(\frac{\theta^2 \|\Phi\|^2}{8T}\right).$$

Hence

$$\mathbb{P}(M_T \geq x) \leq \exp\left(\frac{\theta^2 \|\Phi\|_\infty^2}{8T} - \theta x\right).$$

By optimizing this in θ and applying the same inequality to $-\varphi$, we get for all positive u

$$\mathbb{P}\left(M_T \geq \sqrt{\frac{u\|\Phi\|_\infty^2}{2T}}\right) \leq e^{-u} \quad \text{and} \quad \mathbb{P}\left(|b_\varphi - \bar{b}_\varphi| \geq \sqrt{\frac{u\|\Phi\|_\infty^2}{2T}}\right) \leq 2e^{-u}$$

Therefore taking $u = \log|\Phi| + \log(2\delta^{-1})$ and then applying the union bound we obtain the result.

7.2. *Proof of Proposition 1*

Since $\{N_{(i,t)} > \ell\} = \{|A_{i,t}^\ell| \geq 1\}$, the Markov inequality implies that

$$\mathbb{P}(N_{i,t} > \ell) \leq \mathbb{E}\big[\big|A_{i,t}^\ell\big|\big].$$

So let us prove by induction that $\mathbb{E}[|A_{i,t}^\ell|] \leq (\bar{m})^\ell$ for all $\ell \geq 1$. For $\ell = 1$, we have $\mathbb{E}[|A_{i,t}^1|] = \mathbb{E}[|V_{i,t}|] = \bar{m}_i \leq \bar{m}$. Next for $\ell > 1$,

$$\begin{aligned}\mathbb{E}\big[\big|A_{i,t}^\ell\big|\big|A_{i,t}^{\ell-1}\big] &\leq \sum_{(j,s)\in A_{i,t}^{\ell-1}} \mathbb{E}\big[\big|V_{j,s}^{\to s}\big|\big] \\ &\leq \sum_{(j,s)\in C_{i,t}(\ell-1)} \bar{m}_j \leq \big|A_{i,t}^{\ell-1}\big|\,\bar{m}.\end{aligned}$$

To conclude the proof take the overall expectation and use the induction assumption given by $\mathbb{E}[|A_{i,t}^{\ell-1}|] \leq (\bar{m})^{\ell-1}$.

7.3. *Proof of Theorem 2*

For any fixed $n \geq 1$, for all site (i,t) let

$$G_{i,t}^n = \bigcup_{m=1}^n A_{i,t}^m$$

We adopt the convention that if $G_{i,t}^n = \varnothing$, $\mathbb{T}(G_{i,t}^n) = t$ and we consider the variable $T_{i,t}^n = t - \mathbb{T}(G_{i,t}^n)$ as well as its Laplace transform $\Psi_i^n(\theta) = \mathbb{E}(e^{\theta T_{i,t}^n})$.

Let us prove by induction that $\Psi_i^n(\theta)$ is finite and that

$$\Psi^n(\theta) = \sup_i \Psi_i^n(\theta) \leq \bar{\lambda}\big(1 + \varphi(\theta) + \cdots + \varphi(\theta)^{n-2}\big)\mathbf{1}_{n>1} + \varphi(\theta)^{n-1} g(\theta), \tag{7.6}$$

where $\bar{\lambda} = \sup_{i\in I} \lambda_i(\varnothing)$ and

$$g(\theta) = \sup_{i\in I} \sum_{v\in\mathcal{V}} e^{\theta T(v)} \lambda_i(v).$$

Note that $g(\theta)$ is finite as soon as $\varphi(\theta)$ is and that $0 \leq \bar{\lambda} \leq 1$.

For $n = 1$, since for all i, $\mathbb{T}(G_{i,t}^1) = \mathbb{T}(A_{i,t}^1) = \mathbb{T}(K_{i,t}) = t - T(V_{i,t})$

$$\begin{aligned}\Psi_i^1(\theta) &= \mathbb{E}\big(\exp\big[\theta T(V_{i,t})\big]\big) \\ &= \sum_{v\in\mathcal{V}} e^{\theta T(v)} \lambda_i(v) \\ &\leq g(\theta).\end{aligned}$$

Next by induction, let us assume (7.6) at level n for all i and let us prove it at level $n+1$. Note that because the $G_{i,t}^n$ are computed recursively, we have that when $K_{i,t}$ is not empty,

$$\mathbb{T}\big(G_{i,t}^{n+1}\big) = \min_{(k,r)\in K_{i,t}} \mathbb{T}\big(G_{k,r}^n\big).$$

Therefore if $K_{i,t} = \varnothing$, $T_{i,t}^{n+1} = 0$ and

$$\mathbb{E}\big(\exp\big[\theta T_{i,t}^{n+1}\big]|K_{i,t}\big) = 1.$$

This happens with probability $\lambda_j(\varnothing)$. If $K_{i,t} \neq \varnothing$,

$$\begin{aligned}\mathbb{E}\big(\exp\big[\theta\big(t - \mathbb{T}\big(G_{i,t}^{n+1}\big)\big)\big]|K_{j,t}\big) &= \mathbb{E}\Big(\exp\Big[\theta \max_{(k,r)\in K_{i,t}}\big(t - \mathbb{T}\big(G_{k,r}^n\big)\big)\Big]|K_{i,t}\Big)\\ &\leq \sum_{(k,r)\in K_{i,t}} e^{\theta(t-r)}\mathbb{E}\big(\exp\big[\theta\big(r - \mathbb{T}\big(G_{k,r}^n\big)\big)\big]|K_{i,t}\big).\end{aligned}$$

Since (see the algorithm) $K_{i,t}$ only depends on $U_{j,t}^1$ and $G_{k,r}^n$ only depends on the $U_{k',r'}^1$ for $k' \in I$, $r' \leq r$ and $r < t$, it follows that $\mathbb{T}(G_{k,r}^n)$ is independent of $K_{i,t}$. Hence if $K_{i,t} \neq \varnothing$

$$\begin{aligned}\mathbb{E}\big(\exp\big[\theta T_{i,t}^{n+1}\big]|K_{j,t}\big) &\leq \sum_{(k,r)\in K_{i,t}} e^{\theta(t-r)}\Psi_k^n(\theta)\\ &\leq \Big[\sum_{(k,r)\in K_{i,t}} e^{\theta(t-r)}\Big]\Psi^n(\theta)\\ &\leq \big[|K_{i,t}|e^{\theta(t-\mathbb{T}(K_{j,t}))}\big]\Psi^n(\theta)\\ &\leq |V_{i,t}|e^{\theta T(V_{i,t})}\Psi^n(\theta).\end{aligned}$$

We obtain by taking the overall expectation that

$$\Psi_i^{n+1}(\theta) \leq \bar{\lambda} + \varphi(\theta)\Psi^n(\theta),$$

so that $\sup_{i\in I}\Psi_i^{n+1}(\theta)$ is finite and (7.6) holds at level $n+1$ by induction.

To conclude, it is sufficient to remark that by the monotone convergence theorem, $\Psi_i^n(\theta) \to_{n\to\infty} \Psi_i(\theta)$ which are therefore upper bounded by $\bar{\lambda}/(1-\varphi(\theta))$. This concludes the proof.

7.4. *Proof of Lemma 1*

We use the perfect simulation algorithm to construct these chains. Let $\mathbf{U}^0 = (U_{i,t}^{0,1}, U_{i,t}^{0,2})_{i\in I, t\in\mathbb{Z}}, \ldots, \mathbf{U}^{2k+1} = (U_{i,t}^{2k+1,1}, U_{i,t}^{2k+1,2})_{i\in I,t\in\mathbb{Z}}$ be independent fields of independent random variables with uniform distribution on $[0,1]$. We assume that these sequences are defined in the same probability space and set $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ to be this common probability space.

The perfect simulation algorithm performed with the same field $\mathbf{U}^0$ on each site (i,t) yields the construction of $\mathbf{X} = (X_{i,t})_{i\in I,t\in\mathbb{Z}}$.

For any n, the chain $\mathbf{X}^n$ is also built similarly via the perfect simulation algorithm but with the field $\mathbf{U}^n$ except on a small portion of time where we use $\mathbf{U}^0$. More precisely, we use the following variables

$$\big(\big(U_{i,t}^{n,1}, U_{i,t}^{n,2}\big)_{i\in I, t\leq(n-2)B}, \big(U_{i,t}^{0,1}, U_{i,t}^{0,2}\big)_{i\in I,(n-2)B<t\leq nB}, \big(U_{i,t}^{n,1}, U_{i,t}^{n,2}\big)_{i\in I,t>nB}\big),$$

for $1 \leq n \leq 2k$ and for $n = 2k+1$,

$$\big(\big(U_{i,t}^{n,1}, U_{i,t}^{n,2}\big)_{i\in I, t\leq(2k-1)B}, \big(U_{i,t}^{0,1}, U_{i,t}^{0,2}\big)_{i\in I,(2k-1)B<t\leq T}, \big(U_{i,t}^{n,1}, U_{i,t}^{n,2}\big)_{i\in I,t>T}\big).$$

Since all chains are simulated with the same set of weights $(\lambda_i)_{i\in I}$ and transitions $(p_i^v)_{i\in I, v\in\mathcal{V}}$, they have obviously the same distribution. Since the algorithms use disjoint sets of uniform variables for the odd (resp. even) chains, they are obviously independent and therefore Items 1–3 follows easily from the construction.

Let $G_{i,t}$ be the genealogy of site (i,t) in the chain $\mathbf{X}$ and $\mathbb{T}_{i,t} = \mathbb{T}(G_{i,t})$. For any n, any $i \in F$ and any $t \in I_n$, if $\mathbb{T}_{i,t} > (n-2)B$, then we use exactly the same set of uniform variables to produce the values of $X_{i,t}$ and $X^n_{i,t}$ and their values are equal.

Therefore on $\Omega_{\text{good}} = \bigcap_{i\in F}\bigcap_{n=1}^{2k+1}\bigcap_{t\in I_n}\{\mathbb{T}_{i,t} > (n-2)B\}$, $X_{F,I_n} = X^n_{F,I_n}$ for all $n = 1,\dots,2k+1$. Note that Ω_{good} only depends on $\mathbf{X}$.

It remains to control $\tilde{\mathbb{P}}(\Omega^c_{\text{good}})$. By a union bound, and the application of Theorem 2, we obtain

$$\begin{aligned}
\tilde{\mathbb{P}}\big(\Omega^c_{\text{good}}\big) &\le \sum_{i\in F}\sum_{n=1}^{2k+1}\sum_{t\in I_n}\mathbb{P}\big(\mathbb{T}_{i,t} \le (n-2)B\big)\\
&\le \sum_{i\in F}\sum_{n=1}^{2k+1}\sum_{t\in I_n}\mathbb{P}\big(t - \mathbb{T}_{i,t} \ge t - (n-2)B\big)\\
&\le \sum_{i\in F}\sum_{n=1}^{2k+1}\sum_{t\in I_n} e^{-\theta(t-(n-2)B)}\Psi(\theta)\\
&\le |F|(2k+1)\frac{e^{-\theta(B-m+1)}}{1-e^{-\theta}}\Psi(\theta).
\end{aligned}$$

In particular if we choose $B = m + \theta^{-1}(2\log(T) + \log(|F|)$,

$$\tilde{\mathbb{P}}\big(\Omega^c_{\text{good}}\big) \le \frac{2k+1}{T^2}\frac{\Psi(\theta)}{1-e^{-\theta}},$$

which concludes the proof.

7.5. *Proof of Theorem 3*

Take $B = m + \theta^{-1}(2\log(T) + \log(|F|))$, $k = \lfloor \frac{T}{2B} \rfloor$ and use the probability space $(\tilde\Omega, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ and the stochastic chains $\mathbf{X},\dots,\mathbf{X}^{2k+1}$ given by Lemma 1. By Lemma 1-Item 1 we can assume that Z is also defined on $(\tilde\Omega, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$. Define also a partition $J_1,\dots,J_{2k+1}$ of $1:T$ as follows:

$$J_n = \big\{1 + (n-1)B,\dots,nB\big\} \quad \text{for } 1 \le n \le 2k, \quad \text{and} \quad J_{2k+1} = \{1 + 2kB,\dots,T\}.$$

For each $1 \le n \le 2k+1$, write $S_n = \frac{1}{T}\sum_{t\in J_n} f(X^n_{F,t-m:t-1})$ and note that S_n only depends on the t's in I_n as defined in Lemma 1. Since $|J_n| \le B$ for all $1 \le n \le 2k+1$, it holds $|S_n| \le MB/T$.

Observe that Lemma 1-Item 1 and 4 ensure that on Ω_{good},

$$Z = \sum_{n=1}^{2k+1}\big(S_n - \mathbb{E}(S_n)\big),$$

so that for any $w > 0$, we have

$$\begin{aligned}
\tilde{\mathbb{P}}(Z > w) &\le \tilde{\mathbb{P}}\big(\Omega^c_{\text{good}}\big) + \tilde{\mathbb{P}}\left(\sum_{n=1}^{2k+1}\big(S_n - \mathbb{E}(S_n)\big) > w\right)\\
&\le \frac{c'(\theta)}{T} + \tilde{\mathbb{P}}\left(\sum_{n=1}^{2k+1}\big(S_n - \mathbb{E}(S_n)\big) > w\right).
\end{aligned}$$

Moreover, if we denote $Z_1 = \sum_{n=1}^{k+1}(S_{2n-1} - \mathbb{E}(S_{2n-1}))$ and $Z_2 = \sum_{n=1}^{k}(S_{2n} - \mathbb{E}(S_{2n}))$, then

$$\tilde{\mathbb{P}}\left(\sum_{n=1}^{2k+1}\big(S_n - \mathbb{E}(S_n)\big) > u + v\right) \le \tilde{\mathbb{P}}(Z_1 > u) + \tilde{\mathbb{P}}(Z_2 > v),$$

for all $u + v = w$.

Lemma 1-Item 3 implies that $S_2, \dots, S_{2k}$ are independent, so that by the classical Hoeffding inequality, we have for any $x > 0$, $\tilde{\mathbb{P}}(Z_1 > \sqrt{kB^2M^2T^{-2}x/2}) \le e^{-x}$, and similarly for $\tilde{\mathbb{P}}(Z_1 > \sqrt{(k+1)B^2M^2T^{-2}x/2}) \le e^{-x}$. Hence

$$\tilde{\mathbb{P}}\big(Z > \sqrt{kB^2M^2T^{-2}x/2} + \sqrt{(k+1)B^2M^2T^{-2}x/2}\big) \le \frac{c'(\theta)}{T} + 2e^{-x}.$$

But $k \le T(2B)^{-1}$ and $k+1 \le (T+2B)(2B)^{-1} \le T/B$. This leads directly to the first result.

For the second result, note that we can restrict ourselves to Ω_{good} once and for all at the beginning and use the union bound only on the auxiliary independent chains, which explains why we pay $|\mathcal{F}|$ only in front of the deviation e^{-x}.

7.6. *Proof of Theorem 4*

Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ be the probability space and $\mathbf{X}, \dots, \mathbf{X}^{2k+1}$ be the stochastic chains given by Lemma 1. By Lemma 1-Item 1 we can assume that Z is also defined on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$. We write $\tilde{\mathbb{E}}$ to denote the expectation taken with respect the probability measure $\tilde{\mathbb{P}}$.

Now, let B, k, $J_1, \dots, J_{2k+1}$ as in the proof of Theorem 3 and define for $1 \le n \le 2k+1$, the random matrix $\Sigma_n = ((\Sigma_n(f,g))_{f,g\in\mathcal{F}}$ as follows:

$$\Sigma_n(f,g) = \frac{1}{T}\sum_{t\in J_n}\big(f\big(X^n_{F,t-m:t-1}\big)g\big(X^n_{F,t-m:t-1}\big) - \mathbb{E}\big(f\big(X^n_{F,t-m:t-1}\big)g\big(X^n_{F,t-m:t-1}\big)\big)\big).$$

Clearly $\tilde{\mathbb{E}}(\Sigma_n) = 0$. To apply Theorem 1.3 of Tropp [30], we need to find a deterministic self-adjoint matrix A_n such that $A_n^2 - \Sigma_n^2$ is non negative. This means that for all vector $x \in \mathbb{R}^{\mathcal{F}}$,

$$x^\intercal\big[A_n^2 - \Sigma_n^2\big]x \ge 0.$$

By taking $A_n = \sigma I_n$, it is sufficient to prove that

$$x^\intercal \Sigma_n^2 x \le \sigma^2\|x\|^2.$$

But

$$\begin{aligned}
x^\intercal\Sigma_n^2x &= \sum_{f,g\in\mathcal{F}} x_f x_g \frac{1}{T^2}\sum_{t,t'\in J_n}\sum_{h\in\mathcal{F}}\big(f\big(X^n_{F,t-m:t-1}\big)h\big(X^n_{F,t-m:t-1}\big) - \mathbb{E}\big(f\big(X^n_{F,t-m:t-1}\big)h\big(X^n_{F,t-m:t-1}\big)\big)\\
&\quad\times\big(h\big(X^n_{F,t'-m:t'-1}\big)g\big(X^n_{F,t'-m:t'-1}\big) - \mathbb{E}\big(h\big(X^n_{F,t'-m:t'-1}\big)g\big(X^n_{F,t'-m:t'-1}\big)\big)\\
&= \frac{1}{T^2}\sum_{t,t'\in J_n}\sum_{h\in\mathcal{F}}\Big[\sum_f x_f\big(f\big(X^n_{F,t-m:t-1}\big)h\big(X^n_{F,t-m:t-1}\big) - \mathbb{E}\big(f\big(X^n_{F,t-m:t-1}\big)h\big(X^n_{F,t-m:t-1}\big)\big)\Big]\\
&\quad\times\Big[\sum_g x_g\big(g\big(X^n_{F,t'-m:t'-1}\big)h\big(X^n_{F,t'-m:t'-1}\big) - \mathbb{E}\big(g\big(X^n_{F,t'-m:t'-1}\big)h\big(X^n_{F,t'-m:t'-1}\big)\big)\Big]\\
&\le \frac{1}{T^2}\sum_{t,t'\in J_n}\sum_{h\in\mathcal{F}}\|x\|^2\sqrt{\sum_f\big(f\big(X^n_{F,t-m:t-1}\big)h\big(X^n_{F,t-m:t-1}\big) - \mathbb{E}\big(f\big(X^n_{F,t-m:t-1}\big)h\big(X^n_{F,t-m:t-1}\big)\big)^2}\\
&\quad\times\sqrt{\sum_g\big(g\big(X^n_{F,t'-m:t'-1}\big)h\big(X^n_{F,t'-m:t'-1}\big) - \mathbb{E}\big(g\big(X^n_{F,t'-m:t'-1}\big)h\big(X^n_{F,t'-m:t'-1}\big)\big)^2}\\
&\le \frac{4\|x\|^2|\mathcal{F}|}{T^2}\sum_{t,t'\in J_n}\sum_{h\in\mathcal{F}}M^4\\
&\le \frac{4|\mathcal{F}|^2B^2M^4}{T^2}\|x\|^2.
\end{aligned}$$

Hence $\sigma = \frac{2|\mathcal{F}|BM^2}{T}$ works. Denote $Z_1 = \sum_{n=1}^{k+1}\Sigma_{2n-1}$ and $Z_2 = \sum_{n=1}^{k}\Sigma_{2n}$. Lemma 1 implies that on Ω_{good},

$$Z = Z_1 + Z_2,$$

so that by the triangle inequality we have for any $u > 0$ and $v > 0$,

$$\tilde{\mathbb{P}}\big(\|Z\| > u + v\big) \le \tilde{\mathbb{P}}\big(\Omega_{\text{good}}^c\big) + \tilde{\mathbb{P}}\big(\|Z_1\| > u\big) + \tilde{\mathbb{P}}\big(\|Z_2\| > v\big).$$

Since by Lemma 1-Item 3, $\Sigma_2, \Sigma_4, \dots, \Sigma_{2k}$ are i.i.d. random matrices, we can apply Theorem 1.3 of Tropp [30] to deduce that for any $v > 0$,

$$\tilde{\mathbb{P}}\big(\|Z_2\| > \sqrt{8k\sigma^2 v}\big) \le 2|\mathcal{F}|e^{-v},$$

Similarly, we have that for any $x > 0$,

$$\tilde{\mathbb{P}}\big(\|Z_2\| > \sqrt{8(k+1)\sigma^2 u}\big) \le 2|\mathcal{F}|e^{-u},$$

and as a consequence, it follows that for any $x > 0$,

$$\tilde{\mathbb{P}}\big(\|Z\| > \sqrt{8k\sigma^2 x} + \sqrt{8(k+1)\sigma^2 x}\big) \le \frac{c'(\theta)}{T} + 4|\mathcal{F}|e^{-x}.$$

Since $k^{1/2} + (k+1)^{1/2} \le (4T/B)^{1/2}$, the result follows from the inequality above.

7.7. *Proof of Lemma 2*

The proof is done for the lower bound. The argument is similar for the upper bound. We use induction on the time length of v. If $v = \varnothing$, f is constant and $\mathbb{E}(f(X_v)) = \mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(f(X_v))$. Let $Q = \mathcal{B}(1/2)^{\otimes v}$.

If the time length of v is strictly positive, let t be the maximal time of v and let $w_t = \{(i,t) \text{ for } i \text{ such that } (i,t) \in v\}$.

$$\begin{aligned}
\mathbb{E}\big(f(X_v)\big) &= \mathbb{E}\big[\mathbb{E}\big(f(X_v)|X_{-\infty:t-1}\big)\big] \\
&= \mathbb{E}\Bigg(\sum_{x_{w_t}\in\{0,1\}^{w_t}} f\big((X_{v\setminus w_t}, x_{w_t})\big)\mathbb{P}(X_{w_t} = x_{w_t}|X_{-\infty:t-1})\Bigg) \\
&= \mathbb{E}\Bigg(\sum_{x_{w_t}\in\{0,1\}^{w_t}} f\big((X_{v\setminus w_t}, x_{w_t})\big)\prod_{i/(i,t)\in w_t}\mathbb{P}(X_{i,t} = x_{i,t}|X_{-\infty:t-1})\Bigg) \\
&\ge (2\mu)^{|w_t|}\mathbb{E}\Bigg(\sum_{x_{w_t}\in\{0,1\}^{w_t}} f\big((X_{v\setminus w_t}, x_{w_t})\big)Q(X_{i,t} = x_{i,t})\Bigg)
\end{aligned}$$

But $\sum_{x_{w_t}\in\{0,1\}^{w_t}} f((X_{v\setminus w_t}), x_{w_t}))Q(X_{i,t} = x_{i,t})$ is a cylindrical function on $v \setminus w_t$ with time length strictly smaller than v, so by induction,

$$\begin{aligned}
\mathbb{E}\big(f(X_v)\big) &\ge (2\mu)^{|w_t|}(2\mu)^{|v\setminus w_t|}\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}\Bigg(\sum_{x_{w_t}\in\{0,1\}^{w_t}} f\big((X_{v\setminus w_t}), x_{w_t}\big)\Bigg)Q(X_{i,t} = x_{i,t})) \\
&\ge (2\mu)^{|v|}\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}\big(f(X_v)\big),
\end{aligned}$$

which concludes the proof.

7.8. *Proof of Theorem 5*

For any $a \in \mathbb{R}^{\Phi}$ such that $\|a\| = 1$, we have by Cauchy–Schwarz inequality

$$\kappa \le a^{\intercal}\mathbb{E}(G)a \le a^{\intercal}Ga + \|a\|\big\|\big(G - \mathbb{E}(G)\big)a\big\| \le a^{\intercal}Ga + \big\|G - \mathbb{E}(G)\big\|, \tag{7.7}$$

so that the result follows from Theorem 4 with $x = \log(4|\mathcal{F}|/\delta)$ and $\mathcal{F} = \Phi$.

7.9. *Proof of Theorem 6*

First of all, remark that thanks to Lemma 2 and since φ in this case depends on a neighborhood of size 1, one has that

$$\mathbb{E}(G_{\varphi,\varphi}) = \mathbb{E}\big(\varphi(X)^2\big) \geq 2\mu 1/2 = \mu$$

and similarly for $\varphi \neq \varphi'$, $\varphi\varphi'$ is positive and depends on a neighborhood of size 2, hence

$$(1-\mu)^2 \geq \mathbb{E}(G_{\varphi,\varphi'}) \geq \mu^2.$$

Moreover let us apply our version of Hoeffding's inequality, i.e. the second result of Theorem 3 on all the $\varphi^2 = \varphi$, $\varphi\varphi'$ and $-\varphi\varphi'$ for $\varphi \neq \varphi'$. Hence there exists and event of probability larger than $1 - \frac{c'(\theta)}{T} - \delta$ such that for all φ, φ',

$$\big|G_{\varphi,\varphi'} - \mathbb{E}(G_{\varphi,\varphi'})\big| \leq R_T,$$

with

$$R_T = \sqrt{c''(\theta)\frac{(m + \log T + \log|F|)}{T}\log\left(\frac{4|\Phi|^2}{\delta}\right)},$$

which means that there exists a constant c_1 depending only on the distribution such that for T large enough (depending on θ and $|F|$)

$$R_T = c_1 T^{-1}\big(m + \log T + \log|F|\big)^{1/2}\big(\log m + \log|F| + \log\delta^{-1}\big)^{1/2}.$$

Therefore on this event, for all a and J such that $|J| \leq s$ and $|a_{J^c}|_1 \leq c|a_J|_1$, and if $\mu^2 \geq R_T$,

$$\begin{aligned}
a^\intercal G a &= \sum_{\varphi\in\Phi} a_\varphi^2 G_{\varphi,\varphi} + \sum_{\varphi\neq\varphi'\in\Phi} a_\varphi a_{\varphi'} G_{\varphi,\varphi'} \\
&\geq (\mu - R_T)\sum_{\varphi\in\Phi} a_\varphi^2 + \big(\mu^2 - R_T\big)\sum_{\substack{\varphi\neq\varphi'\in\Phi \\ a_\varphi a_{\varphi'}\geq 0}} a_\varphi a_{\varphi'} + \big((1-\mu)^2 + R_T\big)\sum_{\substack{\varphi\neq\varphi'\in\Phi \\ a_\varphi a_{\varphi'}<0}} a_\varphi a_{\varphi'} \\
&\geq \big(\mu - \mu^2\big)\|a\|^2 + \mu^2\sum_{\varphi,\varphi'\in\Phi} a_\varphi a_{\varphi'} + (1-2\mu)\sum_{\substack{\varphi\neq\varphi'\in\Phi \\ a_\varphi a_{\varphi'}<0}} a_\varphi a_{\varphi'} - R_T|a|_1^2 \\
&\geq \big(\mu - \mu^2\big)\|a\|^2 + \mu^2\left(\sum_{\varphi\in\Phi} a_\varphi\right)^2 - \big((1-2\mu) - R_T\big)|a|_1^2 \\
&\geq \big(\mu - \mu^2\big)\|a\|^2 - \big((1-2\mu) + R_T\big)\big[|a_J|_1 + |a_{J^c}|_1\big]^2 \\
&\geq \big(\mu - \mu^2\big)\|a\|^2 - \big((1-2\mu) + R_T\big))(1+c)^2|a_J|_1^2 \\
&\geq \big(\mu - \mu^2\big)\|a_J\|^2 - \big((1-2\mu) + R_T\big)(1+c)s\|a_J\|^2,
\end{aligned}$$

which is the desired result.

7.10. *Proof of Corollary 1*

We shall prove only for the short effect dictionary. The other cases are treated similarly. For this choice of dictionary $\|\Phi\|_\infty = 1$ and $|\Phi| = |F|$. Hence, by applying Theorem 5 and Theorem 1 both with $\delta = T^{-1}$ one deduces that, for T large enough, on an event of probability larger then $1 - c_1/T$, the following oracle inequality holds

$$\big\|\hat{f} - p_i(\cdot)\big\|_T^2 \leq \inf_{a\in\mathbb{R}^\Phi}\left\{\big\|f_a - p_i(\cdot)\big\|_T^2 + 4\kappa^{-1}\big|S(a)\big|\frac{(\log|F| + \log(2T))}{2T}\right\}, \tag{7.8}$$

where c_1 depends only on the distribution of $\mathbf{X}$ and

$$\kappa = \kappa' - c_1' T^{-1/2}|F|^{1/2}\big(m + \log(T) + \log|F|\big)^{1/2}\big(\log|F| + \log\delta^{-1}\big)^{1/2},$$

with c_1' depending only on the distribution of $\mathbf{X}$ and κ' given by (5.1).

Now, for the choices given by (5.2), (5.3) and (5.4), then, as seen previously $\kappa' = c_2' \log(T))^{-c_3'}$, for positive constants c_2' and c_3' depending only on m and μ and

$$\kappa = \frac{c_2'}{(\log T)^{c_3'}}\big(1 - o(1)\big).$$

By plugging κ into (7.8), the result follows.

Acknowledgements

This research has been conducted as part of FAPESP project *Research, Innovation and Dissemination Center for Neuromathematics* (grant 2013/07699-0). This work was also supported by the French government, through the UCA[Jedi] "Investissements d'Avenir" managed by the National Research Agency (ANR-15-IDEX-01), by the structuring program *C@UCA* of Université Côte d'Azur and by the interdisciplinary axis MTC-NSC of the University of Nice Sophia-Antipolis. We would like to thank anonymous referees for their helpful comments that improved the manuscript.

References

[1] S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **45** (2015) 1535–1567. MR3357870 https://doi.org/10.1214/15-AOS1315
[2] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin, 2011. MR2807761 https://doi.org/10.1007/978-3-642-20192-9
[3] E. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** (2005) 4203–4215. MR2243152 https://doi.org/10.1109/TIT.2005.858979
[4] S. Chen, A. Shojaie, E. Shea-Brown and D. Witten. The multivariate Hawkes process in high dimensions: Beyond mutual excitation. Preprint Arxiv, 2017.
[5] J. Chevallier. Mean-field limit of generalized Hawkes processes. *Stochastic Process. Appl.* **127** (2015) 3870–3912. MR3718099 https://doi.org/10.1016/j.spa.2017.02.012
[6] R. Cofré and B. Cessac. Exact computation of the maximum-entropy potential of spiking neural-network models. *Phys. Rev. E* **89** (2014).
[7] F. Comets, R. Fernandez and P. A. Ferrari. Processes with long memory: Regenerative construction and perfect simulation. *Ann. Appl. Probab.* **12** (2002) 921–943. MR1925446 https://doi.org/10.1214/aoap/1031863175
[8] A. Duarte, A. Galves, E. Löcherbach and G. Ost. Estimating the interaction graph of stochastic neural dynamics. *Bernoulli*. **25** (2019) 771–792. MR3892336 https://doi.org/10.3150/17-bej1006
[9] R. Fernandéz, P. Ferrari and A. Galves. Coupling, renewal and perfect simulation of chains of infinite order, 2001.
[10] S. Gaïffas and A. Guilloux. High-dimensional additive hazards models and the Lasso. *Electron. J. Stat.* **6** (2012) 522–546. MR2988418 https://doi.org/10.1214/12-EJS681
[11] S. Gaïffas and G. Matulewicz. Sparse inference of the drift of a high-dimensional Ornstein–Uhlenbeck process. *J. Multivariate Anal.* **169** (2019) 1–20. MR3875583 https://doi.org/10.1016/j.jmva.2018.08.005
[12] A. Galves, N. L. Garcia, E. Löcherbach and E. Orlandi. Kalikow-type decomposition for multicolor infinite range particle systems. *Ann. Appl. Probab.* **23** (2013) 1629–1659. MR3098444 https://doi.org/10.1214/12-aap882
[13] A. Galves and E. Löcherbach. Infinite systems of interacting chains with memory of variable length – a stochastic model for biological neural nets. *J. Stat. Phys.* **151** (2013) 896–921. MR3055382 https://doi.org/10.1007/s10955-013-0733-9
[14] A. Galves and E. Löcherbach. Modeling networks of spiking neurons as interacting processes with memory of variable length. *Journal de la Société Française de Statistiques* **157** (2016) 17–32. MR3491721
[15] N. R. Hansen, P. Reynaud-Bouret and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli* **21** (2015) 83–143. MR3322314 https://doi.org/10.3150/13-BEJ562
[16] P. Hodara and E. Löcherbach. Hawkes Processes with variable length memory and an infinite number of components. *Adv. Appl. Probab.* **49** (2017) 84–107. MR3631217 https://doi.org/10.1017/apr.2016.80
[17] X. J. Hunt, P. Reynaud-Bouret, V. Rivoirard, L. Sansonnet and R. Willett. A data-dependent weighted LASSO under Poisson noise. *IEEE Transactions on Information Theory.* **65** (2018) 1589–1613. MR3923187 https://doi.org/10.1109/TIT.2018.2869578
[18] X. Jiang, G. Raskutti and R. Willett. Minimax optimal rates for Poisson inverse problems with physical constraints. *IEEE Trans. Inform. Theory* **61** (2015) 4458–4474. MR3372365 https://doi.org/10.1109/TIT.2015.2441072
[19] S. Kalikow. Random Markov processes and uniform martingales. *Israel J. Math.* **71** (1990) 33–54. MR1074503 https://doi.org/10.1007/BF02807249
[20] R. C. Kelly, M. A. Smith, R. E. Kass and T. S. Lee. Accounting for network effects in neuronal responses using L1 regularized point process models. *NIPS – Adv. Neural Inf. Process. Syst.* **23** (2010) 1099–1107.
[21] A. B. Kock and L. Callot. Oracle inequalities for high dimensional vector autoregressions. *J. Econometrics* **186** (2015) 325–344. MR3343790 https://doi.org/10.1016/j.jeconom.2015.02.013
[22] M. Lerasle and D. Y. Takahashi. Sharp oracle inequalities and slope heuristic for specification probabilities estimation in general random fields. *Bernoulli* **22** (2016) 325–344. MR3449785 https://doi.org/10.3150/14-BEJ660
[23] B. Mark, G. Raskutti and R. Willett. Network estimation from point process data. Preprint Arxiv, 2019. MR3951378 https://doi.org/10.1109/TIT.2018.2875766
[24] B. Mark, G. Raskutti and R. Willett. Estimating network structure from incomplete event data. Oral presentation AISTATS, 2019.

[25] C. Pouzat and A. Chaffiol. Automatic spike train analysis and report generation. An implementation with R, R2HTML and STAR. *J. Neurosci. Meth.* (2009).

[26] P. Reynaud-Bouret and E. Roy. Some non asymptotic tail estimates for Hawkes processes. *Bull. Belg. Math. Soc. Simon Stevin* **13** (2007) 883–896. MR2293215

[27] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.* **61** (2008) 1025–1045. MR2417886 https://doi.org/10.1002/cpa.20227

[28] L. Sacerdote and M. T. Giraudo. *Stochastic Integrate and Fire Models: A Review on Mathematical Methods and Their Applications*. *Lecture Notes in Mathematics* **2058**, 99–148. Springer, Berlin, 2013. MR3051031 https://doi.org/10.1007/978-3-642-32157-3_5

[29] J. Tropp. *Sampling Theory, a Renaissance: Compressive Sampling and Other Developments. Convex Recovery of a Structured Signal from Independent Random Linear Measurements*. *G. Pfander. Ser. Applied and Numerical Harmonic Analysis.* Birkhaeuser, Basel, 2015. MR3467419

[30] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** (2012) 389–434. MR2946459 https://doi.org/10.1007/s10208-011-9099-z

[31] S. van de Geer. *Estimation and Testing Under Sparsity. École d'été de Saint-Flour XLV*. Springer, Berlin, 2016. MR3526202 https://doi.org/10.1007/978-3-319-32774-7

[32] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** (2009) 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

[33] G. Viennet. Inequalities for absolutely regular sequences: Application to density estimation. *Probab. Theory Related Fields* **107** (1997) 467–492. MR1440142 https://doi.org/10.1007/s004400050094