Bayesian estimation of nonlinear Hawkes processes

Déborah Sulem, Vincent Rivoirard, Judith Rousseau

Submitted to Bernoulli

Bayesian estimation of nonlinear Hawkes processes

DÉBORAH SULEM [1], VINCENT RIVOIRARD [2,†] and JUDITH ROUSSEAU [1,*]

[1] *University of Oxford, E-mail:* deborah.sulem@stats.ox.ac.uk; [*] judith.rousseau@stats.ox.ac.uk
[2] *Ceremade, CNRS, UMR 7534, Université Paris-Dauphine, PSL University, 75016 Paris, France.*
E-mail: [†] Vincent.Rivoirard@dauphine.fr

Multivariate point processes (MPPs) are widely applied to model the occurrences of events, e.g., natural disasters, online message exchanges, financial transactions or neuronal spike trains. In the Hawkes process model, the probability of occurrences of future events depend on the past of the process. This model is particularly popular for modelling interactive phenomena such as disease expansion. In this work we consider the nonlinear multivariate Hawkes model, which allows to account for *excitation* and *inhibition* between interacting entities. We provide theoretical guarantees for applying nonparametric Bayesian estimation methods in this context. In particular, we obtain concentration rates of the posterior distribution on the parameters, under mild assumptions on the prior distribution and the model. These results also lead to convergence rates of Bayesian estimators. Another object of interest in event-data modelling is to infer the *graph of interaction* - or Granger causal graph. In this case, we provide consistency guarantees; in particular, we prove that the posterior distribution is consistent on the graph adjacency matrix of the process, as well as a Bayesian estimator based on an adequate loss function.

MSC2020 subject classifications: Primary 62G20, 62G05; secondary 60G65
Keywords: nonlinear Hawkes processes; nonparametric Bayesian inference; Granger-causal Graph

1. Introduction

1.1. Nonlinear Hawkes processes

The Hawkes model is a popular temporal point process (PP) for modelling the occurrences of event-type phenomena. Extending the Poisson cluster process (Møller and Rasmussen, 2005), this model allows the probability of occurrence of a new event to depend on the history of the process. The first construction by Hawkes (1971) aimed at modelling the *self-excitatory* behaviour of earthquakes' strikes with aftershocks, and is called the *linear Hawkes process*. Since then, it has been extensively used, partly due to its interpretable parameters and branching structure representation (Reynaud-Bouret and Roy, 2007). This notably leads to tractable inference and simulation methods (Bacry et al., 2020; Chen, Witten and Shojaie, 2017; Hansen, Reynaud-Bouret and Rivoirard, 2015).

Hawkes processes have been largely and successfully applied in various contexts of correlated event-data, including online social popularity (Farajtabar et al., 2015), stock prices moves (Embrechts, Liniger and Lin, 2011), topic modelling (Du et al., 2015), DNA motifs occurrences (Carstensen et al., 2010; Gusto and Schbath, 2005; Reynaud-Bouret and Schbath, 2010), and neuronal activity modelling (Chornoboy, Schramm and Karr, 1988; Lambert et al., 2017; Reynaud-Bouret et al., 2014). They are used to infer both diffusion phenomena on networks and the structure of time-dependent networks (Miscouridou, Caron and Teh, 2018). Related and extended models include the mutually-regressive PP (Apostolopoulou et al., 2019), the age-dependent (Raad, Ditlevsen and Löcherbach, 2020) and marked (Karabash and Zhu, 2015) Hawkes processes, the dynamic contagion process (Dassios and

Zhao, 2011), the reactive PP (Ertekin, Rudin and McCormick, 2015), the self-correcting PP (Isham and Westcott, 1979) and the Dirichlet-Hawkes process (Du et al., 2015). More recently, neural point processes inspired by the Hawkes model have also been proposed, e.g., by Du et al. (2016); Mei and Eisner (2017).

In a multivariate temporal PP, each dimension represents an entity, a location or a type of event - it is equivalent to a *marked* point process with finite mark space. For $K \in \mathbb{N}\backslash\{0\}$, the PP can be described as a counting process $N = (N_t)_t = (N_t^1, \dots, N_t^K)_{t\geqslant 0}$, where N_t^k denotes the number of events that have occurred until time t at location k. Its dynamics are characterised by a conditional intensity function $(\lambda_t)_t = (\lambda_t^1, \dots, \lambda_t^K)_{t\geqslant 0}$, which is informally the infinitesimal rate of event conditionally on the past of the process, i.e, for $k = 1, \dots, K$, $\lambda_t^k dt = \mathbb{P}\left[N_t^k \text{ has a jump in } [t, t+dt] | \mathcal{G}_t\right]$, where $\mathcal{G}_t$ is the history of the process up to time t. In the nonlinear Hawkes model, only one dimension N^k of the process can jump at each time t and the intensity process has the following form

$$\lambda_t^k = \phi_k\left(\nu_k + \sum_{l=1}^{K} \int_{-\infty}^{t^-} h_{lk}(t-s) dN_s^l\right), \quad k = 1, \dots, K. \tag{1}$$

In (1), the parameter $\nu_k > 0$ denotes the *background* - or *spontaneous* - rate of events, and models exogeneous influences. The endogenous effects on the process are parametrised by *interaction functions* $(h_{lk})_{l,k=1}^K$ - or *triggering kernels*. More precisely, for $(l,k) \in [K]^2$, the function $h_{lk} : \mathbb{R} \to \mathbb{R}$ models the influence of component N^l onto component N^k. It can be decomposed into an *excitating* contribution $(h_{lk}^+ = \max(h_{lk}, 0))$ and an *inhibiting* contribution $(h_{lk}^- = \max(-h_{lk}, 0))$. Finally, the *link* or *activation function* $\phi_k : \mathbb{R} \to \mathbb{R}^+$ ensures that the intensity is a non-negative process, and is generally chosen to be monotone non-decreasing. If all the interaction functions h_{lk} are non-negative and all the link functions equal the identity functions, (1) corresponds to the linear Hawkes model.

The dependence on past events in the intensity (1) leads to a notion of *causality*. For Hawkes processes, a Granger-causal relationship between two components of the process corresponds to a non-null interaction function (Eichler, Dahlhaus and Dueck, 2017). We can define the *connectivity graph* parameter $\delta \in \{0,1\}^{K^2}$ such that for each (l,k), $\delta_{lk} = 1$ if the function h_{lk} in (1) is non null and $\delta_{lk} = 0$ otherwise. We note that this parameter is redundant with $(h_{lk})_{l,k=1}^K$.

To the best of our knowledge, the estimation of the parameters of nonlinear Hawkes processes $\nu = (\nu_k)_k$, $h = (h_{lk})_{l,k=1}^K$, $\delta = (\delta_{lk})_{l,k=1}^K$ - as well as additional parameters of the link functions $(\phi_k)_k$ has not been theoretically analysed, neither in the frequentist nor in the Bayesian frameworks. In the nonparametric setting, the existing results apply to linear Hawkes processes for the estimation of (ν, h) (Donnet, Rivoirard and Rousseau, 2020) and for the estimation of the connectivity graph δ (Hansen, Reynaud-Bouret and Rivoirard, 2015; Chen, Witten and Shojaie, 2017). In the nonlinear model, Chen et al. (2017) study the estimation of the cross-covariances of the process, and Wang et al. (2016) estimate a piecewise-constant link function assuming a parametric form on the interaction functions.

In this work, we analyse the theoretical properties of Bayesian methods for estimating ν, h, δ and additional parameters of the nonlinear functions $(\phi_k)_k$. We consider a prior distribution on the parameters, say Π, and our aim is to study posterior concentration rates in such models. More precisely, we wish to determine $\epsilon_T = o(1)$ and conditions on the model and on Π such that

$$\mathbb{E}_{f_0}[\Pi(d(f, f_0) > \epsilon_T | N)] \xrightarrow[T\to\infty]{} 1,$$

where $f = (\nu, h)$, $d(.,.)$ is some loss function on the parameter space, and $\Pi(.|N)$ denotes the posterior distribution given an observation of the process on $[0, T]$. In the last equation, we assume that the data N is generated by a Hawkes process with *true parameter* f_0, and we denote $\mathbb{P}_{f_0}$ its generating distribution

and $\mathbb{E}_{f_0}$ the associated expectation. In particular, a consequence of such result is the construction of estimators on ν, h which converge in the frequentist sense at the rate ϵ_T. We also obtain posterior consistency results on the graph parameter δ, and construct a consistent risk-minimising estimator.

1.2. Related works

There is a rich literature on Hawkes processes in probability, statistics, and more recently in machine learning and deep learning. The stability properties of the nonlinear Hawkes model have been studied under several assumptions (Brémaud and Massoulié, 1996; Karabash, 2012), together with the rate of convergence to the stationary solution (Brémaud, Nappo and Torrisi, 2002) and the Bartlett spectrum (Massoulié, 1998). Regenerative properties of Hawkes processes were investigated for the models with finite (Costa et al., 2020) and infinite (Graham, 2021; Raad, 2019) memory. Recently Bacry et al. (2013); Gao and Zhu (2018a,b) derived functional central limit theorems and large deviations principles for ergodic processes. Malliavin-Stein calculus was applied by Torrisi (2016, 2017) to establish Gaussian and Poisson approximations of functionals of the linear Hawkes process, and later by Hillairet et al. (2021) to obtain Berry-Esséen bounds. Stationary distributions of high dimensional Hawkes processes were also studied, notably in the mean-field limit (Delattre and Fournier, 2016; Delattre, Fournier and Hoffmann, 2016; Raad, Ditlevsen and Löcherbach, 2020).

Many statistical works have been dedicated to designing robust and efficient estimation procedures in the linear Hawkes model. In the seminal work of Ogata (1988), the interaction functions are given in a parametric form and estimated by maximising the likelihood function. In parametric models, an Expectation-Maximisation algorithm was proposed in Veen and Schoenberg (2008) to compute the maximum likelihood estimator while MCMC methods were designed for sampling from the posterior distribution (Rasmussen, 2013). The EM algorithm was extended by Lewis and Mohler (2011) to nonparametric Hawkes models using a penalised likelihood objective. Another nonparametric approach was introduced by Reynaud-Bouret and Schbath (2010) for the linear univariate model by using a model selection strategy. In the multivariate Hawkes model, Lasso-type estimates were designed by Hansen, Reynaud-Bouret and Rivoirard (2015). Still for linear models, Bayesian approaches have also been implemented for nonparametric Hawkes models, see for instance Du et al. (2015). In Donnet, Rivoirard and Rousseau (2020) the authors study asymptotic properties of the posterior distribution in the linear model.

Causality graphs for discrete-time events were introduced by Granger (1969) and extended to marked point processes by Didelez (2008), with an explicit definition in the case of multivariate Hawkes processes by Eichler, Dahlhaus and Dueck (2017). In linear parametric models, some approaches optimise a least-square objective based on the intensity process (Bacry et al., 2020, 2013). For nonparametric Hawkes processes, Xu, Farajtabar and Zha (2016) apply an EM algorithm based on a penalized likelihood objective leading to temporal and group sparsity. Still in the linear model, Lasso-type estimates proposed by Hansen, Reynaud-Bouret and Rivoirard (2015) for nonparametric Hawkes processes naturally lead to sparse connectivity graphs. This procedure has been generalised to high-dimensional processes by Chen, Witten and Shojaie (2017) by adding an edge screening step.

1.3. Our contributions

This paper considers a general multivariate Hawkes model with a nonlinear and nonparametric form of the intensity function, and provides theoretical guarantees on Bayesian estimation methods. We cover a large range of link functions ϕ_k, which covers most of the nonlinear Hawkes models considered in the literature (Costa et al., 2020; Hansen, Reynaud-Bouret and Rivoirard, 2015; Gerhard, Deger and

Truccolo, 2017; Carstensen et al., 2010; Chen et al., 2017; Menon and Lee, 2018; Mei and Eisner, 2017; Deutsch and Ross, 2022; Truccolo et al., 2005), such as the ReLU $\phi_k(x) = (x)_+ = \max(x, 0)$, clipped exponentials $\phi_k(x) = \min(e^x, \Lambda)$, the sigmoid $\phi_k(x) = (1 + e^{-x})^{-1}$, and the softplus $\phi_k(x) = \log(1 + e^x)$. These models have been notably introduced for neuronal spike-train data modelling, where intense-activity periods alternate with resting states called *refractory periods*[1]. The ReLU function directly extends the original linear Hawkes model to handle negative interaction functions. In Hansen, Reynaud-Bouret and Rivoirard (2015); Costa et al. (2020) it is called the *standard* nonlinear Hawkes model, as it is the closest to the linear Hawkes process. Exponential and sigmoidal functions appear in several applied works (Gerhard, Deger and Truccolo, 2017; Carstensen et al., 2010), where smoothness, saturation and thresholding effects are desirable properties. The softplus function is often preferred in machine learning algorithms as a soft approximation of ReLU (Mei and Eisner, 2017).

The first question to answer is the identifiability of $f = (\nu, h)$, which is treated in Section 2.2. Building on these results, we study posterior concentration rates in terms of the L_1-norm on f in Section 3.1. Our aim is to describe the posterior concentration rates in terms of conditions on the prior Π and on the true parameter $f_0 = (\nu_0 = (\nu_k)_k, h_0 = (h^0_{lk})_{l,k})$ which are simple to verify and under rather weak assumptions on the link functions. Interestingly, we eventually reduce the problem to conditions on the prior and the f_0 similar to those found in the literature on density and nonlinear regression estimation (see Theorem 3.2), which makes them easy to verify in a wide range of prior models. From this we derive convergence rates of Bayesian estimators of f_0 (Corollary 3.8) and posterior consistency on δ_0 (Theorem 3.9), with δ_0 the true graph parameter associated to h_0.

We also extend our results to the case where the link functions are partially unknown, in the special case of shifted ReLU link functions. More precisely we consider models in the form $\phi_k(x) = \theta_k + (x)_+$, with $\theta_k > 0$ unknown. For such models we show identifiability of the parameters (f, θ), $\theta = (\theta_k)_{1 \leqslant k \leqslant K}$ and derive a general posterior concentration rate result similar to Theorem 3.2 on both f and θ.

To the best of our knowledge, these results are the first theoretical properties on the nonparametric estimation of both f_0 and δ_0 in the frequentist and Bayesian literature of nonlinear Hawkes processes. Besides, for partially known link functions, in the particular setting of the shifted ReLU model, we also provide the first result on the estimation of the additional parameter θ. We note that recently, computational methods for a related setting have been developed in Zhou et al. (2021a,b); Malem-Shinitski, Ojeda and Opper (2022); Zhou et al. (2022). In the latter works, a sigmoidal nonlinear Hawkes model is defined with $\phi_k(x) = \theta_k(1 + e^{-x})^{-1}$ and unknown parameter $\theta = (\theta_k)_k$. However, although the theoretical analysis of the latter model is beyond the scope of this paper, it is similar in spirit to our models. In fact, our techniques could potentially be applied to this multiplicative parametrisation, which we leave for future work.

Our results are related to those of Donnet, Rivoirard and Rousseau (2020), obtained in the case of linear Hawkes processes. However, the analysis of the process and our proofs for estimating the parameter rely on *renewal* properties, newly introduced by Costa et al. (2020) in the univariate ReLU nonlinear Hawkes model. One key novelty of our work is to leverage the concept of *excursions* in the context of statistical analysis. This concept allows to decompose the trajectory of the process into independent, observable subintervals, and also to analyse the process on specific events where the parameter estimation is simplified. Developing these tools for nonlinear processes is fundamental since classical technical arguments used for linear Hawkes processes and based on Poisson branching structures cannot be applied in this case. We believe that these new proof techniques have an interest in themselves, in addition to weakening some of the assumptions on the prior distribution considered in Donnet, Rivoirard and Rousseau (2020).

[1] A refractory period is a time interval during which a neuron is unlikely to emit a spike train.

The rest of the paper is organised as follows. In Section 2, we define the multivariate stationary nonlinear Hawkes process, present the identifiability results and describe the Bayesian framework. Section 3 presents the posterior concentration results on f and θ and consistency on δ results. Section 4 is dedicated to the construction of prior distributions that satisfy the assumptions of the theorems. The most novel aspects of the proofs are reported in Section 5. Appendix A contains some technical lemmas. Finally, supplementary proofs and results can be found in the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023).

Notations. For a function h, we denote $\|h\|_1 = \int_{\mathbb{R}} |h(x)|dx$ the L_1-norm, $\|h\|_2 = \sqrt{\int_{\mathbb{R}} h^2(x)dx}$ the L_2-norm, $\|h\|_\infty = \sup_{x\in\mathbb{R}} |h(x)|$ the supremum norm, and $h^+ = max(h, 0)$, $h^- = max(-h, 0)$ its positive and negative parts. For a $K \times K$ matrix A, we denote $r(A)$ its spectral radius and $\|A\|$ its spectral norm. For a vector $u \in \mathbb{R}^K$, $\|u\|_1 = \sum_{k=1}^K |u_k|$. The notation $k \in [K]$ is used for $k \in \{1, \cdots, K\}$. For a set B and $k \in [K]$, we denote $N^k(B)$ the number of events of N^k in B and $N^k|_B$ the point process measure restricted to the set B. For random processes, the notation $\overset{\mathcal{L}}{=}$ corresponds to equality in distribution. We also denote $\mathcal{N}(u, \mathcal{H}_0, d)$ the covering number of a set $\mathcal{H}_0$ by balls of radius u w.r.t. a metric d. For any $k \in [K]$, let $\mu_k^0 = \mathbb{E}_0[\lambda_t^k(f_0)]$ be the mean of $\lambda_t^k(f_0)$ under the stationary distribution $\mathbb{P}_0$. For a set Ω, its complement is denoted Ω^c. We also use the notations $u_T \lesssim v_T$ if $|u_T/v_T|$ is bounded when $T \to \infty$, $u_T \gtrsim v_T$ if $|v_T/u_T|$ is bounded and $u_T \asymp v_T$ if $|u_T/v_T|$ and $|v_T/u_T|$ are bounded.

2. Problem setup

2.1. Definition and stationary distribution

In this section, we first recall the formal definition of a multivariate Hawkes process. We consider a probability space $(\mathcal{X}, \mathcal{G}, \mathbb{P})$ and a MPP $N = (N_t)_{t\in\mathbb{R}} = (N_t^1, \ldots, N_t^K)_{t\in\mathbb{R}}$. Let $\{\mathcal{G}_t\}_{t\in\mathbb{R}}$ be the filtration such that $\mathcal{G}_t = \sigma(N_s, s \leqslant t)$ and for $T > 0$, we assume that $\mathcal{G}_T \subset \mathcal{G}$. We say that $(N_t)_t$ is a multivariate Hawkes process with parameter $f = ((\nu_k)_{k=1}^K, (h_{lk})_{l,k=1}^K, (\theta_k)_{k=1}^K)$ adapted to $\mathcal{G}$ if

i) almost surely, $\forall k, l \in [K]$, $(N_t^k)_t$ and $(N_t^l)_t$ never jump simultaneously;
ii) for all $k \in [K]$, the $\mathcal{G}_t$-predictable intensity process of N^k at $t \in \mathbb{R}$ is given by

$$\lambda_t^k(f) = \phi_k\left(\nu_k + \sum_{l=1}^K \int_{-\infty}^{t^-} h_{lk}(t-s)dN_s^l\right), \quad k = 1, \ldots, K.$$

We consider finite-memory Hawkes processes for which interaction functions have a bounded support included in $[0, A]$ with $A > 0$ known - chosen arbitrarily large in practice. We recall that in (1), if for all k, ϕ_k is the identity function and for all l, h_{lk} is non-negative, this PP model corresponds to the classical linear Hawkes process with parameter $\nu = (\nu_k)_{k=1}^K$ and $h = (h_{lk})_{k,l=1}^K$ and intensity process:

$$\tilde{\lambda}_t^k(\nu, h) := \nu_k + \sum_{l=1}^K \int_{t-A}^{t^-} h_{lk}(t-s)dN_s^l. \tag{2}$$

With this notation, the nonlinear intensity can be written as $\lambda_t^k(f) = \phi_k(\tilde{\lambda}_t^k(\nu, h))$. For a nonlinear Hawkes process, the existence and uniqueness of a stationary distribution is proved under some assumptions on the parameters f and the link functions $\phi = (\phi_k)_k$. In the following lemma, we provide two sufficient conditions, which are variants of existing work. We recall that a function ϕ is L-Lipschitz, if for any $(x, x') \in \mathbb{R}^2$, $|\phi(x) - \phi(x')| \leqslant L|x - x'|$.

Lemma 2.1. *Let N be a Hawkes process with parameter f and link functions $(\phi_k)_k$ such that for any $k \in [K]$, $\phi_k : \mathbb{R} \to \mathbb{R}^+$ is monotone non-decreasing and L-Lipschitz, with $L > 0$. If one of the following conditions is satisfied:*

(C1) *The matrix S^+ with entries $S^+_{lk} = L\left\|h^+_{lk}\right\|_1$ satisfies $r(S^+) < 1$;*
(C2) *For any $k \in [K]$, ϕ_k is bounded, i.e., $\exists \Lambda_k > 0, \forall x \in \mathbb{R}$, $\phi_k(x) \leqslant \Lambda_k$.*

then there exists a unique stationary version of the process N with finite average.

In the previous lemma, the second stationarity condition **(C2)** directly comes from Theorem 7 by Brémaud and Massoulié (1996) and is applied to our (less general) context of Lipschitz and non-decreasing link functions. The first condition **(C1)** is obtained in Theorem 1 of Deutsch and Ross (2022), in a more restricted Hawkes model where $\phi_k(x) = (x)_+$ and the interaction functions are of the form $h_{lk} = K_{lk}g(t)$ with $g \geqslant 0$ and $K_{lk} \in \mathbb{R}$, but the same arguments can be applied to prove the stationarity of the process in our more general nonlinear model. However, in the context of inference, we will consider a slightly stronger condition:

(C1bis) The matrix S^+ with entries $S^+_{lk} = L\left\|h^+_{lk}\right\|_1$ satisfies $\left\|S^+\right\| < 1$.

From now on, we will assume that we observe on a window $[-A, T]$ a stationary Hawkes process with link functions $(\phi_k)_k$ and true parameters $f_0 = ((\nu^0_k)_{k=1}^K, (h^0_{lk})_{l,k=1}^K)$. We denote $\mathbb{P}_0$ the stationary distribution of N and $\mathbb{P}_0(.|\mathcal{G}_0)$ its conditional distribution given $\mathcal{G}_0$. We note that $\mathbb{P}_0$ is a well-defined transformation of the probability distribution $\mathbb{P}$ (through its alternative definition in Lemma S10.2 of the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023)). For $f = ((\nu_k)_{k=1}^K, (h_{lk})_{l,k=1}^K)$ satisfying the assumptions of Lemma 2.1, the log-likelihood of the processs on $[0, T]$ conditionally on $\mathcal{G}_0$ (i.e., conditionally on $N|_{[-A,0)}$) is given by

$$L_T(f) := \sum_{k=1}^K \left[\int_0^T \log(\lambda^k_t(f)) dN^k_t - \int_0^T \lambda^k_t(f) dt \right].$$

Then, for any parameter f, we define the conditional distribution $\mathbb{P}_f$ from the likelihood function

$$d\mathbb{P}_f(.|\mathcal{G}_0) = e^{L_T(f) - L_T(f_0)} d\mathbb{P}_0(.|\mathcal{G}_0). \tag{3}$$

We denote $\mathbb{E}_0$ (resp. $\mathbb{E}_f$) the expectation associated with $\mathbb{P}_0$ (resp. $\mathbb{P}_f$).

2.2. Identifiability of the parameters

In this section, we provide some conditions on the model which ensure that the parameters of the nonlinear Hawkes model defined in (1) are identifiable. To do so we consider the following weak assumption.

***Assumption* 2.2.** For $f = (\nu, h)$, there exists $\varepsilon > 0$ such that for any $k \in [K]$, ϕ_k restricted to $I_k = (\nu_k - \max_{l \in [K]} \left\|h^-_{lk}\right\|_\infty - \varepsilon, \nu_k + \max_{l \in [K]} \left\|h^+_{lk}\right\|_\infty + \varepsilon)$ is injective.

Proposition 2.3. *Let N be a nonlinear Hawkes process as defined in (1) with link functions $(\phi_k)_k$ and parameter $f = (\nu, h)$ satisfying the conditions of Lemma 2.1 and Assumption 2.2. If N' is a Hawkes processes with the same link functions $(\phi_k)_k$ and parameter $f' = (\nu', h')$, then if N and N' have the same distribution, i.e., $N \overset{\mathcal{L}}{=} N'$, then $\nu = \nu'$ and $h = h'$.*

Note that if the ϕ_k's are injective on $\mathbb{R}$, which holds in particular for the sigmoid and the softplus functions, then Assumption 2.2 is verified for all f. However our result is more general and also covers link functions which are only injective on a sub-interval of $\mathbb{R}$ such as ReLU or shifted ReLU ($\phi_k(x) = \theta_k + \max(x, 0)$) and clipped exponentials ($\phi_k(x) = \min(e^x, \Lambda_k)$). In this case, Assumption 2.2 holds over a restricted parameter space for f. More precisely, ϕ_k needs to be injective over an interval which includes all the possible values of $\nu_k + h_{lk}(s)$, for any $l \in [K]$ and $s \in [0, A]$.

***Remark* 2.4.** One consequence of Assumption 2.2 is that for any $t > 0$ such that $N[t - A, t) \leqslant 1$, then $\lambda_t^k(f) > 0$ (since ϕ_k is non-negative and monotone non-decreasing) for all $k \in [K]$. However, Assumption 2.2 still allows to model the *refractory periods* of biological neurons, i.e., when the neurons cannot or are very unlikely to fire again during a period after firing. Indeed, one can have $\lambda_t^k(f)$ very small for t such that $N^k[t - A, t) = 1$, depending on f and ϕ_k.

Proposition 2.3 supports the feasibility of the parameter estimation when the nonlinear functions ϕ_k's are fully known. It can however be extended to the setup where the link functions are partially known. In the next proposition, we consider the case of $\phi_k(x) = \theta_k + \psi_k(x)$ where ψ_k is a function such that $\lim_{x \to -\infty} \psi_k(x) = 0$ and $\theta_k \geqslant 0$ is an unknown parameter, for each $k \in [K]$. In this case, we denote $\lambda_t(f, \theta)$ the intensity process.

Proposition 2.5. *Let N be a Hawkes process with parameter $f = (\nu, h)$ and link function $\phi_k(x; \theta_k) = \theta_k + \psi_k(x)$ with $\theta_k \geqslant 0$ for any $k \in [K]$ satisfying the conditions of Lemma 2.1 and Assumption 2.2. We also assume that for all $k \in [K]$, $\lim_{x \to -\infty} \psi_k(x) = 0$ and*

$$\exists l \in [K],\ x_1 < x_2, \quad \text{such that } h_{lk}^-(x) > 0, \quad \forall x \in [x_1, x_2]. \tag{4}$$

Then if N' is a Hawkes processes with link functions $\phi_k(x; \theta_k') = \theta_k' + \psi_k(x)$, $\theta_k' \geqslant 0$ and parameter $f' = (\nu', h')$,

$$N \overset{\mathcal{L}}{=} N' \implies \nu = \nu', \quad h = h', \quad \text{and} \quad \theta = \theta', \quad \theta = (\theta_k)_{k=1}^K, \quad \theta' = (\theta_k')_{k=1}^K.$$

Besides, in this case we have $\mathbb{P}_f[\inf_{t \geqslant 0} \lambda_t^k(f, \theta) = \theta_k] = 1$.

The proofs of Propositions 2.3 and 2.5 are reported in Section S8 in the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023). In Proposition 2.5, the condition (4) implies that each component k receives some inhibition (i.e., $\exists l$, $h_{lk}^- \neq 0$). In particular, we will use this condition in the shifted ReLU model where $\psi_k(x) = (x)_+$. We note that θ_k is not identifiable when no inhibition is received by N^k (i.e., when $\forall l$, $h_{lk}^- = 0$). More precisely, the following lemma - proved in Section S8 in the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023) - states that in a mutually-exciting ReLU model, the parametrisation of the process is not unique. Informally, our models present a singularity at the parameter "$h^- = 0$".

Lemma 2.6. *Let N be a Hawkes process with parameter $f = (\nu, h)$ and link functions $\phi_k(x; \theta_k) = \theta_k + (x)_+$, $\theta_k \geqslant 0$, $k \in [K]$ satisfying Assumption 2.2, and let $k \in [K]$. If $\forall l \in [K], h_{lk} \geqslant 0$, then for any $\theta_k' \geqslant 0$ such that $\theta_k + \nu_k - \theta_k' > 0$, let N' be the Hawkes process driven by the same underlying Poisson process Q as N (see Lemma Lemma S10.2 in the Supplementaty Material (Sulem, Rivoirard and Rousseau, 2023)) with parameter $f' = (\nu', h')$ and link functions $\phi_k(x; \theta_k') = \theta_k' + (x)_+, k \in [K]$ with $\nu' = (\nu_1, \ldots, \nu_k + \theta_k - \theta_k', \ldots, \nu_K) \neq \nu$, $h' = h$, and $\theta' = (\theta_1, \ldots, \theta_k', \ldots, \theta_K) \neq \theta$. Then for any $t \geqslant 0, \lambda_t^k(f, \theta) = \lambda_t^k(f', \theta')$, and therefore $N \overset{\mathcal{L}}{=} N'$.*

2.3. Bayesian inference

We can now present our Bayesian estimation framework. We assume that the observed Hawkes process N satisfies the conditions of Lemma 2.1, i.e., the link functions ϕ_k's are monotone non-decreasing, L-Lipschitz with $L > 0$ and either we consider a bounded model $\phi_k(x) \leqslant \Lambda, \forall k, \Lambda > 0$ (condition **(C2)**) or we assume $\left\|S_0^+\right\| < 1$ (condition **(C1bis)**) with $S_0^+ = (L\left\|h_{lk}^{0+}\right\|_1)_{l,k\in[K]^2}$. We define the parameter space for $f = ((\nu_k)_{k=1}^K, (h_{lk})_{l,k=1}^K)$ and the functional space as follows. Let

$$\mathcal{H}' = \{h : [0,A] \to \mathbb{R};\ \|h\|_\infty < \infty\}, \quad \mathcal{H} = \left\{h = (h_{lk})_{l,k=1}^K \in \mathcal{H}'^{K^2};\ (h,\phi) \text{ satisfy } \textbf{(C1bis)} \text{ or } \textbf{(C2)}\right\},$$

$$\mathcal{F} = \left\{f = (\nu, h) \in (\mathbb{R}_+\backslash\{0\})^K \times \mathcal{H};\ f \text{ satisfies Assumption } 2.2\right\}.$$

We recall that for an unbounded link function, condition **(C1bis)** corresponds to $\left\|S^+\right\| < 1$ with $S^+ = (L\left\|h_{lk}^+\right\|_1)_{l,k\in[K]^2}$. We also recall that $A > 0$ is fixed. In the graph estimation problem (see Section 3.2), the parameter of interest is $\delta_0 \in \{0,1\}^{K^2}$ where $h_{lk}^0 = 0 \iff \delta_{lk}^0 = 0$. With a slight abuse of notations, we sometimes write $f = ((\nu_k)_k, (h_{lk})_{l,k}, (\delta_{lk})_{l,k})$ with $\delta \in \{0,1\}^{K^2}$.

***Remark* 2.7.** With ReLU-type link functions, we have $\mathcal{H} = \{h = (h_{l,k}) \in (\mathcal{H}')^{K^2};\ \left\|S^+\right\| < 1\}$ and $\mathcal{F} = \{f \in (\mathbb{R}_+\backslash\{0\})^K \times \mathcal{H};\ \left\|h_{lk}^-\right\|_\infty < \nu_k,\ (l,k) \in [K]^2\}$. With clipped exponential links $\phi_k(x) = \min(e^x, \Lambda_k)$, we have $\mathcal{H} = \mathcal{H}'^{K^2}$ and $\mathcal{F} = \left\{f \in \mathbb{R}_+\backslash\{0\}^K \times \mathcal{H}'^{K^2};\ \nu_k + \left\|h_{lk}^+\right\|_\infty < \log\Lambda_k,\ (l,k) \in [K]^2\right\}$.

We now define our metric on the parameter space $\mathcal{F}$. For any $f = (\nu, h),\ f' = (\nu', h') \in \mathcal{F}$, we define the following L_1-distances:

$$\left\|\nu - \nu'\right\|_1 = \sum_{k=1}^K |\nu_k - \nu'_k|, \quad \left\|h - h'\right\|_1 = \sum_{l=1}^K \sum_{k=1}^K \|h_{lk} - h'_{lk}\|_1, \quad \|f - f'\|_1 = \left\|\nu - \nu'\right\|_1 + \left\|h - h'\right\|_1.$$

Finally, we consider a prior distribution Π on $\mathcal{F}$ and define the posterior distribution on $B \subset \mathcal{F}$ as

$$\Pi(B|N) = \frac{\int_B \exp(L_T(f))d\Pi(f)}{\int_{\mathcal{F}} \exp(L_T(f))d\Pi(f)} = \frac{\int_B \exp(L_T(f) - L_T(f_0))d\Pi(f)}{\int_{\mathcal{F}} \exp(L_T(f) - L_T(f_0))d\Pi(f)} =: \frac{N_T(B)}{D_T}, \tag{5}$$

denoting $N_T(B)$ and D_T our numerator and denominator of the posterior with the form above.

3. Main results

In this section, we state our most important results on the posterior distribution on the parameter f and the restriction on the connectivity graph δ, leading respectively to convergence rates and consistency of some Bayesian nonparametric estimators.

3.1. Posterior concentration rates

We first prove that under mild assumptions on the link functions and the true parameter, we can describe the posterior concentration rate ϵ_T with respect to the L_1-distance on $\mathcal{F}$ defined in Section 2.3, in terms

of standard conditions on the prior. We then consider the case where the link functions ϕ_k depend on an unknown parameter, in the special case of shifted ReLU: $\phi_k(x;\theta_k^0) = \theta_k^0 + (x)_+$, for which we prove posterior concentration on both f_0 and θ_0. To do so, we use the following assumption on the true parameter, which is a strengthening of the identifiability condition in Assumption 2.2.

***Assumption* 3.1.** For $f_0 = (\nu_0, h_0)$, we assume that there exists $\varepsilon > 0$ such that for any $k \in [K]$, ϕ_k restricted to $I_k = (\nu_k^0 - \max_{l\in[K]} \left\|h_{lk}^{0-}\right\|_\infty - \varepsilon, \nu_k^0 + \max_{l\in[K]} \left\|h_{lk}^{0+}\right\|_\infty + \varepsilon)$ is bijective from I_k to $J_k = \phi_k(I_k)$ and its inverse is L'- Lipschitz on J_k, with $L' > 0$. We also assume that one of the two following conditions is satisfied:

i) For any $k \in [K]$, $\inf_{x\in\mathbb{R}} \phi_k(x) > 0$.

ii) For any $k \in [K]$, $\phi_k > 0$ and $\sqrt{\phi_k}$ and $\log \phi_k$ are L_1-Lipschitz with $L_1 > 0$.

The first part of Assumption 3.1, which is a slight strengthening of Assumption 2.2, holds in all cases described previously. The second part considers two cases: (i) the ϕ_k's are lower bounded by a positive constant, which holds for instance when $\phi_k(x;\theta_k) = \theta_k + \psi_k(x)$ with $\theta_k > 0$ and $\psi_k \geqslant 0$ and (ii) the ϕ_k's can approach 0 but satisfy an additional smoothness condition which holds in particular if the derivatives ϕ_k' are bounded and the functions $\log \phi_k$'s are Lipschitz. It is notably the case for the commonly used Hawkes models (Costa et al., 2020; Hansen, Reynaud-Bouret and Rivoirard, 2015; Gerhard, Deger and Truccolo, 2017; Carstensen et al., 2010; Chen et al., 2017; Menon and Lee, 2018; Mei and Eisner, 2017), see Example 1 below. Note that this assumption excludes the standard ReLU function $\phi_k(x) = (x)_+$, which we will treat separately in Proposition 3.5.

***Example* 1.** The following nonlinear models verify Assumption 3.1. Let $s, t, \Lambda > 0$.

- **Positive or shifted ReLU**-type functions: $\phi_1(x) = \max(sx, t) \geqslant t > 0$, which is injective on $[t/s, +\infty)$, s-Lipschitz and its inverse on $[t, +\infty)$, $\phi_1^{-1}(x) = s^{-1}x$ is s^{-1}-Lipschitz.
- **Clipped exponential** functions: $\phi_2(x) = \min(e^{sx}, \Lambda)$, which is injective on $(-\infty, s^{-1}\log\Lambda]$ and $s\Lambda$-Lispchitz. Its inverse on $(0, \Lambda]$, $\phi_2^{-1}(x) = s^{-1}\log x$ is Lipschitz on any compact of $(0, \Lambda]$ and $\sqrt{\phi_2}(x) = \sqrt{\min(e^{sx}, \Lambda)} = \min(e^{sx/2}, \sqrt{\Lambda})$ and $\log\phi_2 = \min(sx, \log\Lambda)$ are respectively $s\Lambda$–Lispchitz and s-Lipchitz;
- **Sigmoid** functions: $\phi_3(x) = (1 + e^{-s(x-t)})^{-1}$, which is injective on $\mathbb{R}$ and s-Lipschitz. Its inverse $\phi_3^{-1}(x) = t + \frac{1}{s}\log\frac{x}{1-x}$ is Lipschitz on any compact of $(0, 1)$, $\sqrt{\phi_3}$ is s-Lipschitz and $\frac{\phi_3'(x)}{\phi_3(x)} \leqslant s$ thus $\log\phi_3$ is s-Lipschitz;
- **Softplus** functions: $\phi_4(x) = \log(1 + e^{s(x-t)})$, which is injective on $\mathbb{R}$, s-Lipschitz and its inverse $\phi_4^{-1}(x) = \frac{1}{s}\log(e^x - 1) + t$ is Lipschitz on any compact of $\mathbb{R}_+^*$. Finally $\sqrt{\phi_4}$ and $\log\phi_4$ are s-Lipschitz.

To state our first result, we also define the following neighbourhoods in f_0 in supremum and L_2-norms respectively, for $B > 0$:

$$B_\infty(\epsilon_T) = \{f \in \mathcal{F};\ \nu_k^0 \leqslant \nu_k \leqslant \nu_k^0 + \epsilon_T,\ h_{lk}^0 \leqslant h_{lk} \leqslant h_{lk}^0 + \epsilon_T,\ (l,k) \in [K]^2\}.$$

$$B_2(\epsilon_T, B) = \{f \in \mathcal{F};\ \max_k |\nu_k - \nu_k^0| \leqslant \epsilon_T,\ \max_{l,k} \|h_{lk} - h_{lk}^0\|_2 \leqslant \epsilon_T,\ \max_{l,k} \|h_{lk}\|_\infty < B\}.$$

In particular, $B_\infty(\epsilon_T)$ is chosen so that for any $f \in B_\infty(\epsilon_T)$, $k \in [K]$ and $t \geqslant 0$, the intensities verify $\lambda_t^k(\nu, h) \geqslant \lambda_t^k(\nu_0, h_0)$. Finally we define

$$\kappa_T = 10(\log T)^r \tag{6}$$

with $r = 0$ if $(\phi_k)_k$ satisfies Assumption 3.1 (i), $r = 1$ if $(\phi_k)_k$ satisfies Assumption 3.1 (ii).

Theorem 3.2. *Let N be a Hawkes process with known link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) satisfy Assumption 3.1. Let $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$ and Π be a prior distribution on $\mathcal{F}$. We assume that the following conditions are satisfied for T large enough.*

(A0) *There exists $c_1 > 0$ such that $\Pi(B_\infty(\epsilon_T)) \geqslant e^{-c_1 T\epsilon_T^2}$.*

(A1) *There exist subsets $\mathcal{H}_T \subset \mathcal{H}$ and $c_2 > 0$ such that, with $\Upsilon_T = \{\nu = (\nu_k)_k,\ 0 < \nu_k \leqslant e^{c_2 T\epsilon_T^2},\ \forall k\}$, $\Pi(\mathcal{H}_T^c) + \Pi(\Upsilon_T^c) = o(e^{-(\kappa_T + c_1)T\epsilon_T^2})$.*

(A2) *There exist $\zeta_0 > 0$ and $x_0 > 0$ such that $\log \mathcal{N}(\zeta_0\epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant x_0 T\epsilon_T^2$.*

Then, for $M > 0$ large enough, we have

$$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 > M\sqrt{\kappa_T}\epsilon_T | N)\right] = o(1). \tag{7}$$

The proof of Theorem 3.2 is provided in Section 5.2.

***Remark* 3.3.** In Theorem 3.2, if we replace $B_\infty(\epsilon_T)$ by $B_2(\epsilon_T, B)$ for some $B > 0$ in **(A0)**, then the concentration rate in (7) is $\sqrt{\log\log T\kappa_T}\epsilon_T$ instead of $\sqrt{\kappa_T}\epsilon_T$. Replacing $B_\infty(\epsilon_T)$ by $B_2(\epsilon_T, B)$ can be useful for some families of priors, as seen in the case of mixtures of Beta distributions in Section S5.1 in the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023).

***Remark* 3.4.** Our concentration rate in (7) holds under the stationary distribution $\mathbb{P}_0$, implying in this case that the "initial condition" $N|_{[-A,0]} \subset \mathcal{G}_0$ also comes from the stationary law. However, in practice, one might observe a process on $[-A, T]$ with an arbitrary distribution on $[-A, 0]$. Under the conditions of Lemma 2.1, the dynamics of the resulting process are *stable* (in the sense of Definition 1 of Brémaud and Massoulié (1996)), using the results in Brémaud and Massoulié (1996). In particular, its distribution $\mathbb{P}_0(.|\mathcal{G}_0)$ converges exponentially fast to the stationary distribution $\mathbb{P}_0$. Therefore, we expect that (7) would still hold under $\mathbb{P}_0(.|\mathcal{G}_0)$, i.e., under a more general initial distribution on $[-A, 0]$.

An interesting aspect of Theorem 3.2 is that the assumptions on the prior **(A0)**, **(A1)** and **(A2)**, are similar to those of simpler estimation problems like density estimation, regression or linear Hawkes processes. This allows to directly derive explicit forms of the posterior concentration rates in the nonlinear Hawkes model under common families of priors, such as Gaussian processes, hierarchical Gaussian processes, basis expansions or mixture models (see van der Vaart and van Zanten (2009, 2008); Arbel, Gayraud and Rousseau (2013); Rousseau (2010) or Section 2.3.2 of Donnet, Rivoirard and Rousseau (2020)). In Section 4, we illustrate this using splines and mixture models. Additionally, our Theorem 3.2 avoids the unpleasant assumption on the prior in Theorem 3 of Donnet, Rivoirard and Rousseau (2020) which requires that for some $u_0 > 0$, $\Pi(\|S\| > 1 - u_0(\log T)^{1/6}\epsilon_T^{1/3}) \leqslant e^{-2c_1 T\epsilon_T^2}$. This is thanks to our novel proof techniques using regeneration times under the true model $\mathbb{P}_0$ (see Section 5.1).

Theorem 3.2 provides posterior concentration rates for a large class of link functions, as discussed earlier. In particular, it covers the case of shifted ReLU link functions, i.e, $\phi_k(x) = \theta_k + (x)_+$ where $\theta_k > 0$ is a *baseline* rate, which can be arbitrarily small. This link function can be seen as an alternative to the exponential function with positive baseline rate in Gerhard, Deger and Truccolo (2017) In Gerhard, Deger and Truccolo (2017), neurons firing rates are modelled using nonlinear Hawkes processes with a positive link function, which still allows to account for the refractory periods of neurons (for which the firing rate is small). Moreover, while in the case of the shifted ReLU model, Theorem 3.2 assumes that the baseline rates $\theta = (\theta_k)_k$ are known, we show in the next proposition that we can also estimate

θ. Besides, we additionally provide a posterior concentration result when using the standard ReLU function $\phi_k(x) = (x)_+$, under a stronger assumption on the model. We note that for this latter choice of link function, the intensity function is only *non-negative* and the likelihood function is equal to 0 in parts of neighbourhoods of h_0, which causes several issues in the control of the Kullback-Leibler divergence.

Before stating our results, we define neighbourhoods in θ_0, also in supremum and L_2-norms, respectively $B_\infty^\Theta(\epsilon_T) = \{\theta \in \Theta;\ \|\theta - \theta_0\|_\infty \leqslant \epsilon_T\}$ and $B_2^\Theta(\epsilon_T, B) = \{\theta \in \Theta;\ \|\theta - \theta_0\|_2 \leqslant \epsilon_T\}$, and in this case we define $\kappa_T = 10(\log T)^r$ with $r = 0$ in the shifted ReLU model (Case 2 of the following proposition) and $r = 2$ in the standard ReLU model (Case 1).

Proposition 3.5. *Let N be a nonlinear Hawkes process with link functions $(\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ satisfying Assumption 2.2. Let $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$ and Π be a prior distribution on $\mathcal{F}$.*

- ***Case 1 (Standard ReLU):** $\phi_k(x) = (x)_+$**, for all** $k \in [K]$**.** Under the Assumptions **(A0)**, **(A1)** and **(A2)** of Theorem 3.2, if f_0 verifies the following additional assumption*

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_0\left(\int_0^T \frac{\mathbb{1}_{\lambda_t^k(f_0)>0}}{\lambda_t^k(f_0)}dt\right) < +\infty, \quad k \in [K], \tag{8}$$

  then for $M > 0$ large enough, (7) holds.
- ***Case 2 (Shifted ReLU with θ_0 unknown):** $\phi_k(x;\theta_k^0) = \theta_k^0 + (x)_+$, $\theta_k^0 > 0$,* ***for all*** $k \in [K]$**.** *Let Π_θ be a prior distribution on $\Theta = \{\theta = (\theta_k)_k;\ \theta_k > 0\}$. If the Assumptions **(A0)**, **(A1)** and **(A2)** of Theorem 3.2 are satisfied when replacing $B_\infty(\epsilon_T)$ by $B_\infty(\epsilon_T) \cap B_\infty^\Theta(\epsilon_T)$ for T large enough, and if (4) is verified, then for $M > 0$ large enough,*

$$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 + \|\theta - \theta_0\|_1 > M\sqrt{\kappa_T'}\epsilon_T | N)\right] = o(1).$$

***Remark* 3.6.** In Case 1 of Proposition 3.5 only, $B_\infty(\epsilon_T)$ cannot be replaced by $B_2(\epsilon_T)$ in assumption **(A0)**. This is due to the fact that we need to consider parameters f such that the likelihood at f is positive (i.e., $\exp(L_T(f)) > 0$) in order to control the Kullback-Leibler divergence (see Lemma A2 and Lemmas S7.1 and S7.3 in the Supplementary Material(Sulem, Rivoirard and Rousseau, 2023)). In this argument, we also need the additional assumption (8). The latter is a non trivial condition on the intensity of the true model, which we do not expect to hold in many situations. For instance it does not hold if $\tilde{\lambda}_t(f_0)$ is Lipschitz in a neighbourhood of t such that $\tilde{\lambda}_t(f_0) = 0$. We expect that this can happen with significant probability as soon as one interaction functions h_{lk}^0 is Lipschitz and h_{lk}^{0-} is non-null. It is however not clear if this condition is sharp, i.e., if Bayesian or other likelihood-based methods would be suboptimal without this assumption (from our construction of tests, it is easy to construct frequentist estimates of f which converge at the rate $\sqrt{\epsilon_T}$ defined by the testing condition in Ghosal and van der Vaart (2007)). This also motivates the study of the shifted ReLU model, as an alternative of interest for modelling positive intensity functions. Nonetheless, in Lemma 4.3, we provide sufficient conditions in a finite-histogram model so that (8) holds. Finally, we note that using Theorem 1.2 of Costa et al. (2020) and notation τ_1, τ_2 for the regeneration times defined in Lemma 5.1, (8) is equivalent to $\mathbb{E}_0\left(\int_{\tau_1}^{\tau_2} \frac{\mathbb{1}_{\lambda_t^k(f_0)>0}}{\lambda_t^k(f_0)}dt\right) < +\infty$.

***Remark* 3.7.** In Theorem 3.2, we in fact obtain the posterior concentration rate on $((\phi_k(\nu_k))_k, h)$, i.e.,

$$\mathbb{E}_0\left[\Pi\left(\sum_k |\phi_k(\nu_k) - \phi_k(\nu_k^0)| + \|h - h_0\|_1 > M\sqrt{\kappa_T}\epsilon_T \middle| N\right)\right] = o(1),$$

for M a large enough constant. Moreover, if the ϕ_k's are partially known of the form $\phi_k(x;\theta_k) = \theta_k + \psi(x)$ where $\theta_k \geqslant 0$ and ψ is given, then we obtain

$$\mathbb{E}_0\left[\Pi(\|h - h_0\|_1 + \sum_k |\theta_k + \psi(\nu_k) - \theta_k^0 - \psi(\nu_k^0)| > M\sqrt{\kappa_T}\epsilon_T \middle| N)\right] = o(1).$$

In the next corollary, we deduce from the previous results the convergence rate of the posterior means

$$(\hat{\nu}, \hat{h}) = \mathbb{E}^{\Pi}[f|N] = \int_{\mathcal{F}} f d\Pi(f|N), \quad \text{and} \quad \hat{\theta} = \mathbb{E}^{\Pi}[(\theta)|N] \quad \text{when } \theta_0 \text{ is unknown (in the shifted ReLU model).}$$

Corollary 3.8. *Under the assumptions of Theorem 3.2 or Case 1 of Proposition 3.5, if* $\int_{\mathcal{F}} \|f\|_1 \, d\Pi(f) < +\infty$*, then for* $M > 0$ *large enough, it holds that*

$$\mathbb{P}_0\left[\|\hat{\nu} - \nu_0\|_1 + \|\hat{h} - h_0\|_1 > M\sqrt{\kappa_T}\epsilon_T\right] = o(1).$$

Under the assumptions of Case 2 of Proposition 3.5, we have

$$\mathbb{P}_0\left[\|\hat{\nu} - \nu_0\|_1 + \|\hat{h} - h_0\|_1 + \|\hat{\theta} - \theta_0\|_1 > M\sqrt{\kappa_T}\epsilon_T\right] = o(1).$$

The proof of Theorem 3.2 can be found in Section 5.2, and the proofs of Proposition 3.5 and Corollary 3.8 are reported in Section S1 and S4 in the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023).

3.2. Consistency on the connectivity graph

In this section, we state our consistency results on the connectivity or Granger causality graph $\delta \in \{0,1\}^{K^2}$, which characterises the fact that interaction functions between pairs of dimensions are null or not, i.e., $\delta_{lk} = 0 \iff h_{lk} = 0$, $(l,k) \in [K]^2$. We note that the definition of Granger causality graph for the linear Hawkes model (see for instance Definition 3.3 in Eichler, Dahlhaus and Dueck (2017)) also holds for the nonlinear model. This leads us to consider the following hierarchical spike-and-slab prior structure. Writing $h_{lk} = \delta_{lk}h_{lk} = \delta_{lk}S_{lk}\bar{h}_{lk}$, with $S_{lk} = \|h_{lk}\|_1$ and $\bar{h}_{lk}$ such that $\left\|\bar{h}_{lk}\right\|_1 = 1$, we define a family of priors:

$$\begin{aligned} &\delta \sim \pi_\delta, \quad \mathcal{I}(\delta) = \{(l,k) \in [K]^2;\, \delta_{lk} = 1\}, \\ &(h_{lk},\, (l,k) \in \mathcal{I}(\delta))|\delta \sim \Pi_{h|\delta}(\cdot|\delta) \quad \text{and} \quad \forall (l,k) \notin \mathcal{I}(\delta),\, h_{lk} = 0, \end{aligned} \tag{9}$$

with π_δ a probability distribution on $\{0,1\}^{K^2}$. We can either determine $\Pi_{h|\delta}$ as a distribution on the set of $(h_{lk}, (l,k) \in \mathcal{I}(\delta))$ and obtain the marginal distribution of $S = (S_{lk})_{lk}$, or construct it as in Donnet, Rivoirard and Rousseau (2020) - see also the prior construction in Section 4. Adapting **(A0)** to the above structure, we recall that δ_0 corresponds to the true connectivity parameter and we consider the following assumption

$$\textbf{(A0')} \quad \Pi(B_\infty(\epsilon_T)|\delta = \delta_0) \geqslant e^{-c_1 T\epsilon_T^2/2}, \quad \pi_\delta(\delta = \delta_0) \geqslant e^{-c_1 T\epsilon_T^2/2}.$$

For instance, one can choose $\pi_\delta = \mathcal{B}(p)^{K^2}$ with $0 < p < 1$, implying that the δ_{lk}'s are i.i.d. Bernoulli random variables. Then for any fixed p, **(A0')** is verified as soon as $\Pi_{h|\delta}(B_\infty(\epsilon_T)|\delta = \delta_0) \geqslant e^{-c_1 T\epsilon_T^2/2}$ holds. This formalism allows us to consider the posterior distribution of δ which is a key object to infer the connectivity graph. The next theorem is our posterior consistency result, which is a consequence of Theorem 3.2 and Proposition 3.5 and holds for all previously considered link functions ϕ.

Theorem 3.9. *Let N be a Hawkes process with function $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$, $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence and Π be a prior distribution on $\mathcal{F}$ satisfying the conditions of Theorem 3.2 or Proposition 3.5 (replacing **(A0)** by **(A0')**). Then,*

$$\mathbb{E}_0\left[\Pi(\delta_{lk} \neq \delta^0_{lk},\ \forall (l,k) \in \mathcal{I}(\delta_0)|N)\right] = o(1), \quad \mathcal{I}(\delta_0) = \{(l,k) \in [K]^2;\ \delta^0_{lk} = 1\}. \tag{10}$$

If in addition the following holds

$$\forall \delta \in \{0,1\}^{K^2},\ \forall C > 0,\ \forall (l,k) \in \mathcal{I}(\delta) \cap \mathcal{I}(\delta_0)^c,\ \Pi_{h|\delta}\left(S_{lk} \leqslant C\epsilon_T\,|\,\delta\right) = o\left(e^{-(\kappa_T + c_1)T\epsilon_T^2}\right), \tag{11}$$

*with $c_1 > 0$ defined in **(A0')**, then $\mathbb{E}_0\left[\Pi(\delta \neq \delta_0|N)\right] = o(1)$.*

The first part of Theorem 3.9 in (10) is directly obtained from Theorem 3.2 or Proposition 3.5 (Cases 1 and 2) and says that the posterior probability of $\delta_{lk} = 1$ converges to 1, if the edge $l \to k$ is in $I(\delta_0)$, i.e., $\delta^0_{lk} = 1$. The second and more difficult part of Theorem 3.9 is to infer a non-edge $\delta^0_{lk} = 0$. The condition (11) forces the conditional prior distribution $\Pi_{h|\delta}$ to be exponentially small around 0 for all h_{lk} such that $\delta_{lk} = 1$. We note that it also implies that if $h^0_{lk} \neq 0$ and is small, then it may not be detected nor estimated properly. In Section 4, we present two common families of priors on the S_{lk}'s that verify (11).

Interestingly, if the model is more constrained, a much weaker condition on the prior distribution on S_{lk} is required which avoids this issue on the estimation of small "signals" h^0_{lk}. We now consider two restricted Hawkes models, where the interaction functions are either all the same, or only depend on the "receiver" node. For simplicity of exposition, we consider the case of fully known link functions satisfying the assumptions of Theorem 3.2, however our next proposition remains valid for the ReLU and shifted ReLU models under the assumptions of Proposition 3.5.

- **All-equal model:** we assume that $\forall (l,k) \in [K]^2$, $h_{lk} = \delta_{lk}\tilde{h}$, with $\tilde{h} \in \mathcal{H}'$ so that $\mathcal{F} = \{f = (\nu, \delta, \tilde{h}) \in \mathbb{R}_+\backslash\{0\}^K \times \{0,1\}^{K^2} \times \mathcal{H}';\ (f,\phi)$ satisfy **(C1bis)** or **(C2)** and Assumption 3.1$\}$. When $\delta \neq 0$, then $\tilde{h} \sim \Pi_{\tilde{h}}$ is a probability distribution on $\mathcal{H}' \cap \{\tilde{h} \neq 0\}$.
- **Receiver node dependent model:** we assume that $\forall (l,k) \in [K]^2, h_{lk} = \delta_{lk}h_k$ with $h_k \in \mathcal{H}'$, so that $\mathcal{F} = \{f = (\nu, \delta, (h_k)_k);\ h_k \in \mathcal{H}',\ \forall k,\ (f,\phi)$ satisfy **(C1bis)** or **(C2)** and Assumption 3.1$\}$. We also assume that the prior distribution Π can be written as a product of priors $(\Pi_k)_k$ where for each k, Π_k is a distribution on $(\nu_k, h_k, \delta_{lk},\ l \in [K])$, restricted to $\mathcal{F}$. We denote $\delta_{\cdot k} = (\delta_{lk},\ l \in [K])$.

Proposition 3.10. *We consider a restricted Hawkes model either defined above as the **All-equal model** or as the **Receiver node dependent model**. Let N be a Hawkes process with function $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ and let Π be a prior distribution on $\mathcal{F}$ such that the prior on ν has positive and continuous density wrt the Lebesgue measure. We also assume that there exists $0 < p_1 < 1/2$ such that for any $(l,k) \in [K]^2$, $p_1 \leqslant \Pi(\delta_{lk} = 1) \leqslant 1 - p_1$.*

- *In the **All-equal model**:*
  1. *If there exists $(l,k) \in [K]^2$ such that $\delta^0_{lk} \neq 0$, then if $\Pi_{\tilde{h}}(h_0 \leqslant \tilde{h} \leqslant h_0 + \epsilon_T) \geqslant e^{-c_1 T\epsilon_T^2/2}$ and if **(A1)**, **(A2)** hold, then $\mathbb{E}_0\left[\Pi(\delta \neq \delta_0|N)\right] = o(1)$.*

2. *If $\delta_0 = 0$, then if there exists $\mathcal{H}_T \subset \mathcal{H}$ such that for all $\delta \neq 0$, $\Pi_{h|\delta}(\mathcal{H}_T^c|\delta) = o(T^{-K/2})$, if **(A2)** holds with $\epsilon_T = \sqrt{\log T/T}$, and if*

$$\forall C > 0,\ \Pi_{\tilde{h}}\left(0 < \|\tilde{h}\|_1 \leqslant C\sqrt{\log T/T}\right) = o((\log T)^{-K/2}), \tag{12}$$

then $\mathbb{E}_0[\Pi(\delta \neq 0|N)] = o(1)$.

- *In the **Receiver node dependent model**: under **(A0')**, **(A1)**, **(A2)**, for any $k \in [K]$,*
  1. *If there exists $l \in [K]$ such that $\delta^0_{lk} \neq 0$, then $\mathbb{E}_0\left[\Pi(\delta_{k_1 k} \neq \delta^0_{k_1 k}|N)\right] = o(1)$, $\forall k_1 \in [K]$.*
  2. *If $\delta^0_{\cdot k} = 0$, if there exists $\tilde{\mathcal{H}}_T \subset \mathcal{H}_1$ such that $\Pi_k(\tilde{\mathcal{H}}_T{}^c) = o(T^{-K/2})$, and if for $M > 0$ large enough and $x_0 > 0$, $\zeta_0 > 0$,*

$$\mathcal{N}\left(\zeta_0 M\sqrt{\log T/T}, \tilde{\mathcal{H}}_T, \|.\|_1\right) \leqslant T^{x_0 M},$$

and if (12) holds with h_k instead of $\tilde{h}$, then $\mathbb{E}_0\left[\Pi(\delta_{\cdot k} \neq \delta^0_{\cdot k}|N)\right] = o(1)$.

Consequently, in those restricted Hawkes models, the above proposition states that the posterior distribution is consistent at δ_0 under the much weaker assumption(12) on the prior compared to (11) of Theorem 3.9. In fact, in the **All-equal model** (resp. the **Receiver node dependent model**), if the true graph has no edge (resp. no edge arriving on node k), then the posterior distribution on h (resp. h_k) concentrates at the paranetric rate $\sqrt{\log T/T}$. This gives a sharp lower bound on the marginal density of N, i.e., on the denominator D_T in (5). We note that (12) is a mild condition which is verified in particular when the prior distribution on $\tilde{S} = \|\tilde{h}\|_1$ (resp. $S_k = \|h_k\|_1$) conditionally on $\tilde{S} \neq 0$ (resp. $S_k \neq 0$) has a density wrt the Lebesgue measure bounded by $\tilde{S}^{-a}$ (resp. S_k^{-a}) with $a > 0$ near 0.

We now study the consistency of Bayesian estimators of the connectivity graph. From Theorem 3.9 or Proposition 3.10, we can directly obtain that the graph estimator based on the $0-1$ loss function defined as $\hat{\delta}^{\Pi}_{lk}(N) = 1 \iff \Pi(\delta_{lk} = 1|N) > \Pi(\delta_{lk} = 0|N)$, is consistent, i.e., $\mathbb{P}_0\left[\hat{\delta}^{\Pi}(N) \neq \delta_0\right] = o(1)$. This result is obtained with the prior condition (11) in the non-restricted model, which as previously explained can deteriorate the inference of small and non-null interaction functions. We thus propose an alternative graph estimator based on a loss function penalising small signals, which therefore allows us to use prior distributions which do not verify (11). For any graph estimator $\hat{\delta} = (\hat{\delta}_{lk})_{l,k} \in \{0,1\}^{K^2}$ and parameter $f = (\nu, h, \delta) \in \mathcal{F}$, we define

$$L(\hat{\delta}, f) = \sum_{l,k=1}^{K} \mathbb{1}_{\hat{\delta}_{lk}=0}\mathbb{1}_{\delta_{lk}=1} + \mathbb{1}_{\hat{\delta}_{lk}=1}(\mathbb{1}_{\delta_{lk}=0} + \mathbb{1}_{\delta_{lk}=1}F(\|h_{lk}\|_1)),$$

with $F : \mathbb{R}^+ \to [0,1]$ a monotone non-increasing function, with $F(0) = 1$. For a prior Π, the risk of the estimator $\hat{\delta}$ is defined as

$$r(\hat{\delta}, \Pi|N) = \int_{\mathcal{F}} L(\hat{\delta}, f)d\Pi(f|N) = \sum_{l,k} \mathbb{1}_{\hat{\delta}_{lk}=0}\Pi(\delta_{lk} = 1|N) + \mathbb{1}_{\hat{\delta}_{lk}=1}\left[\Pi(\delta_{lk} = 0|N) + \mathbb{E}^{\Pi}(\mathbb{1}_{\delta_{lk}=1}F(\|h_{lk}\|_1)|N)\right].$$

Then the associated risk-minimising estimator, $\hat{\delta}^{\Pi,L}(N) = \arg\min_{\delta \in \{0,1\}^{K^2}} r(\delta, \Pi|N)$, verifies

$$\hat{\delta}^{\Pi,L}_{lk}(N) = 1 \iff \mathbb{E}^{\Pi}[(1 - F(\|h_{lk}\|_1))\mathbb{1}_{\delta_{lk}=1}|N] \geqslant \Pi(\delta_{lk} = 0|N). \tag{13}$$

In the next theorem, we prove that our estimator $\hat{\delta}^{\Pi,L}(N)$ is consistent under the true model $\mathbb{P}_0$ if the penalisation function F satisfies an exponential condition.

Theorem 3.11. *Let N be a Hawkes process with function $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$, $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence and Π be a prior distribution on $\mathcal{F}$ satisfying the conditions of Theorem 3.2 or Proposition 3.5 (replacing **(A0)** by **(A0')**). Then, if there exists $a > 0$ such that*

$$0 \leqslant 1 - F(M\sqrt{\kappa_T}\epsilon_T) \leqslant e^{-(c_1+a+\kappa_T)T\epsilon_T^2}, \tag{14}$$

for T large enough and with $M > 0$ defined in Theorem 3.2, then for any $(l,k) \in \mathcal{I}(\delta_0)$ such that $1 - F(\|h_{lk}^0\|_1) \geqslant 2e^{-(\kappa_T+c_1)T\epsilon_T^2}$, we have $\mathbb{P}_0\left[\hat{\delta}^{\Pi,L}(N) \neq \delta_0\right] = o(1)$.

***Remark* 3.12.** The assumption on the penalisation function (14) is verified in particular if (i) F is truncated, i.e., $F(x) = \mathbb{1}_{[0,\epsilon]}(x)$ for some (arbitrarily small) $\epsilon > 0$, or if (ii) F is exponentially decreasing around 0, i.e., $F(x) = 1 - \exp\{-\frac{1}{x^p}\}$ with $p > 1/\beta$ if $\epsilon_T = T^{-\beta/2\beta+1}(\log T)^q$ for some $q \geqslant 0$ (see Corollary 4.2 for instance). We note that the choice of penalisation function F determines the detection level of our risk-minimising graph estimator for "small signals". With (i), we will detect "signals" $\|h_{lk}^0\|_1 > \epsilon$ and with (ii), we can detect $\|h_{lk}^0\|_1 > T^{-(p(2\beta+1))^{-1}}$. We also note that this assumption is related to (11), however, since it applies on the penalisation function F and not on the prior distribution, it does not alter the posterior distribution, thus the estimation of ν_0 and h_0.

The proofs of Theorem 3.9, Proposition 3.10 and Theorem 3.11 can be found in Section S3 in the Supplementary material (Sulem, Rivoirard and Rousseau, 2023).

4. Prior models

In this section, we construct prior distributions Π that satisfy the assumptions of our main results stated in Section 3 and obtain explicit posterior concentration rates for Hölder-smooth classes of interaction functions. For ease of exposition, we consider link functions ϕ_k's injective on (m_k, M_k), with $m_k, M_k \in \mathbb{R} \cup \{-\infty, +\infty\}$.

First, we consider a prior on $\nu = (\nu_k)_k$ of the form: $\nu_k \overset{i.i.d.}{\sim} \pi_\nu(\nu_k|(h_{lk})_{l\in[K]}) \propto \pi_\nu(\nu_k)\mathbb{1}_{(m_k,M_k)}(\nu_k)$ with π_ν a positive and continuous probability density on $(0, +\infty)$. To verify **(A1)**, we can for instance choose π_ν such that $\pi_\nu(\nu_k > x) \leqslant x^{-a}$ with $a > 1$. Then it is enough to choose c_2 such that $c_2 > (\kappa_T + c_1)/a$. Moreover in Case 2 of Proposition 3.5 (i.e., shifted ReLU with unknown shift θ_0), we consider a prior on θ such that $\theta_k \overset{i.i.d.}{\sim} \pi_\theta$ with π_θ a density wrt the Lebesgue measure on $(0, +\infty)$.

For the prior on h, we consider the hierarchical structure (9) introduced in Section 3.2 and for the sake of simplicity we assume that $\delta_{lk} \overset{i.i.d.}{\sim} \mathcal{B}(p)$, $\forall (l,k) \in [K]^2$, $p \in (0,1)$, although as previously mentioned, more general priors on δ could be considered. We recall that $I(\delta) = \{(l,k) \in [K]^2;\ \delta_{lk} = 1\}$. We then consider two parametrisation setups. In the first one, $h = (h_{lk},\ (l,k) \in I(\delta))$ is drawn from a truncated distribution of the form

$$d\Pi_h(h|\delta) \propto d\Pi_h^{\otimes|I(\delta)|}(h)\mathbb{1}_{\|S^+\|<1}(h), \tag{15}$$

or simply $d\Pi_h(h|\delta) \propto d\Pi_h^{\otimes|I(\delta)|}(h)$ in the case of a bounded link function (condition **(C2)**), where Π_h is a prior distribution on one function. In the second parametrisation setup,

$$h_{lk} = S_{lk}\bar{h}_{lk}, \quad \left\|\bar{h}_{lk}\right\|_1 = 1, \quad [\bar{h}_{lk}|(l,k) \in I(\delta)] \overset{i.i.d.}{\sim} \Pi_{\bar{h}}, \quad S|\delta \sim \Pi_{S|\delta}, \tag{16}$$

with $\Pi_{\bar{h}}$ is a prior distribution on one L_1-normalised function and $\Pi_{S|\delta}$ is a prior distribution on matrices with non-zero entries δ and, under **(C1bis)**, satisfying $\left\|S^+\right\| < 1$.

Examples of the parametrisation setup (15) are Gaussian processes (or hierarchical Gaussian processes) priors, and prior distributions based on an expansion on some basis, such as Legendre, Fourier, wavelets, splines, etc. As mentioned earlier, the prior assumptions **(A0)-(A2)** are very common in the literature, which allows to directly apply existing results, as we illustrate on spline priors in Section 4.1. In Donnet, Rivoirard and Rousseau (2020), a similar construction is provided using a mixture of Betas distributions in the linear Hawkes model, which leads to the minimax rate of assumption up to a logarithmic factor. We report this construction in the nonlinear model in Section S5.1 in the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023) and obtain the same estimation rate up to logarithmic terms. The difficulty in this parametrisation might be to prove condition (11) in Theorem 3.9 for estimating the connectivity graph. In Section 4.2, we illustrate the second parametrisation setup (16) with random histogram priors, which is a setup where condition (11) can be more easily verified. We also consider a prior based on mixtures of Beta distributions in the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023). We denote $\mathcal{H}(\beta, L_0)$ the class of β-smooth functions with radius L_0.

4.1. Spline priors for Π_h

A nonparametric prior Π_h satisfying the assumptions of Theorem 3.2 can be constructed using the family of splines or free knot splines. Without loss of generality, we assume that $A = 1$. For $J \geqslant 1$, let $t_0 = 0 < t_1 < \cdots < t_J = 1$ define a partition of $[0,1]$ and $I_j = (t_{j-1}, t_j)$, $j \in [J]$. We consider splines of order $q \geqslant 0$, i.e., piecewise polynomial functions (on the partition) of degree q and for $q \geqslant 2$, $q-2$ times continuously differentiable. For a given partition, this defines a vector space of dimension $V = q + J - 1$ (see for instance Stone (1994); Ghosal, Ghosh and van der Vaart (2000)).

For the sake of simplicity, we present the construction of regular partitions, where $t_j = j/J$, however random partitions can be dealt with following the computations of Section 2.3.1 of Donnet, Rivoirard and Rousseau (2020). Let $B = (B_1, \cdots, B_V)$ be the B-spline basis of order q, as defined in Ghosal, Ghosh and van der Vaart (2000). Recall that for any $j \in [V]$, B_j has support included in an interval of length q/J, $B_j \geqslant 0$ and $\sum_j B_j(x) = 1$ for all $x \in [0,1]$. We can then define

$$h_{w,J}(x) = w^T B(x), \quad w \in \mathbb{R}^V, \quad J \sim \mathcal{P}(\lambda),$$

where $\mathcal{P}(\lambda)$ is the Poisson distribution with mean λ, and consider the following hierarchical construction of Π_h

$$w_j \overset{i.i.d.}{\sim} \pi_w, \quad 1 \leqslant j \leqslant V = q + J - 1, \tag{17}$$

with π_w a positive and continuous density on $\mathbb{R}$ satisfying $\pi_w(x) \lesssim e^{-a_1|x|^{a_2}}$ for some $a_1, a_2, \lambda > 0$.

Using Lemma 4.1 of Ghosal, Ghosh and van der Vaart (2000), if h_0 is $\mathcal{H}(\beta, L_0)$ for some $\beta \leqslant q$ and $L_0 > 0$, then setting $J_T = J_0(T/\log T)^{1/(2\beta+1)}$, $\epsilon_T = (T/\log T)^{-\beta/(2\beta+1)}$, there exist $w_0 \in \mathbb{R}^{V_T}$, $V_T = q + J_T - 1$ and $C > 0$ such that $\|h_0 - h_{w_0,J_T}\|_\infty \leqslant C\epsilon_T$. Moreover using Lemma 4.2 and Lemma 4.3 of Ghosal, Ghosh and van der Vaart (2000), we have $\|w_0\|_\infty \leqslant C_0$, for some C_0, and obtain that $\{w \in \mathbb{R}^{V_T}, \|w - w_0\|_\infty \leqslant \epsilon_T\} \subset B_\infty(\epsilon_T)$, which leads to **(A0)**. Similarly, from Lemma 4.2 of Ghosal, Ghosh and van der Vaart (2000), $\|h_{w,J} - h_{w',J}\|_1 \lesssim \|w - w'\|_\infty$ and with $\mathcal{H}_T = \{h_{w,J};\ \|w\|_\infty \leqslant T^{B_0},\ J \leqslant J_1 J_T\}$ for some $B_0 > 0$ and $J_1 > 0$, **(A1)** and **(A2)** are also verified. We finally obtain the following result.

Corollary 4.1. *Let N be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) verify the conditions of Lemma 2.1, and Assumption 3.1. Under the above spline prior,*

if for any $(l,k) \in [K]^2, h^0_{lk} \in \mathcal{H}(\beta, L_0)$ *with* $\beta \in (0, q+1]$ *and* $L_0 > 0$*, then for* $M > 0$ *large enough, we have*

$$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 > M\,(T/\log T)^{-\beta/(2\beta+1)}\,(\log T)^{q_0}\big|N)\right] = o(1),$$

where $q_0 = 0$ *if* ϕ *verifies Assumption 3.1(i) and* $q_0 = 1/2$ *if* ϕ *verifies Assumption 3.1(ii).*

To estimate the connectivity graph δ_0, one can either use the penalised estimator (13), which from the above computations and Corollary 3.11 is consistent, or use the estimator based on the 0-1 loss function if (11) can be verified. In the next section, we consider a prior based on random histograms and illustrate how the latter condition (11) can be satisfied.

4.2. Random histograms prior

Random histograms are a special case of splines with $q = 0$. These piecewise constant functions are of particular interest in the modelling of spike trains emitted by biological neurons, which only interact on certain time periods. We use a similar construction as in Section 2.3.1. of Donnet, Rivoirard and Rousseau (2020), however here the interaction functions are no longer restricted to be non-negative. Using parametrisation (16), the interaction function h_{lk} for $(l,k) \in I(\delta)$ has the form $h_{lk} = S_{lk}\bar{h}_{lk}$ and the $\bar{h}_{lk}$'s are independent and distributed as a random histogram $\bar{h}_{w,\mathbf{t}}$ defined as follows. Given a partition $\mathbf{t} : 0 = t_0 < t_1 < \cdots < t_J = 1$, we define

$$\bar{h}_{w,\mathbf{t}}(x) = \sum_{j=0}^{J-1} \frac{w_j}{t_{j+1} - t_j} \mathbb{1}_{(t_{j-1}, t_j]}, \quad \sum_{j=0}^{J-1} |w_j| = 1, \quad J \sim \mathcal{P}(\lambda),\ \lambda > 0.$$

Similarly to Donnet, Rivoirard and Rousseau (2020), the prior on $(|w_1|, \cdots, |w_J|)$ is constructed by first selecting the non-zero coefficients w_j's, then defining a Dirichlet prior on the vector of non-zero $|w_j|$'s, and finally sampling the sign of the w_j's. Hence,

$$\forall j \in [J],\ w_j = Z_j u_j, \quad Z_j \in \{-1, 0, 1\}, \quad u_j \geqslant 0, \quad \sum_{j=1}^{J} u_j = 1,$$

and $u_j = 0$ if $Z_j = 0$. We can consider $Z_j \overset{i.i.d.}{\sim}$ Multinomial(p_{-1}, p_0, p_1), with $p_{-1} + p_0 + p_1 = 1$, and given $(Z_1, \cdots, Z_J)$, $(u_{i_1}, \cdots, u_{i_{s_z}}) \sim \mathcal{D}(a_{s_z}, \cdots, a_{s_z})$, $\quad s_z = \sum_j |Z_j|$, where $i_1, \cdots, i_{s_z}$ are the indices of the non zero Z_j's and $\alpha_{-1}, \alpha_0, \alpha_1, a_{s_z} > 0$. Finally if the partition $\mathbf{t}$ is random, we consider a Dirichlet prior $\mathcal{D}(\alpha, \cdots, \alpha)$ on $(t_1, t_2 - t_1, \cdots, 1 - t_{J-1})$. We note that this construction is very similar to Section 2.3.1 of Donnet, Rivoirard and Rousseau (2020), and we therefore obtain the same results as in Corollaries 2 and 3 of Donnet, Rivoirard and Rousseau (2020).

Besides, to estimate the connectivity graph using the 0-1 loss (and to establish our posterior consistency result), we can now verify (11). This condition holds if, with $d\Pi_{S|\delta} = \prod_{(l,k)\in I(\delta)} d\Pi_S(S_{lk}) \mathbb{1}_{\|S^+\|<1}$ (under **(C1bis)**), Π_S has a positive and continuous density π_S on either $[\epsilon, 1]$ if $S^0_{lk} > \epsilon$, or if the density near 0 verifies

$$\pi_S(s^p) \propto s^{-p(\alpha-1)} \exp(-a/s^p) \mathbb{1}_{[0,1]}(s), \quad p > \beta, \quad a > 0.$$

We now present a corollary of Theorem 3.2 in the case of random histograms with random partitions, which is proved as in Donnet, Rivoirard and Rousseau (2020).

Corollary 4.2. *Let N be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) verify Assumption 3.1. Under the above random histogram prior, if for any $(l,k) \in [K]^2, h^0_{lk} \in \mathcal{H}(\beta, L_0)$ with $\beta \in (0,1]$ and $L_0 > 0$, then for M large enough, we have*

$$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 > M(T/\log T)^{-\beta/(2\beta+1)}(\log T)^q \middle| N)\right] = o(1),$$

where $q = 0$ if ϕ verifies Assumption 3.1(i), and $q = 1/2$ if ϕ verifies Assumption 3.1(ii).

Finally, in the case of the ReLU model (Proposition 3.5), we can also verify (8), in special case of the true parameter $f_0 = (\nu_0, h_0)$ where each h^0_{lk} lie in the space of finite histograms.

Lemma 4.3. *Let N be a nonlinear Hawkes process with parameter $f_0 = (\nu_0, h_0)$ and ReLU link functions $\phi_k(x) = (x)_+, \forall k$, satisfying Assumption 2.2 (and condition **(C1bis)**). If for all $(l,k) \in [K]^2$, there exists $J_0 \in \mathbb{N}^*$ such that $h^0_{lk}(t) = \sum_{j=1}^{J_0} \omega^{lk}_{j0} \mathbb{1}_{I_j}(t)$, with $\{I_j\}_{j=1}^{J_0}$ a partition of $[0,1]$ and $\forall j \in [J_0]$, $\omega^{lk}_{j0} \in \mathbb{Q}$, then (8) holds.*

***Remark* 4.4.** In the previous lemma, the condition that the weights $w^{lk}_{j0}, (l,k) \in [K]^2, j \in [J]$ are rational numbers is a technical argument that allows to find a lower bound on $\tilde{\lambda}^k_t(f_0)$ when $\lambda^k_t(f_0) > 0$. This results from a density argument of the linear combinations of the weights, which, under these conditions, constrains $\lambda^k_t(f_0)$ to take values on a lattice. Besides, we note that our result is in fact more general and applies to any model with Lipschitz link functions such that $\min_{x \in \mathbb{R}} \phi_k(x) = 0$.

Lemma 4.3 is proved in Section S5.2 in the Supplementary material (Sulem, Rivoirard and Rousseau, 2023).

5. Proofs

In this section, we report the proofs of our main theorems on the posterior concentration properties (Theorems 3.2 and Proposition 3.5), and on the estimation of the connectivity graph (Theorems 3.9 and 3.11). Instead of using the clustering structure of linear Hawkes processes like in Donnet, Rivoirard and Rousseau (2020) or a coupling technique like in Chen et al. (2017), these proofs leverage the renewal properties of nonlinear Hawkes processes notably studied by Costa et al. in Costa et al. (2020). The novelty of our proofs lies in the selection of parts or special "excursions", that allow us to estimate the parameter at a rate equivalent to the one for a linear Hawkes process. In the following section, we first recall the definitions of the concept of excursions and some properties of the process' renewal times.

5.1. Renewal times and excursions

In the following lemma, we introduce the concept of *excursions* for stationary nonlinear Hawkes processes verifying the conditions of Lemma 2.1. This result extends the ones of Costa et al. in Costa et al. (2020) to the multivariate case under condition **(C1bis)** of Lemma 2.1 and to bounded models (condition **(C2)**).

Lemma 5.1. *Let N be a Hawkes process with monotone non-decreasing and Lipschitz link functions $\phi = (\phi_k)_k$ and parameter $f = (\nu, h)$ such that (ϕ, f) verify **(C1bis)** or **(C2)**. Then the point process measure $X_t(.)$ defined as*

$$X_t(.) = N|_{(t-A,t]}, \tag{18}$$

is a strong Markov process with positive recurrent state $\emptyset$. *Let* $\{\tau_j\}_{j\geqslant 0}$ *be the sequence of random times defined as*

$$\tau_j = \begin{cases} 0 & \text{if } j = 0; \\ \inf\left\{t > \tau_{j-1};\ X_{t^-} \neq \emptyset,\ X_t = \emptyset\right\} = \inf\left\{t > \tau_{j-1};\ N|_{[t-A,t)} \neq \emptyset,\ N|_{(t-A,t]} = \emptyset\right\} & \text{if } j \geqslant 1. \end{cases}$$

Then, $\{\tau_j\}_{j\geqslant 0}$ *are stopping times for the process* N. *For* $T > 0$, *we also define*

$$J_T = \max\{j \geqslant 0;\ \tau_j \leqslant T\}. \tag{19}$$

The intervals $\{[\tau_j, \tau_{j+1})\}_{j=0}^{J_T-1} \cup [\tau_{J_T}, T]$ *form a partition of* $[0, T]$. *The point process measures* $(N|_{[\tau_j,\tau_{j+1})})_{1\leqslant j\leqslant J_T-1}$ *are i.i.d. and independent of* $N|_{[0,\tau_1)}$ *and* $N|_{[\tau_{J_T},T]}$; *they are called* excursions *and the stopping times* $\{\tau_j\}_{j\geqslant 1}$ *are called* regenerative *or* renewal *times.*

The proof of the previous lemma is omitted since it is a fairly direct multivariate extension of some elements of Proposition 3.1, Proposition 3.4, Theorem 3.5 and Theorem 3.6 in Costa et al. (2020), recalled in Section S.10 in the Supplementary Material (Sulem et al. (2023)). For the extension to bounded models, we use a direct consequence of the results in Costa et al. Costa et al. (2020) that if N is dominated by a homogeneous Poisson point process, then it also have the regenerative properties of Lemma 5.1. We also note that since A is known, the renewal times τ_j's are observable. In the rest of this article, we denote

$$\Delta\tau_1 = \tau_2 - \tau_1, \tag{20}$$

the length of a generic excursion. For any link functions ϕ_k's and parameter $f = (\nu, h)$, we denote r_f the value of the intensity process at the beginning of each excursion, defined as

$$r_f = (r_1^f, \ldots, r_K^f), \quad r_k^f = \phi_k(\nu_k), \quad k \in [K]. \tag{21}$$

In the next two lemmas, we prove some useful results on the distributions of $\Delta\tau_1$, on the number of points in a generic excursion $N[\tau_1, \tau_2)$ and on the number of excursions in the observation window $[-A, T]$, J_T, defined in (19).

Lemma 5.2. *Under the assumptions of Lemma 5.1, the random variables* $\Delta\tau_1$ *and* $N[\tau_1, \tau_2)$ *admit exponential moments. More precisely, under condition* ***(C1bis)****, with* $m = \left\|S^+\right\| < 1$, *we have*

$$\forall s < \min(\left\|r_f\right\|_1, \gamma/A), \quad \mathbb{E}_f\left[e^{s\Delta\tau_1}\right] < +\infty, \quad \textit{and} \quad \mathbb{E}_f\left[e^{sN[\tau_1,\tau_2)}\right] < +\infty, \quad \gamma = \frac{1-m}{2\sqrt{K}} \log\left(\frac{1+m}{2m}\right).$$

Under condition ***(C2)****, we have* $\forall s < \min_k \Lambda_k)$, $\mathbb{E}_f\left[e^{s\Delta\tau_1}\right] \leqslant \frac{\|\Lambda\|_1^2}{(\min_k \Lambda_k - s)^2}$ *and* $\mathbb{E}_f\left[e^{sN[\tau_1,\tau_2)}\right] < +\infty$. *In particular, this implies that* $\mathbb{E}_f\left[N[\tau_1, \tau_2) + N[\tau_1, \tau_2)^2\right] < +\infty$.

***Remark* 5.3.** The previous lemma provides exponential moments of $\Delta\tau_1$ and $N[\tau_1, \tau_2)$, under the assumption that $\left\|S^+\right\| < 1$ **(C1bis)**, but we conjecture that results of Lemma 5.2 still holds under the more general conditions $r(S^+) < 1$ **(C1)** of Lemma 2.1.

Lemma 5.4. *Under the assumptions of Lemma 5.1, for any $\beta > 0$, there exists a constant $c_\beta > 0$ such that $\mathbb{P}_f\left[J_T \notin [J_{T,\beta,1}, J_{T,\beta,2}]\right] \leqslant T^{-\beta}$, with J_T defined in (19) and*

$$J_{T,\beta,1} = \left\lfloor \frac{T}{\mathbb{E}_f[\Delta\tau_1]}\left(1 - c_\beta\sqrt{\frac{\log T}{T}}\right)\right\rfloor, \quad J_{T,\beta,2} = \left\lfloor \frac{T}{\mathbb{E}_f[\Delta\tau_1]}\left(1 + c_\beta\sqrt{\frac{\log T}{T}}\right)\right\rfloor.$$

The proofs of Lemmas 5.2 and 5.4 are reported in Section S8 in the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023).

5.2. Proof of Theorem 3.2 and Case 1 of Proposition 3.5

In this section, we prove our main posterior concentration theorem, Theorem 3.2, as well as Case 1 of Proposition 3.5, which deals with the specific case of the standard ReLU model. The first step of this proof borrows some ideas from the one of Theorem 3 in Donnet, Rivoirard and Rousseau (2020), but also introduces novel elements built from the renewal properties of the process. In particular, the posterior concentration is first proved in terms of a particular distance on the intensity process (see Proposition 5.5 below), which in fact corresponds to a stochastic (pseudo) distance on the parameter space $\mathcal{F}$. This stochastic distance $\tilde{d}_{1T}$ resembles the L_1 stochastic distance used in Donnet, Rivoirard and Rousseau (2020), except that it is restricted to a subset of the observation window $[-A, T]$ which only contains the beginning of each excursion. More precisely for any excursion index $j \in [J_T - 1]$, we denote $(U_j^{(1)}, U_j^{(2)})$ the times of the first two events after the j-th renewal time τ_j (as defined in Lemma 5.1). We note that by definition, $U_j^{(1)} \in [\tau_j, \tau_{j+1})$, $U_j^{(2)} \in [\tau_j, \tau_{j+2}]$ and $\tau_{j+1} \geqslant U_j^{(1)} + A$. We then define our restricted observation window $A_2(T)$ as

$$A_2(T) := \bigcup_{j=1}^{J_T - 1} [\tau_j, \xi_j], \tag{22}$$

with $\xi_j := U_j^{(2)}$ if $U_j^{(2)} \in [\tau_j, \tau_{j+1})$ and $\xi_j := \tau_{j+1}$ otherwise. We note that the interval $[\tau_j, \xi_j]$ corresponds either to the beginning of the j-th excursion or to the whole excursion $[\tau_j, \tau_{j+1})$ when the latter contains only one event, implying that $U_j^{(2)} \geqslant \tau_{j+1}$. Moreover, since the renewal times (and J_T) are observable, so is $A_2(T)$.

The construction of $A_2(T)$ is a novel and essential element of our proof. Informally, it corresponds to a set of intervals where the parameters can be inferred in a similar way as in the linear Hawkes model and which Lebesgue measure is of order T. More precisely, using the renewal properties from Section 5.1, we will prove, using Lemma A.1, that with probability going to 1, $|A_2(T)| \gtrsim T$ under $\mathbb{P}_0$. We can now define our auxiliary stochastic distance as

$$\tilde{d}_{1T}(f, f') = \frac{1}{T}\sum_{k=1}^{K}\int_0^T \mathbb{1}_{A_2(T)}(t)|\lambda_t^k(f) - \lambda_t^k(f')|dt, \tag{23}$$

and state our intermediate posterior concentration rate result, which holds for all models satisfying the conditions of Theorem 3.2 and the ReLU-type models considered in Proposition 3.5.

Proposition 5.5. *Under the assumptions of Theorem 3.2 or Proposition 3.5, for $M'_T = M'\sqrt{\kappa_T}$ with $M' > 0$ a large enough constant,*

$$\mathbb{E}_0\left[\Pi(\tilde{d}_{1T}(f, f_0) > M'_T\epsilon_T|N)\right] = o(1).$$

The proof of the previous proposition follows the strategy of Donnet, Rivoirard and Rousseau (2020) in Theorem 1, which is based on the now well-known argument by Ghosal and van der Vaart (2007). However, we note that in our setting, this strategy can be applied thanks to the definition of the stochastic distance which restricts the observation window to the set $A_2(T)$. We recall here its main steps. First, we restrict the space of probability events to a subset $\tilde{\Omega}_T$ that has high probability (see below and Lemma A.1). Secondly, we prove a lower bound of the denominator D_T defined in (5), derived from the technical Lemma A.2. Thirdly, we consider a ball centered at the true parameter f_0 of radius $M'_T\epsilon_T$ w.r.t. $\tilde{d}_{1T}$, denoted by $A_{d_1}(M'_T\epsilon_T) \subset \mathcal{F}$. Finally, to find an upper bound of the numerator $N_T(A_{d_1}(M'_T\epsilon_T)^c)$ defined in (5), we partition $A_{d_1}(M'_T\epsilon_T)^c$ into slices $\{S_i\}_i$ on which we can design tests that have exponentially decreasing type I and type II errors (see Lemma S6.1 in the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023)). We then define ϕ as the maximum of the tests on the individual slices S_i. Due to the space constraints, this proof is reported in Section S2 of the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023).

From Proposition 5.5, we prove Theorem 3.2 and Case 1 of Proposition 3.5 using the following classical decomposition (see for instance the proof of Theorem 1 in Donnet, Rivoirard and Rousseau (2020)). Let $A, B \in \mathcal{F}_T \subset \mathcal{F}$, with B possibly data dependent, $\phi \in [0,1]$ be a measurable test, κ_T defined in (6), and $\tilde{\Omega}_T \subset \Omega$. Then,

$$\mathbb{E}_0\left[\Pi(A \cap B|N)\right] \leqslant \mathbb{P}_0\left[\{D_T < e^{-(\kappa_T+c_1)T\epsilon_T^2}\} \cap \tilde{\Omega}_T\right] + \mathbb{E}_0\left[\phi \mathbb{1}_{\tilde{\Omega}_T}\right] + \mathbb{P}_0[\tilde{\Omega}_T^c]$$
$$+ e^{(\kappa_T+c_1)T\epsilon_T^2}\Pi(\mathcal{F}_T^c) + e^{(\kappa_T+c_1)T\epsilon_T^2}\int_{A\cap\mathcal{F}_T} \mathbb{E}_0\left[\mathbb{E}_f\left[(1-\phi)\mathbb{1}_B(f)\mathbb{1}_{\tilde{\Omega}_T}(N)\middle|\mathcal{G}_0\right]\right] d\Pi(f). \quad (24)$$

We first introduce the set $\tilde{\Omega}_T$, which from Lemma A.1, has probability $\mathbb{P}_0\left[\tilde{\Omega}_T^c\right]$ going to 0 at any polynomial rate. For $T > 0$, we denote

$$\mathcal{J}_T := \left\{J \in \mathbb{N};\ \left|\frac{J-1}{T} - \frac{1}{\mathbb{E}_0[\Delta\tau_1]}\right| \leqslant c_\beta\sqrt{\frac{\log T}{T}}\right\},$$

with $c_\beta > 0$ (and $\beta > 0$) chosen in Lemma A.1, and, with $r_0 := r_{f_0} = (r_1^0, \ldots, r_K^0)$ where $r_k^0 = \phi_k(\nu_k^0)$, and $\mu_k^0 = \mathbb{E}_0\left[\lambda_t^k(f_0)\right]$, for any k,

$$\Omega_N = \left\{\max_{k\in[K]}\sup_{t\in[0,T]} N^k[t-A,t) \leqslant C_\beta \log T\right\} \cap \left\{\sum_{k=1}^K \left|\frac{N^k[-A,T]}{T} - \mu_k^0\right| \leqslant \delta_T\right\},$$

$$\Omega_J = \{J_T \in \mathcal{J}_T\}, \quad \Omega_U = \left\{\sum_{j=1}^{J_T-1}(U_j^{(1)} - \tau_j) \geqslant \frac{T}{\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}\left(1 - 2c_\beta\sqrt{\frac{\log T}{T}}\right)\right\},$$

with $\delta_T = \delta_0\sqrt{\frac{\log T}{T}}$, $\delta_0 > 0$ and $C_\beta > 0$ chosen in Lemma A.1 and define

$$\tilde{\Omega}_T = \Omega_N \cap \Omega_J \cap \Omega_U. \quad (25)$$

The sets Ω_N, Ω_J and Ω_U control respectively the number of events, the number of excursions and the length of excursions. First, Ω_N corresponds to realisations of N such that the number of events in any interval of length A is upper bounded by $c_\beta \log T$, and the number of events on $[-A,T]$ is close to its expectation under the stationary distribution $\mathbb{P}_0$. Secondly, Ω_J corresponds to the realisations such that the number of excursions in the observation interval $[0,T]$ divided by T, J_T/T, is close to

its limit $1/\mathbb{E}_0[\Delta\tau_1]$. Thirdly, on Ω_U, the measure of the subset corresponding to the collections of the beginnings of excursions (from τ_j to the first event $U_j^{(1)}$) is of order T.

Next, we bound the denominator of the posterior D_T from (5). From Lemma A.2, together with the lower bound technique of Ghosal and van der Vaart (2007), we have that

$$\mathbb{P}_0\left[D_T < \Pi(B_\infty(\epsilon_T))e^{-\kappa_T T\epsilon_T^2}\right] \leqslant 2\int_{B_\infty(\epsilon_T)} \frac{\mathbb{P}_0[L_T(f) - L_T(f_0) < -\kappa_T T\epsilon_T^2/2]}{\Pi(B_\infty(\epsilon_T))} d\Pi(f) = o(1), \tag{26}$$

which leads to $\mathbb{P}_0\left[D_T < e^{-(\kappa_T + c_1)T\epsilon_T^2}\right] = o(1)$ using asssumption **(A0)**.

Then, we find a lower bound on $|A_2(T)|$ on $\tilde{\Omega}_T$. We recall that the point process measures $(N|_{[\tau_j,\tau_{j+1})})_{1\leqslant j\leqslant J_T-1}$ are i.i.d. and *a fortiori* that the random variables $\{U_j^{(1)} - \tau_j\}_j$ are i.i.d. Moreover, for any $j \in [J_T - 1]$, $t \in [\tau_j, U_j^{(1)})$ and $k \in [K]$, the intensity process is by construction equal to $\lambda_t^k(f_0) = r_k^0 = \phi_k(\nu_k^0)$. Therefore, conditionally on τ_j, $U_j^{(1)}$ has the same distribution as an event from a Poisson point process beginning at τ_j, with intensity $\|r_0\|_1$, since the process is the superposition of K univariate Poisson process with intensity r_k^0, $k \in [K]$. Thus, under $\mathbb{P}_0$, each variable $U_j^{(1)} - \tau_j$ follows an exponential distribution with mean $1/\|r_0\|_1$, and on Ω_U, for T large enough, we have that

$$|A_2(T)| = \sum_{j=1}^{J_T-1}(\xi_j - \tau_j) \geqslant \sum_{j=1}^{J_T-1}(U_j^{(1)} - \tau_j) \geqslant c_0 T, \quad c_0 := \frac{1}{2\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}.$$

Finally, for $R > 0$, we define the balls in L_1 and stochastic distances

$$A_{L_1}(R) := \{f \in \mathcal{F};\ \|f - f_0\|_1 \leqslant R\}, \quad A_{d_1}(R) = \{\tilde{d}_{1T}(f, f_0) \leqslant R\}.$$

We now apply the decomposition (24) with $\phi = 1$, $A := A_{L_1}(M_T\epsilon_T)^c$ and $B := A_{d_1}(M_T'\epsilon_T)$, with $M_T = M\sqrt{\kappa_T}$, $M_T' = M'\sqrt{\kappa_T}$, $M > M'$ and M' defined in Theorem 5.5. As in the proof of Theorem 3 of Donnet, Rivoirard and Rousseau (2020), we are thus left to prove that

$$\sup_{A_{L_1}(M_T\epsilon_T)^c \cap \mathcal{F}_T} \mathbb{P}_f\left[\tilde{\Omega}_T \cap A_{d_1}(M_T'\epsilon_T)|\mathcal{G}_0\right] = o_{\mathbb{P}_0}(e^{-(c_1+\kappa_T)T\epsilon_T^2}), \tag{27}$$

with c_1 defined in assumption **(A0)**. We recall that $\mathbb{P}_f$ is the process distribution associated to parameter f defined in (3). To prove (27), we consider $f \in A_{L_1}(M_T\epsilon_T)^c$ such that $\tilde{d}_{1T}(f, f_0) \leqslant M_T'\epsilon_T$ and for $l \in [K]$ and $j \in [J_T - 1]$, we define

$$Z_{jl} := \int_{\tau_j}^{\xi_j} |\lambda_t^l(f) - \lambda_t^l(f_0)|dt. \tag{28}$$

We note that using Lemma 5.1, the random variables $\{Z_{jl}\}_{j\in[J_T-1]}$ are i.i.d., and from (23) we also have that $T\tilde{d}_{1T}(f, f_0) > \max_{l\in[K]} \sum_{j=1}^{J_T-1} Z_{jl}$. In order to derive a Bernstein-type inequality on the sum of the Z_j's, we first find an upper bound of Z_{1l} and its moments. Using that the link functions ϕ_k's are L-Lipschitz, we have

$$Z_{jl} = \int_{\tau_j}^{\xi_j} |\phi_k(\tilde{\lambda}_t^l(\nu, h)) - \phi_k(\tilde{\lambda}_t^l(\nu_0, h_0))|dt \leqslant L\int_{\tau_j}^{\xi_j} |\tilde{\lambda}_t^l(\nu, h) - \tilde{\lambda}_t^l(\nu_0, h_0)|dt$$

$$\leqslant L(\xi_j - \tau_j)|\nu_l - \nu_l^0| + L\sum_k \int_{U_j^{(1)}}^{\xi_j} |h_{kl} - h_{kl}^0|(t - U_j^{(1)})dt$$

$$\leqslant L(A + U_j^{(1)} - \tau_j)|\nu_l - \nu_l^0| + L\sum_k \|h_{kl} - h_{kl}^0\|_1 \leqslant L(A + 1 + U_j^{(1)} - \tau_j)\,\|f - f_0\|_1\,. \tag{29}$$

Moreover, under $\mathbb{P}_f$, for any $j \in [J]$, $U_j^{(1)} - \tau_j$ follows an exponential distribution with mean $1/\left\|r_f\right\|_1$, therefore, for any $n \in \mathbb{N}$, $\mathbb{E}_f\left[(U_j^{(1)} - \tau_j)^n\right] = \frac{n!}{\|r_f\|_1^n}$. Using the standard inequality $(x+y)^n \leqslant 2^{n-1}(x^n + y^n)$, we thus obtain that

$$\mathbb{E}_f\left[Z_{1l}^n\right] \leqslant 2^{n-1}L^n\left((A+1)^n + \mathbb{E}_f\left[(U_j^{(1)} - \tau_j)^n\right]\right)\|f - f_0\|_1^n$$

$$\leqslant \frac{1}{2}2n!\left(2L\max\left(A+1, \frac{1}{\left\|r_f\right\|_1}\right)\|f - f_0\|_1\right)^{n-2} \times L^2\max\left(A+1, \frac{1}{\left\|r_f\right\|_1}\right)^2 \|f - f_0\|_1^2 \leqslant \frac{1}{2}n!b^{n-2}v^2, \tag{30}$$

with $b := 2L\max\left(A+1, \frac{2}{\|r_0\|_1}\right)\|f - f_0\|_1$ and $v := L\max\left(A+1, \frac{2}{\|r_0\|_1}\right)\|f - f_0\|_1$. In the last inequality, we have used the fact that $\|r_f - r_0\|_1 \lesssim \tilde{d}_{1T}(f, f_0) \leqslant M_T'\epsilon_T$ on $\tilde{\Omega}_T$. This is because $(U_1^{(1)} - \tau_1) + \cdots + (U_{J_T-1}^{(1)} - \tau_{J_T-1}) \geqslant c_0 T/2$, which leads to

$$T\tilde{d}_{1T}(f, f_0) \geqslant \sum_k |r_k^f - r_k^0|\left((U_1^{(1)} - \tau_1) + \cdots + (U_{J_T-1}^{(1)} - \tau_{J_T-1})\right) \geqslant \frac{T\sum_k |r_k^f - r_k^0|}{2\mathbb{E}_0\left[\Delta\tau_1\right]\|r_0\|_1}. \tag{31}$$

It also implies that $\|r_f\|_1 \geqslant \|r_0\|_1 - \left\|r_f - r_0\right\|_1 \geqslant \|r_0\|_1/2$ for T large enough.

Our final argument consists in using the lower bound on $\mathbb{E}_f\left[Z_{1l}\right]$ obtained in Lemma A.4. In this technical lemma, we show that there exists $l \in [K]$ and $C(f_0) > 0$ such that $\mathbb{E}_f\left[Z_{1l}\right] \geqslant C(f_0)\,\|f - f_0\|_1$. Therefore, for this l,

$$\mathbb{P}_f\left[\tilde{\Omega}_T \cap \{\tilde{d}_{1T}(f, f_0) \leqslant M_T'\epsilon_T\}\middle|\mathcal{G}_0\right] \leqslant \mathbb{P}_f\left[\tilde{\Omega}_T \cap \left\{\sum_{j=1}^{J_T-1} Z_{jl} \leqslant M_T' T\epsilon_T\right\}\middle|\mathcal{G}_0\right]$$

$$\leqslant \mathbb{P}_f\left[\tilde{\Omega}_T \cap \left\{\sum_{j=1}^{J_T-1}(Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right]) \leqslant M_T' T\epsilon_T - (J_T - 1)\mathbb{E}_f\left[Z_{jl}\right]\right\}\middle|\mathcal{G}_0\right]$$

$$\leqslant \mathbb{P}_f\left[\bigcup_{J\in\mathcal{J}_T}\left\{\sum_{j=1}^{J-1}(Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right]) \leqslant -\frac{C(f_0)T\left\|f - f_0\right\|_1}{4\mathbb{E}_0[\Delta\tau_1]}\right\}\middle|\mathcal{G}_0\right] \leqslant \sum_{J\in\mathcal{J}_T}\mathbb{P}_f\left[\sum_{j=1}^{J-1}(Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right]) \leqslant -\frac{C(f_0)T\left\|f - f_0\right\|_1}{4\mathbb{E}_0[\Delta\tau_1]}\middle|\mathcal{G}_0\right],$$

where we have used, for the third inequality, that on $\tilde{\Omega}_T$, $J_T - 1 \geqslant \frac{T}{2\mathbb{E}_0[\Delta\tau_1]}$, $\|f - f_0\|_1 \geqslant M_T\epsilon_T$ and $M_T' < M_T$. For each $J \in \mathcal{J}_T$, we can now apply the Bernstein's inequality:

$$\mathbb{P}_f\left[\sum_{j=1}^{J-1}(Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right]) \leqslant x\right] \leqslant \exp\left\{-\frac{x^2}{2(J-1)(v^2 + bx)}\right\},$$

with $x = -\frac{C(f_0)T\|f-f_0\|_1}{4\mathbb{E}_0[\Delta\tau_1]}$. We first upper bound the term $v^2 + bx$:

$$v^2 + b\frac{C(f_0)\|f-f_0\|_1}{4\mathbb{E}_0[\Delta\tau_1]} \leqslant L\max\left(A+1, \frac{2}{\|r_0\|_1}\right)\left(L\max\left(A+1, \frac{2}{\|r_0\|_1}\right) + \frac{C(f_0)}{2\|r_0\|_1\,\mathbb{E}_0[\Delta\tau_1]}\right)\|f-f_0\|_1^2$$
$$= C_1(f_0)\|f-f_0\|_1^2,$$

with $C_1(f_0) := L\max\left(A+1, \frac{2}{\|r_0\|_1}\right)\left(L\max\left(A+1, \frac{2}{\|r_0\|_1}\right) + \frac{C(f_0)}{2\|r_0\|_1\mathbb{E}_0[\Delta\tau_1]}\right)$. Finally, we obtain that

$$\mathbb{P}_f\left[\sum_{j=1}^{J-1}(Z_{jl} - \mathbb{E}_f[Z_{jl}]) \leqslant -\frac{C(f_0)T\|f-f_0\|_1}{4\mathbb{E}_0[\Delta\tau_1]}\middle|\mathcal{G}_0\right] \leqslant \exp\left\{-\frac{C(f_0)^2T^2\|f-f_0\|_1^2}{8(J-1)C_1(f_0)\|f-f_0\|_1^2}\right\} \leqslant \exp\left\{-\frac{C(f_0)^2T}{16C_1(f_0)}\right\},$$

and since $\kappa_T\epsilon_T^2 = o(1)$, we can conclude that

$$\mathbb{P}_f\left[\tilde{\Omega}_T \cap \{\tilde{d}_{1T}(f,f_0) \leqslant M'_T\epsilon_T\}\middle|\mathcal{G}_0\right] \leqslant \frac{2T}{\mathbb{E}_0[\Delta\tau_1]}\exp\left\{-\frac{C(f_0)^2T}{16C_1(f_0)}\right\} = o(e^{-(c_1+\kappa_T)T\epsilon_T^2}),$$

which corresponds to (27) and terminates the proof of Theorem 3.2 and Case 1 of Proposition 3.5.

6. Conclusion

In this paper we have established concentration and consistency properties of the posterior distribution and of Bayesian estimators of the parameter and connectivity graph, in a general class of nonlinear Hawkes processes. These results validate the common use of these models in different applied contexts. In particular, our results include the commonly used sigmoid and softplus models, as well as the more challenging ReLU model, under some additional restrictions on the parameter space. Moreover, we provide the first theoretical results for estimating an additional parameter of the link functions, in the case of shifted ReLU with unknown shift. To prove those results, we have built a new technique for obtaining model identifiability and concentration inequalities based on the decomposition of the process into excursions, recently introduced by Costa et al. Costa et al. (2020). Finally, our results hold under reasonable assumptions on the prior distribution and the true model, and we provide practical examples for which those conditions are verified.

Although rather weak assumptions have been used to prove our results, it is likely that the latter hold in more general contexts. In particular, we believe that one could relax the condition on processes with bounded memory ($A < +\infty$) since the regenerative properties of the nonlinear Hawkes processes also hold for processes with unbounded memory. One major improvement of our results would be to consider high dimensional processes ($K \to \infty$), possibly in restricted models such as sparse models Bacry et al. (2020) or clustering models Raad, Ditlevsen and Löcherbach (2020). Another perspective would be to prove the frequentist minimax rate of estimation, since it would be of great interest to evaluate the optimality of Bayesian procedures in nonlinear Hawkes processes. Some practitioners might also be interested in additional results on the estimation of the link function, through a different parametric or even nonparametric form, like in Wang et al. (2016).

Appendix A: Main lemmas

In this section, we state some important lemmas to prove our main results in Section 5. The proofs of Lemmas A.1, A.2, A.4 and A.5 are provided in Sections S2 and S9 in the Supplementary Material

(Sulem, Rivoirard and Rousseau, 2023). The first two lemmas are controls respectively of the complement of the main event $\tilde{\Omega}_T$ under the true distribution $\mathbb{P}_0$, and of the deviations of the log likelihood ratio $L_T(f_0) - L_T(f)$.

Lemma A.1. *Let $Q > 0$. We consider $\tilde{\Omega}_T$ defined in (25) in Section 5.2. For any $\beta > 0$, we can choose C_β and c_β in the definition of $\tilde{\Omega}_T$ such that $\mathbb{P}_0[\tilde{\Omega}_T^c] \leqslant T^{-\beta}$. Moreover, for any $1 \leqslant q \leqslant Q$, $\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \max_l \sup_{t\in[0,T]} \left(N^l[t-A,t)\right)^q\right] \leqslant 2T^{-\beta/2}$. Finally, the previous results hold when replacing $\tilde{\Omega}_T$ by $\tilde{\Omega}'_T = \tilde{\Omega}_T \cap \Omega_A$ with Ω_A defined in Section S1 of the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023) for the model with shifted ReLU link and unknown shift.*

Lemma A.2. *Under the assumptions of Theorem 3.2 or Proposition 3.5, for any $f \in B_\infty(\epsilon_T)$ and T large enough, we have*

$$\mathbb{P}_0\left[L_T(f_0) - L_T(f) \geqslant \frac{1}{2}\kappa_T T \epsilon_T^2\right] = o(1).$$

with

$$\kappa_T = \begin{cases} 10 & \text{(under Assumption 3.1(i))} \\ 10(\log T) & \text{(under Assumption 3.1(ii))} \\ 10(\log T)^2 & \text{(under Case 1 and condition (8))} \end{cases}$$

***Remark* A.3.** Contrary to the typical approach, the proof of Lemma A.2 is not based on the control of the variance of $L_T(f_0) - L_T(f)$, which is intractable due to the nonlinear form of the log-likelihood function, but on a decomposition of $L_T(f_0) - L_T(f) - KL(f_0, f)$ into a sum of i.i.d. terms T_j defined as:

$$T_j := \sum_k \int_{\tau_j}^{\tau_{j+1}} \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k - \int_{\tau_j}^{\tau_{j+1}} (\lambda_t^k(f_0) - \lambda_t^k(f))dt.$$

The next lemma is a notably used in the proof of Theorem 3.2 in Section 5.2 and bridges the gap between the posterior concentration rate in stochastic distance (see Theorem 5.5) and the rate in L_1-distance (Theorem 3.2).

Lemma A.4. *For $f \in \mathcal{F}_T$ and $l \in [K]$, let*

$$Z_{1l} = \int_{\tau_1}^{\xi_1} |\lambda_t^l(f) - \lambda_t^l(f_0)|dt,$$

where ξ_1 is defined in (22) in Section 5.2. Under the assumptions of Theorem 3.2 and Case 1 of Proposition 3.5, for $M_T \to \infty$ such that $M_T > M\sqrt{\kappa_T}$ with $M > 0$ and for any $f \in \mathcal{F}_T$ such that $\|\nu - \nu_0\|_1 \leqslant \max(\|\nu_0\|_1, \tilde{C})$ with $\tilde{C} > 0$, there exists $l \in [K]$ such that on $\tilde{\Omega}_T$,

$$\mathbb{E}_f[Z_{1l}] \geqslant C(f_0)\|f - f_0\|_1,$$

with $C(f_0) > 0$ a constant that depends only on f_0 and $\phi = (\phi_k)_k$.

Similarly, under the assumptions of Case 2 of Proposition 3.5, for $f \in \mathcal{F}_T$ and $\theta \in \Theta$, let $r_0 = (r_k^0)_k$, $r_f = (r_k^f)_k$ with $r_k^0 = \phi_k(\nu_k^0) = \theta_k^0 + \nu_k^0$, $r_k^f = \phi_k(\nu_k) = \theta_k + \nu_k$, $\forall k$. If $\left\|r_f - r_0\right\|_1 \leqslant \max(\|r_0\|, \tilde{C}')$ with $\tilde{C}' > 0$, then there exists $l \in [K]$ such that on $\tilde{\Omega}_T$,

$$\mathbb{E}_f[Z_{1l}] \geqslant C'(f_0)(\|r_f - r_0\|_1 + \|h - h_0\|_1), \quad C'(f_0) > 0. \tag{32}$$

Finally, this last lemma provides upper bounds on type I and type II errors for the tests used in the proof of Case 2 of Proposition 3.5 in Section S1 of the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023)for estimating the parameter of the link functions θ_0.

Lemma A.5. *Using the notations of Section S1 of the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023), for* $\theta_1 \in \bar{A}(\tilde{M}_T\epsilon_T)^c$, $f_1 \in A_{L_1}(M_T\epsilon_T) \cap \mathcal{F}_T$, *we define*

$$\phi(f_1,\theta_1) = \max_{k\in[K]} \min\left(\mathbb{1}_{N^k(I_k^0(f_1,\theta_1))-\Lambda_k^0(I_k^0(f_1,\theta_1))<-v_T} \vee \mathbb{1}_{|\mathcal{E}|<\frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}}, \mathbb{1}_{N^k(I_k^0(f_1,\theta_1))-\Lambda_k^0(I_k^0(f_1,\theta_1),f_0)>v_T} \vee \mathbb{1}_{|\mathcal{E}|<\frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}} \right),$$

with $I_k^0(f_1,\theta_1)$ *and* $\mathcal{E}$ *defined in (S1.3) and (S1.3) in the Supplementary Material (Sulem, Rivoirard and Rousseau, 2023),* $p_0 = \mathbb{P}_0[j \in \mathcal{E}]$, $\Lambda_k^0(I_0^k(f_1,\theta_1)) = \int_0^T \mathbb{1}_{I_k^0(f_1,\theta_1)}\lambda_t^k(f_0,\theta_0)dt$ *and* $v_T = w_T T\epsilon_T$ *with* $w_T = 2\sqrt{\max_k \theta_k^0(\kappa_T + c_1)} + 2x_0$ *and* $x_0 > 0$. *Then there exists* $u_1 > 2x_0$ *such that*

$$\mathbb{E}_0\left[\phi(f_1,\theta_1)\mathbb{1}'_{\tilde{\Omega}_T}\right] \leqslant e^{-u_1 T\epsilon_T^2}, \qquad \sup_{\|\theta-\theta_1\|+\|f-f_1\|\leqslant\zeta\epsilon_T} \mathbb{E}_0\left[\mathbb{E}_f\left[(1-\phi(f_1,\theta_1))\mathbb{1}_{\tilde{\Omega}'_T}\right]\middle|\mathcal{G}_0\right] = o(e^{-(\kappa_T+c_1)T\epsilon_T^2}).$$

Supplementary Material

The supplementary material contains ten sections and includes proofs and additional results, notably the proofs of Proposition 2.3, Proposition 2.5, Proposition 3.5, Corollary 3.8, Proposition 3.10, Theorem 5.5 and Theorem 3.11 (second case). It also includes an alternative construction of the prior distribution and the proofs of the technical lemmas in Section 5 and Appendix A, and Lemma 2.6. Finally, the last section contains some useful results, in particular some extensions of the results from Costa et al. (2020) related to the regenerative properties of nonlinear Hawkes processes.

Acknowledgements: The project leading to this work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834175). The project is also partially funded by the EPSRC via the CDT OxWaSP. The authors would like to thank the Editor and two anonymous referees for valuable comments and suggestions.

References

Apostolopoulou, I., Linderman, S., Miller, K. and Dubrawski, A. (2019). Mutually regressive point processes. *Advances in Neural Information Processing Systems* **32**.

Arbel, J., Gayraud, G. and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics* **40** 549–570.

Bacry, E., Delattre, S., Hoffmann, M. and Muzy, J.-F. (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Processes and their Applications* **123** 2475–2499.

Bacry, E., Bompaire, M., Gaïffas, S. and Muzy, J.-F. (2020). Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research* **21** 1–32.

Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability* 1563–1588.

Brémaud, P., Nappo, G. and Torrisi, G. L. (2002). Rate of convergence to equilibrium of marked Hawkes processes. *Journal of Applied Probability* 123–136.

Carstensen, L., Sandelin, A., Winther, O. and Hansen, N. R. (2010). Multivariate Hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics* **11** 456.

Chen, S., Witten, D. and Shojaie, A. (2017). Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process. *Electronic Journal of Statistics* **11** 1207–1234.

Chen, S., Shojaie, A., Shea-Brown, E. and Witten, D. (2017). The multivariate Hawkes process in high dimensions: beyond mutual excitation. *arXiv:1707.04928v2*.

Chornoboy, E., Schramm, L. and Karr, A. (1988). Maximum likelihood identification of neural point process systems. *Biological cybernetics* **59** 265–275.

Costa, M., Graham, C., Marsalle, L. and Tran, V. C. (2020). Renewal in Hawkes processes with self-excitation and inhibition. *Advances in Applied Probability* **52** 879–915.

Dassios, A. and Zhao, H. (2011). A dynamic contagion process. *Advances in Applied Probability* **43** 814–846.

Delattre, S. and Fournier, N. (2016). Statistical inference versus mean field limit for Hawkes processes. *Electronic Journal of Statistics* **10** 1223–1295.

Delattre, S., Fournier, N. and Hoffmann, M. (2016). Hawkes processes on large networks. *Ann. Appl. Probab.* **26** 216–261.

Deutsch, I. and Ross, G. J. (2022). Bayesian estimation of multivariate Hawkes processes with inhibition and sparsity. *arXiv preprint arXiv:2201.05009*.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 245–264.

Donnet, S., Rivoirard, V. and Rousseau, J. (2020). Nonparametric Bayesian estimation for multivariate Hawkes processes. *The Annals of Statistics* **48** 2698 – 2727.

Du, N., Farajtabar, M., Ahmed, A., Smola, A. J. and Song, L. (2015). Dirichlet-Hawkes processes with applications to vlustering vontinuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. *KDD '15* 219–228. Association for Computing Machinery, New York, NY, USA.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M. and Song, L. (2016). Recurrent marked temporal point processes: embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1555–1564.

Eichler, M., Dahlhaus, R. and Dueck, J. (2017). Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis* **38** 225–242.

Embrechts, P., Liniger, T. and Lin, L. (2011). Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability* **48** 367–378.

Ertekin, S., Rudin, C. and McCormick, T. H. (2015). Reactive point processes: A new approach to predicting power failures in underground electrical systems. *Ann. Appl. Stat.* **9** 122–144.

Farajtabar, M., Wang, Y., Gomez Rodriguez, M., Li, S., Zha, H. and Song, L. (2015). Coevolve: a joint point process model for information diffusion and network co-evolution. *Advances in Neural Information Processing Systems* **28**.

Gao, F. and Zhu, L. (2018a). Some asymptotic results for nonlinear Hawkes processes. *Stochastic Processes and their Applications* **128** 4051–4077.

Gao, X. and Zhu, L. (2018b). Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Systems* **90** 161–206.

Gerhard, F., Deger, M. and Truccolo, W. (2017). On the stability and dynamics of stochastic spiking neuron models: nonlinear Hawkes process and point process GLMs. *PLOS Computational Biology* **13** e1005390.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28** 500 – 531.

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non iid observations. *The Annals of Statistics* **35** 192-223.

Graham, C. (2021). Regenerative properties of the linear Hawkes process with unbounded memory. *The Annals of Applied Probability* **31** 2844–2863.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* 424–438.

Gusto, G. and Schbath, S. S. (2005). FADO: A statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes model. *Statistical Applications in Genetics and Molecular Biology* **4** n.p. article n° 24.

Hansen, N. R., Reynaud-Bouret, P. and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli* **21** 83–143.

Hawkes, A. G. (1971). Point Spectra of Some Mutually Exciting Point Processes. *Journal of the Royal Statistical Society. Series B (Methodological)* **33** 438–443.

Hillairet, C., Huang, L., Khabou, M. and Réveillac, A. (2021). The Malliavin-Stein method for Hawkes functionals. *arXiv preprint arXiv:2104.01583*.

Isham, V. and Westcott, M. (1979). A self-correcting point process. *Stochastic Processes and their Applications* **8** 335–347.

Karabash, D. (2012). On stability of Hawkes process. *arXiv preprint arXiv:1201.1573*.

Karabash, D. and Zhu, L. (2015). Limit theorems for marked Hawkes processes with application to a risk model. *Stochastic Models* **31** 433–451.

Lambert, R., Tuleau-Malot, C., Bessaih, T., Rivoirard, V., Bouret, Y., Leresche, N. and Reynaud-Bouret, P. (2017). Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *Journal of Neuroscience Methods* **297**.

Lewis, E. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics* **1** 1–20.

Malem-Shinitski, N., Ojeda, C. and Opper, M. (2022). Variational Bayesian Inference for Nonlinear Hawkes Process with Gaussian Process Self-Effects. *Entropy* **24**.

Massoulié, L. (1998). Stability results for a general class of interacting point processes dynamics, and applications. *Stochastic Processes and their Applications* **75** 1-30.

Mei, H. and Eisner, J. M. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems* **30**.

Menon, A. and Lee, Y. (2018). Proper loss functions for nonlinear Hawkes processes. In *Proceedings of the AAAI Conference on Artificial Intelligence* **32**.

Miscouridou, X., Caron, F. and Teh, Y. W. (2018). Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data. *Advances in Neural Information Processing Systems* **31**.

Møller, J. and Rasmussen, J. G. (2005). Perfect simulation of Hawkes processes. *Advances in applied probability* **37** 629–646.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association* **83** 9–27.

Raad, M. B. (2019). Renewal time points for Hawkes processes. *arXiv preprint arXiv:1906.02036*.

Raad, M. B., Ditlevsen, S. and Löcherbach, E. (2020). Stability and mean-field limits of age dependent Hawkes processes. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **56** 1958–1990. Institut Henri Poincaré.

Rasmussen, J. G. (2013). Bayesian Inference for Hawkes Processes. *Methodology and Computing in Applied Probability* **15** 623–642.

Reynaud-Bouret, P. and Roy, E. (2007). Some non asymptotic tail estimates for Hawkes processes. *Bulletin of the Belgian Mathematical Society-Simon Stevin* **13** 883–896.

Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics* **38** 2781-2822.

Reynaud-Bouret, P., Rivoirard, V., Grammont, F. and Tuleau-Malot, C. (2014). Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience* **4** 1–41.

Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of Betas and adaptive nonparametric estimation of the density. *Annals of Statistics* **38** 146–180.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics* **22** 118 – 171.

Sulem, D., Rivoirard, V. and Rousseau, J. (2023). Supplement to "Bayesian estimation of nonlinear Hawkes processes".

Torrisi, G. L. (2016). Gaussian approximation of nonlinear Hawkes processes. *The Annals of Applied Probability* **26** 2106–2140.

Torrisi, G. L. (2017). Poisson approximation of point processes with stochastic intensity, and application to nonlinear Hawkes processes. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **53** 679–700. Institut Henri Poincaré.

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P. and Brown, E. N. (2005). A Point Process Framework for Relating Neural Spiking Activity to Spiking History, Neural Ensemble, and Extrinsic Covariate Effects. *Journal of Neurophysiology* **93** 1074-1089. PMID: 15356183.

van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics* **36** 1435–1463.

van der Vaart, A. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Annals of Statistics* **37** 2655-2675.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space–time branching process models in seismology using an EM-type algorithm. *Journal of the American Statistical Association* **103** 614–624.

Wang, Y., Xie, B., Du, N. and Song, L. (2016). Isotonic Hawkes processes. In *International conference on machine learning* 2226–2234.

Xu, H., Farajtabar, M. and Zha, H. (2016). Learning granger causality for Hawkes processes. *33rd International Conference on Machine Learning, ICML 2016* **4** 2576–2588.

Zhou, F., Kong, Q., Zhang, Y., Feng, C. and Zhu, J. (2021a). Nonlinear Hawkes processes in time-varying system. *arXiv preprint arXiv:2106.04844*.

Zhou, F., Luo, S., Li, Z., Fan, X., Wang, Y., Sowmya, A. and Chen, F. (2021b). Efficient EM-variational inference for nonparametric Hawkes process. *Statistics and Computing* **31** 1–11.

Zhou, F., Kong, Q., Deng, Z., Kan, J., Zhang, Y., Feng, C. and Zhu, J. (2022). Efficient Inference for Dynamic Flexible Interactions of Neural Populations. *Journal of Machine Learning Research* **23** 1–49.