Nonparametric estimation of Hawkes processes with RKHSs

Anna Bonnet[1] and Maxime Sangnier[1]

[1]Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France

March 12, 2025

Abstract

This paper addresses nonparametric estimation of nonlinear multivariate Hawkes processes, where the interaction functions are assumed to lie in a reproducing kernel Hilbert space (RKHS). Motivated by applications in neuroscience, the model allows complex interaction functions, in order to express exciting and inhibiting effects, but also a combination of both (which is particularly interesting to model the refractory period of neurons), and considers in return that conditional intensities are rectified by the ReLU function. The latter feature incurs several methodological challenges, for which workarounds are proposed in this paper. In particular, it is shown that a representer theorem can be obtained for approximated versions of the log-likelihood and the least-squares criteria. Based on it, we propose an estimation method, that relies on two common approximations (of the ReLU function and of the integral operator). We provide a bound that controls the impact of these approximations. Numerical results on synthetic data confirm this fact as well as the good asymptotic behavior of the proposed estimator. It also shows that our method achieves a better performance compared to related nonparametric estimation techniques and suits neuronal applications.

Keywords: Nonlinear Hawkes process, nonparametric estimation, kernel method.

1 Introduction

Hawkes processes are a class of past-dependent point processes, widely used in many applications such as seismology [Ogata, 1988], criminology [Olinde and Short, 2020] and neuroscience [Reynaud-Bouret et al., 2013] for their ability to capture complex dependence structures. In their multidimensional version [Ogata, 1988], Hawkes processes can model pairwise interactions between different types of events, allowing to recover a connectivity graph between different features. Originally developed by Hawkes [1971] in order to model self-exciting phenomena, where each event increases the probability of a new event occurring, many extensions have been proposed ever since. In particular, nonlinear Hawkes processes have been introduced notably to detect inhibiting interactions, when an event can decrease the probability of another one appearing. Hawkes processes with inhibition are notoriously more complicated to handle due to the loss of many properties of linear Hawkes processes such as the cluster representation and the branching structure of the process [Hawkes and Oakes, 1974].

Since the first article on nonlinear Hawkes processes [Brémaud and Massoulié, 1996] proving in particular their existence, many works have focused on inhibition in the past few years. Among them, limit theorems have been established in [Costa et al., 2020] while Duval et al. [2022] obtained mean-field results on the behaviour of two neuronal populations. Regarding statistical inference, in the frequentist setting we can mention the exact maximum likelihood procedure of Bonnet et al. [2023], the least-squares approach by Bacry et al. [2020] and the nonparametric approach based on Bernstein-type polynomials by Lemonnier and Vayatis [2014]. While the first one proposes an exact inference procedure, it is restricted to exponential kernels. The other two methods do not make such an assumption but they consider a strong approximation which requires the intensity function to remain almost always positive: this can be inacurrate when the inhibiting effects are strong, providing estimation errors in such settings, as empirically shown in [Bonnet et al., 2023]. In the Bayesian framework, Sulem et al. [2024] proposed a nonparametric estimation procedure for kernel functions with bounded support. The authors then developed a variational inference procedure [Sulem et al., 2023] in order to reduce the computational cost of their method. Finally, Deutsch and Ross [2022] investigated a parametric inference method based on a new reparametrisation of the process.

Motivated by applications in neuroscience, we aim at developing a flexible model to capture complex neuronal interactions from spike train data. The current knowledge suggests several specificities regarding neuronal interactions that require appropriate modeling. Firstly, several works imply indeed that neuronal interactions can describe both inhibiting and exciting effects [Berg and Ditlevsen, 2013, Bonnet et al., 2023]. Moreover, neurons also exhibit a refractory period [Lovelace and Myers, 1994], that is a recovery period following a spike during which the neuron cannot emit another spike. In the end, it is likely that a neuron has a short-term self-inhibiting effect due to its refractory period combined with a long-term self-exciting behaviour, as illustrated by the synthetic interaction functions (or triggering kernels) in Figure 1 (top left, blue curve). This kind of interaction is very challenging to model with classical kernel functions, such as exponential ones [Bonnet et al., 2023].

In order to account for the specificities of neuronal activity, we propose a nonparametric estimation method to model and estimate complex interaction functions, that can in particular change signs along time (see the green curve in Figure 1). The flexibility of the approach relies on the mild assumption that the interaction functions belong to a reproducing kernel Hilbert space (RKHS). Our method is supported by theoretical guarantees, including representer theorems and approximation bounds. An empirical study highlights the performance of our approach compared to alternative ones including the exponential model [Bonnet et al., 2023] and Bernstein-type polynomial approximation [Lemonnier and Vayatis, 2014].

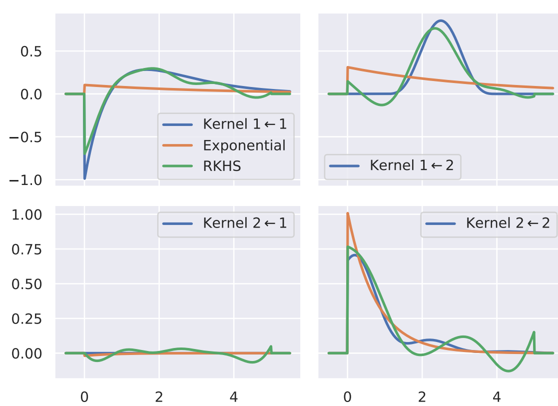

Figure 1: Example of estimation of the triggering kernels (blue) of a Hawkes process with the exponential model (orange) and the proposed method (green).

2 Related work

Nonparametric estimation of point processes, and particularly of Hawkes processes, has been largely studied in the literature, and is still an active field of research. The earliest focus has been made on the Bayesian inference for Poisson processes. This one is often based on the Gaussian-Cox model, which is an inhomogeneous Poisson process with a stochastic intensity function modulated by a Gaussian process (then rectified by a link function in order to guarantee non-negativity of the intensity).

Examples include [Møller et al., 1998, Cunningham et al., 2008, Mohler et al., 2011, Linderman and Adams, 2015] (which both require domain gridding), [Linderman and Adams, 2014] for a combination of exciting point processes with random graph models, and [Adams et al., 2009, Samo and Roberts, 2015] for inference methods based on tractable Markov Chain Monte Carlo algorithms. Concurrently, a variational inference schemes were proposed [Lloyd et al., 2015, Lian et al., 2015], which scale better than the method introduced by Samo and Roberts [2015].

Efficient nonparametric Bayesian estimation of linear Hawkes processes is well exemplified in the work by Zhang et al. [2019], which is based on sampling random branching structures with a Gibbs procedure. Since variational inference enjoys faster convergence than Gibbs sampling, the method proposed by Lloyd et al. [2015] has then been extended to inference of Hawkes processes [Zhang et al., 2020b], where interactions functions are square transformations of Gaussian processes, then to model a nonparametric baseline intensity in addition to the triggering kernels [Zhou et al., 2021], while still enjoying a scalable feature to suit large real world data. Other works modeling both the baseline intensity and the triggering kernels as transformations of Gaussian processes exist, for instance related to mean-field variational algorithms [Zhou et al., 2019, 2020].

A different line of works regarding Bayesian inference of Hawkes processes is modeling interaction kernels as piecewise constant functions and setting priors on the function levels [Donnet et al., 2020, Browning et al., 2022, Sulem et al., 2023, 2024]. The algorithmic part relies on a Markov chain Monte Carlo estimator of the posterior distribution, close to [Zhang et al., 2019].

Nonparametric frequentist estimation methods are not left out: a famous approach relies on relating the interaction functions with the second order statistics of its counting process, leading to an estimation method based on the solution to a Wiener-Hopf equation [Bacry et al., 2012, Bacry and Muzy, 2016]. Neural networks were also used to estimate general point processes [Shchur et al., 2019] (see references therein) and Hawkes processes [Pan et al., 2021], requiring a sufficient amount of data and of computational resources. Models are based on Long Short-Term Memory (LSTM) recurrent neural networks [Du et al., 2016, Mei and Eisner, 2017], independent neural networks for each triggering kernel [Joseph et al., 2022, Joseph and Jain, 2024a,b] (which is more expressive than LSTMs since it allows to recover the components of the Hawkes process and in particular its Granger causality graph), and transformers [Zuo et al., 2020, Meng et al., 2024, Wu et al., 2024].

Nonparametric inference methods also include basis decomposition or linear combination of predefined functions such as piece-wise constant functions [Marsan and Lengliné, 2008, Reynaud-Bouret and Schbath, 2010, Hansen et al., 2015, Eichler et al., 2017], exponential functions coming from Bernstein-type polynomials [Lemonnier and Vayatis, 2014], smoothing splines Lewis and Mohler [2011], Zhou et al. [2013] (estimated by solving Euler-Lagrange equations), Gaussian functions [Xu et al., 2016], and cosine series [Chen and Hall, 2016]. The bulk of these methods are designed for linear Hawkes processes and rely either on the expectation-maximization technique introduced by Lewis and Mohler [2011] or on standard numerical optimization algorithms. Some other approaches are noticeable, such as those related to kernel estimators [Chen and Hall, 2016] and autoregressive processes [Kirchner, 2017, Kirchner and Bercher, 2018, Kurisu, 2017].

Despite the vast literature regarding nonparametric estimation of point processes, only a few works addressed RKHSs. Flaxman et al. [2017a,b] proposed an inference method for inhomogeneous Poisson processes, while online estimation of linear Hawkes processes with time discretization is addressed in [Yang et al., 2017]. It is important to note that online estimation is not aimed at minimizing the estimation criterion (for instance the negative log-likelihood or the least-squares contrast) but rather at minimizing its regret. Thus, this approach is, from an optimization point of view, suboptimal when the data is available offline.

Up to our knowledge, no work tackles inference of triggering kernels from an RKHS in a batch setting (while in many applications, in particular in neuroscience, the data is available offline rather than online) and even less in inhibition setting. In addition, nonparametric frequentist methods able to handle inhibition are either based on Bernstein-type polynomials [Lemonnier and Vayatis, 2014] or on neural networks [Mei and Eisner, 2017, Meng et al., 2024, Joseph and Jain, 2024a]. As a result, our contribution lies in proposing an offline learning procedure of nonlinear Hawkes processes, for which triggering kernels come from an RKHS. The method rests on representer theorems and approximation bounds, proved in Appendices A to C and E. The proposed estimation method is implemented in Python and will be freely available on GitHub.[1]

3 Inference of Hawkes processes

Let $N = (N^{(1)}, \dots, N^{(d)})$ be a multivariate point process on $\mathbb{R}_+$, defined by its conditional intensities $\lambda^{(1)}, \dots, \lambda^{(d)}$:

$$\forall j \in [\![1, d]\!], \forall t \in \mathbb{R}_+ : \quad \lambda^{(j)}(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\left(N^{(j)}([t, t + \Delta t)) = 1 \mid \mathcal{F}_t\right)}{\Delta t},$$

where $\mathcal{F}_t = \sigma\left(N^{(j)}\left([0, s)\right), s \leq t, j \in [\![1, d]\!]\right)$ is the internal history of N [Brémaud and Massoulié, 1996]. It is assumed that N is a nonlinear Hawkes process characterized by:

$$\forall j \in [\![1, d]\!], \forall t \in \mathbb{R}_+ : \quad \lambda^{(j)}(t) = \varphi\left(\mu_j + \sum_{\ell=1}^{d} \int_{[0,t)} g_{j\ell}(t - s)\, N_\ell(\mathrm{d}s)\right),$$

where $\varphi : \mathbb{R} \to \mathbb{R}_+$ is a non-negative link function, $\mu_1, \dots, \mu_d > 0$ are baseline intensities, and $g_{11}, g_{12} \dots, g_{dd}$ are interaction functions from $\mathbb{R}$ to $\mathbb{R}$.

This work is aimed at designing an inference method for N based on RKHSs. Concretely, let $\mathcal{H}$ be a prescribed RKHS of reproducing kernel $k : \mathbb{R}^2 \to \mathbb{R}$ [Berlinet and Thomas-Agnan, 2004] and $A > 0$ some arbitrary bound. From now on, it is assumed that for all $(j, \ell) \in [\![1, d]\!]^2$, $g_{j\ell} = (h_{j\ell} + b_{j\ell})\, \mathbf{1}_{[0,A]}$ (with $\mathbf{1}_{[0,A]}$ being the indicator function of $[0, A]$), where $h_{j\ell} \in \mathcal{H}$ is the functional part of the interaction function $g_{j\ell}$ and $b_{j\ell} \in \mathbb{R}$ its offset. As commonly done (see for instance [Staerman et al., 2023]), it is assumed that events cannot have an influence far in the future, which translates here by the interaction function $g_{j\ell}$ having a bounded support. With a slight abuse of notation, we now note $\theta = ((\mu_j)_{1 \leq j \leq d}, (h_{j\ell})_{1 \leq j,\ell \leq d}, (b_{j\ell})_{1 \leq j,\ell \leq d})$ the parameter to be estimated and $\Theta^+ = \mathbb{R}_+^d \times \mathcal{H}^{d^2} \times \mathbb{R}^{d^2}$ its corresponding set (we allow baseline to cancel for optimization purposes).

Now, let, for all $j \in [\![1, d]\!]$, $\left(T_n^{(j)}\right)_{n \geq 1}$ be a sorted realization of $N^{(j)}$ in $\mathbb{R}_+$, and $N_t^{(j)} = N^{(j)}\left([0, t)\right)$ the number of these times in the interval $[0, t)$ (for any $t \geq 0$). With this notation, conditional intensities read:

$$\forall j \in [\![1, d]\!], \forall t \in \mathbb{R}_+ : \quad \lambda^{(j)}(t) = \varphi\left(\mu_j + \sum_{\ell=1}^{d} \sum_{i=1}^{N_t^{(\ell)}} g_{j\ell}\left(t - T_i^{(\ell)}\right)\right),$$

which will be left-continuous on event times here. Assuming that times are observed in the interval $[0, T]$ ($T > 0$ is a fixed horizon), there exist two common methods for estimating a Hawkes process. The first one is by minimizing the negative log-likelihood [Ozaki, 1979, Daley and Vere-Jones, 2003]:

$$\forall \theta \in \Theta^+, \quad L(\theta) = \sum_{j=1}^{d} \left[\int_0^T \lambda^{(j)}(t)\, \mathrm{d}t - \sum_{n=1}^{N_T^{(j)}} \log\left(\lambda^{(j)}\left(T_n^{(j)}\right)\right)\right],$$

and the second is by minimizing an approximated least-squares contrast [Reynaud-Bouret et al., 2014, Bacry et al., 2020]:

$$\forall \theta \in \Theta^+, \quad J(\theta) = \sum_{j=1}^{d} \left[\int_0^T \lambda^{(j)}(t)^2\, \mathrm{d}t - 2 \sum_{n=1}^{N_T^{(j)}} \lambda^{(j)}\left(T_n^{(j)}\right)\right].$$

[1] https://github.com/msangnier/kernelhawkes

Throughout this paper and as it is common for kernelized estimation (in order to prevent overfitting the training data), a quadratic penalization is added to the contrast: let $\eta > 0$ be some regularization parameter, the regularized estimation problem addressed in this paper being

$$\underset{\theta \in \Theta^+}{\text{minimize}} \;\; C(\theta) + \frac{\eta}{2} \sum_{1 \le j, \ell \le d} \|h_{j\ell}\|_{\mathcal{H}}^2, \tag{P1}$$

where C is either L or J.

As highlighted in the forthcoming sections, while the previous optimization problem will appear to be convex, two pitfalls prevent the easy derivation of kernelized estimators of Hawkes processes. The first one is the integral operator in L and J, and the second one is the nonlinear link function φ. The next sections discuss these problems and present workarounds.

3.1 Linear Hawkes processes

The linear Hawkes process is the original model of past-dependent processes [Hawkes, 1971], where interaction functions are supposed to have non-negative values: for all $(j, \ell) \in [\![1, d]\!]^2$, $g_{j\ell} : \mathbb{R} \to \mathbb{R}_+$, which makes it possible to get rid of the link function φ. Then, conditional intensities are:

$$\forall j \in [\![1, d]\!], \forall t \in \mathbb{R}_+ : \quad \lambda^{(j)}(t) = \mu_j + \sum_{\ell=1}^{d} \sum_{i=1}^{N_t^{(\ell)}} \left(h_{j\ell}\left(t - T_i^{(\ell)}\right) + b_{j\ell} \right) \mathbf{1}_{[0,A]}\left(t - T_i^{(\ell)}\right).$$

From a numerical point of view, it seems easier to estimate linear than nonlinear Hawkes processes because of losing the nonlinear function φ. This is partially true regarding the first difficulty (integrating conditional intensities) but comes with the price of non-negativity constraints on $g_{j\ell}$, for all $(j, \ell) \in [\![1, d]\!]^2$. As a result, implementation of maximum likelihood and least-squares estimators is not trivial and only approximate derivations can be produced. The first example is presented now, the second is a byproduct of Proposition 2 (see the discussion below Proposition 2).

The first example consists in considering, for all $j \in [\![1, d]\!]$, a Riemann approximation of the integral term: $\int_0^T \lambda^{(j)}(t)^2 \, \mathrm{d}t \approx \frac{T}{M} \sum_{n=1}^{M} \lambda^{(j)}(\tau_n)^2$, with $\tau_n = \frac{n-1}{M} T$ (M being an integer greater than two), and in discretizing, for all $\ell \in [\![1, d]\!]$, the non-negativity constraint $g_{j\ell} \ge 0$ to $h_{j\ell}(x_n) + b_{j\ell} \ge 0$ for all $x_n = \frac{n-1}{P-1} A$ (P being an integer greater than two).

Proposition 1 (Representer theorem for discretized least-squares estimation of linear Hawkes processes)**.** *Let φ be the identity function and*

$$\forall \theta \in \Theta^+, \quad J_M(\theta) = \sum_{j=1}^{d} \left[\frac{T}{M} \sum_{n=1}^{M} \lambda^{(j)}(\tau_n)^2 - 2 \sum_{n=1}^{N_T^{(j)}} \lambda^{(j)}\left(T_n^{(j)}\right) \right].$$

Then, denoting $\Theta = \mathbb{R}^d \times \mathcal{H}^{d^2} \times \mathbb{R}^{d^2}$, if the optimization problem

$$\begin{aligned} &\underset{\theta \in \Theta}{\text{minimize}} \;\; J_M(\theta) + \frac{\eta}{2} \sum_{1 \le j, \ell \le d} \|h_{j\ell}\|_{\mathcal{H}}^2 \\ &\text{s.t.} \quad \begin{cases} \forall j \in [\![1, d]\!], \mu_j \ge 0 \\ \forall (j, \ell) \in [\![1, d]\!]^2, \forall n \in [\![1, P]\!], h_{j\ell}(x_n) + b_{j\ell} \ge 0, \end{cases} \end{aligned}$$

has a solution θ, it is of the form:

$$\forall (j, \ell) \in [\![1, d]\!]^2, \quad h_{j\ell} = \eta^{-1} \left[2 q_{j\ell} + \sum_{n=1}^{P} \beta_n^{(j\ell)} k(\cdot, x_n) - \sum_{n=1}^{M} \alpha_n^{(j)} r_{\ell n} \right],$$

where $(\alpha^{(j)})_{1\le j\le d} \in (\mathbb{R}^M)^d$, $(\beta^{(j\ell)})_{1\le j,\ell\le d} \in (\mathbb{R}_+^P)^{d\times d}$ *and for all* $(j,\ell) \in [\![1,d]\!]^2$*:*

$$\begin{cases} q_{j\ell} = \sum_{\substack{1\le n\le N_T^{(j)}\\ 1\le i\le N_T^{(\ell)}}} k\left(\cdot, T_n^{(j)} - T_i^{(\ell)}\right) \mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A} \\ r_{\ell n} = \sum_{i=1}^{N_T^{(\ell)}} k\left(\cdot, \tau_n - T_i^{(\ell)}\right) \mathbf{1}_{0<\tau_n - T_i^{(\ell)}\le A}. \end{cases}$$

Proposition 1 is usually called a representer theorem. It states that even though the optimization problem considered is semi-parametric, a solution of it is supported by the data, and can thus be described with an amount of parameters depending on the number of observations. Here, Proposition 1 tells that for discretized least-squares estimation of linear Hawkes processes, a solution can be expressed with $d\,(M+dP)$ parameters.

Proposition 1 is rather a weak result in that it highly relies on discretization (embodied by arbitrarily large integers M and P), which entails high computational cost. However, it has the merit of showing that nonparametric estimation is possible in an ideal situation. In addition, let us remark that in the expected case where for all $(j,\ell) \in [\![1,d]\!]^2$ and $n \in [\![1,P]\!]$, $h_{j\ell}(x_n) + b_{j\ell} > 0$, then by complementary slackness of Karush-Kuhn-Tucker conditions, $\beta_n^{(j\ell)} = 0$, leading to a much simpler form of $h_{j\ell}$.

Regarding discretized maximum likelihood estimation, such a result remains unestablished, the logarithm function preventing us from isolating saddle points of the Lagrangian function.

3.2 Nonlinear Hawkes processes

As it is common in the literature (in order to be consistent with linear Hawkes processes), we focus on nonlinear processes built on the ReLU link function $\varphi : x \in \mathbb{R} \mapsto \max(0,x)$. Moreover, we consider the convention $\log(x) = -\infty$ for all $x \le 0$. Thereafter, we denote $N_T = \sum_{j=1}^d N_T^{(j)}$ the total number of observed points from the process.

If it turns out that obtaining representer properties for nonlinear Hawkes process estimation based on Problem (P1) is quite intricate, Propositions 2 and 3 present representer theorems for approximations of the objective functions.

Proposition 2 (Representer theorem for approximated maximum likelihood estimation)**.** *The negative log-likelihood* L *can be approximated by a function* $L_0 : \theta \in \Theta^+ \to \mathbb{R}$ *such that if the approximated regularized maximum likelihood problem*

$$\arg\min_{\theta\in\Theta^+} L_0(\theta) + \frac{\eta}{2} \sum_{1\le j,\ell\le d} \|h_{j\ell}\|_{\mathcal{H}}^2$$

admits a minimizer, it has a solution θ *of the form:*

$$\forall (j,\ell) \in [\![1,d]\!]^2, \quad h_{j\ell} = \alpha_0^{(j\ell)} r_\ell + \sum_{u=1}^{N_T^{(j)}} \alpha_u^{(j\ell)} q_{uj\ell},$$

for some $d\,(N_T+d)$ *real values* $\left\{\alpha_u^{(j\ell)}, (j,\ell) \in [\![1,d]\!]^2, u \in [\![0, N_T^{(j)}]\!]\right\}$*, where for all* $(j,\ell) \in [\![1,d]\!]^2$*:*

$$\begin{cases} r_\ell = \sum_{v=1}^{N_T^{(\ell)}} \int_0^T k\left(\cdot, t - T_v^{(\ell)}\right) \mathbf{1}_{0<t-T_v^{(\ell)}\le A}\,\mathrm{d}t \\ q_{uj\ell} = \sum_{v=1}^{N_T^{(\ell)}} k\left(\cdot, T_u^{(j)} - T_v^{(\ell)}\right) \mathbf{1}_{0<T_u^{(j)}-T_v^{(\ell)}\le A}, \quad \forall u \in [\![1, N_T^{(j)}]\!]. \end{cases}$$

The approximation L_0 has been proposed by Lemonnier and Vayatis [2014]. It consists in computing, for all $j \in [\![1,d]\!]$, $\int_0^T \lambda^{(j)}(t)\,\mathrm{d}t$ as if φ were the identity function (thus adding a negative contribution when $\lambda^{(j)}$ is null). In practice, this approximation is detrimental when there is a lot of inhibition (that is when $\lambda^{(j)}$ is often null) but painless for exciting and moderately inhibiting processes. In particular, if the process to estimate is a linear (exciting) Hawkes process, then it is very likely that L and L_0 coincide around the true parameter, such that the estimator obtained with L_0 is exactly that obtained by minimizing the negative log-likelihood with non-negativity constraints on $g_{j\ell}$, for all $(j, \ell) \in [\![1, d]\!]^2$.

Let us remark that the representer theorem does not hold directly for L (neither for least-squares inference based on J) because, for all $j \in [\![1, d]\!]$, the term $\int_0^T \lambda^{(j)}(t)\,\mathrm{d}t$ (respectively $\int_0^T \lambda^{(j)}(t)^2\,\mathrm{d}t$) is no longer linear in $h_{j\ell}$, $\ell \in [\![1, d]\!]$. This shortcoming has already been observed for estimation of non-homogeneous processes using RKHSs [Flaxman et al., 2017a, Yang et al., 2017].

Proposition 3 (Representer theorem for approximate least-squares estimation)**.** *The least-squares contrast J can be upper bounded by a function $J^+ : \theta \in \Theta^+ \to \mathbb{R}$ such that if the approximated regularized least-squares problem*

$$\arg\min_{\theta \in \Theta^+} J^+(\theta) + \frac{\eta}{2} \sum_{1 \le j, \ell \le d} \|h_{j\ell}\|_{\mathcal{H}}^2$$

admits a minimizer, it has a solution θ of the form given by Proposition 2.

Despite being based on approximations of the objective functions of interest, both Propositions 2 and 3 make explicit a representation of the functional parts of kernelized estimators of Hawkes processes. Fortified by this positive result, we consider now semi-parametric candidates of the form given by Proposition 2, both for maximum likelihood and least-squares estimation, and we thus define $\Theta_{\|}^+$ to be the set of parameters $\theta = ((\mu_j)_{1 \le j \le d}, (h_{j\ell})_{1 \le j, \ell \le d}, (b_{j\ell})_{1 \le j, \ell \le d}) \in \Theta^+$ such that for all $1 \le j, \ell \le d$, $h_{j\ell}$ has the form given by Proposition 2. Section 3.3 presents the practical implementation of the proposed estimator.

3.3 Kernelized expression of contrasts

According to the previous sections, we propose here to concede three approximations for estimating Hawkes processes with kernels: i) interactions functions are assumed to have the form proposed by Proposition 2, that is $h_{j\ell} = \alpha_0^{(j\ell)} r_\ell + \sum_{u=1}^{N_T^{(j)}} \alpha_u^{(j\ell)} q_{uj\ell}$, for all $(j, \ell) \in [\![1, d]\!]^2$; ii) the integral term is replaced by a Riemann approximation, that is $\int_0^T \lambda^{(j)}(t)\,\mathrm{d}t \approx \frac{T}{M} \sum_{n=1}^M \lambda^{(j)}(\tau_n)$ for all $j \in [\![1, d]\!]$ (respectively for $\int_0^T \lambda^{(j)}(t)^2\,\mathrm{d}t$) with $\tau_n = \frac{n-1}{M} T$; iii) the ReLU function φ is replaced by the softplus function $\tilde{\varphi} : x \mapsto \log(1 + \mathrm{e}^{\omega x})/\omega$, where $\omega > 0$ is a hyperparameter tuning the proximity of the approximation, which is a smooth upperbound of φ. This alteration, which ensures the differentiability of the objective function of Problem (P1), is common in the literature, regarding neural point processes [Mei and Eisner, 2017, Zuo et al., 2020, Zhang et al., 2020a, Mei et al., 2022], but also for classification with the nondifferentiable hinge loss [Chapelle, 2007, Wang et al., 2019, Luo et al., 2021]. Then, considering either $\varphi_1 = \tilde{\varphi}$ and $\varphi_2 = \log \circ \tilde{\varphi}$ (for the negative log-likelihood L), or $\varphi_1 = \tilde{\varphi}^2$ and $\varphi_2 = 2\tilde{\varphi}$ (for the least-squares criterion J), both as point-wise functions, the objective function of the regularized estimation Problem (P1) can be approximated by:

$$
\begin{aligned}
&\forall \theta \in \Theta^+_{\parallel}, \\
&F_{M,\omega}(\theta) = \sum_{j=1}^{d} \left[\frac{T}{M} \sum_{n=1}^{M} \varphi_1 \left(\mu_j + \sum_{\ell=1}^{d} \sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(\tau_n - T_i^{(\ell)}\right) + b_{j\ell} \right) \mathbf{1}_{0<\tau_n - T_i^{(\ell)} \le A} \right) \right. \\
&\qquad \left. - \sum_{n=1}^{N_T^{(j)}} \varphi_2 \left(\mu_j + \sum_{\ell=1}^{d} \sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(T_n^{(j)} - T_i^{(\ell)}\right) + b_{j\ell} \right) \mathbf{1}_{0<T_n^{(j)} - T_i^{(\ell)} \le A} \right) \right] + \frac{\eta}{2} \sum_{1\le j,\ell\le d} \|h_{j\ell}\|_{\mathcal{H}}^2 \\
&\qquad = \sum_{j=1}^{d} \left[\frac{T}{M} \mathbb{1}^\top \varphi_1 \left(\mu_j \mathbb{1} + Q^{(j)}\alpha^{(j)} + Bb^{(j)} \right) - \mathbb{1}^\top \varphi_2 \left(\mu_j \mathbb{1} + K^{(j)}\alpha^{(j)} + E^{(j)} b^{(j)} \right) \right] \\
&\qquad + \frac{\eta}{2} \sum_{1\le j,\ell\le d} {\alpha^{(j\ell)}}^\top K^{(j\ell)} \alpha^{(j\ell)},
\end{aligned}
$$

where for all $j \in [\![1, d]\!]$,

$$
\alpha^{(j)} = \begin{bmatrix} \alpha^{(j1)} \\ \vdots \\ \alpha^{(jd)} \end{bmatrix} \in \mathbb{R}^{d(N_T^{(j)}+1)} \quad \text{and} \quad b^{(j)} = \begin{bmatrix} b_{j1} \\ \vdots \\ b_{jd} \end{bmatrix} \in \mathbb{R}^d,
$$

and matrices are made explicit in Appendix D. Since $\tilde{\varphi}$ is differentiable, this also holds true for φ_1, φ_2 and $F_{M,\omega}$. Then, gradients can be computed easily and have the form expressed in Appendix D.

If the first of our three approximations is legitimate in nonparametric inference (as justified by the previous section), one may wonder how much the last two deteriorate the estimation. The forthcoming paragraph shows that the numerical impact is bounded by $O(1/M) + O(1/\omega)$. We think this is a reasonable price to pay to overcome the two obstacles of integration and nondifferential optimization. For this statement to be quantified (by Propositions 4 and 5 respectively for maximum likelihood estimation and least-squares minimization), we move to Ivanov regularized optimization problems, as it is common for statistical analysis of empirical risk minimizers (see for instance [Bartlett and Mendelson, 2002]).

For this purpose, let $B > 0$ and $C > 0$ be two fixed bounds, and Ω be the set of parameters $\theta = ((\mu_j)_{1\le j\le d}, (h_{j\ell})_{1\le j,\ell\le d}, (b_{j\ell})_{1\le j,\ell\le d}) \in \Theta^+$ with bounded norms:

$$
\Omega = \left\{ \theta \in \Theta^+ : \|\mu\|_\infty \le B, \|b\|_\infty \le B, \sqrt{\sum_{1\le j,\ell\le d} \|h_{j\ell}\|_{\mathcal{H}}^2} \le C \right\}.
$$

We consider now $L_{M,\omega}$ and $J_{M,\omega}$ to be the unregularized approximate objectives (intuitively $F_{M,\omega}$ with $\eta = 0$):

$$
\begin{aligned}
L_{M,\omega}(\theta) \text{ [respectively } J_{M,\omega}(\theta)] = &\sum_{j=1}^{d} \left[\frac{T}{M} \sum_{n=1}^{M} \varphi_1 \left(\mu_j + \sum_{\ell=1}^{d} \sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(\tau_n - T_i^{(\ell)}\right) + b_{j\ell} \right) \mathbf{1}_{0<\tau_n - T_i^{(\ell)} \le A} \right) \right. \\
&\left. - \sum_{n=1}^{N_T^{(j)}} \varphi_2 \left(\mu_j + \sum_{\ell=1}^{d} \sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(T_n^{(j)} - T_i^{(\ell)}\right) + b_{j\ell} \right) \mathbf{1}_{0<T_n^{(j)} - T_i^{(\ell)} \le A} \right) \right].
\end{aligned}
$$

Proposition 4 (Approximation quality for maximum likelihood estimation)**.** *Consider Problem (P1) as a maximum likelihood problem and let*

$$
\hat{\theta} \in \arg\min_{\theta \in \Omega} L_{M,\omega}(\theta) \quad \textit{and} \quad \bar{\theta} \in \arg\min_{\theta \in \Omega} L(\theta).
$$

Assume that the kernel k is bounded (by some $\kappa > 0$) and L_k-Lipschitz continuous ($L_k > 0$):

$$\forall x \in \mathbb{R}: \quad k(x,x) \leq \kappa^2 \quad \textit{and} \quad \forall (y,y') \in \mathbb{R}^2, |k(x,y) - k(x,y')| \leq L_k |y - y'|,$$

and that $L(\hat{\theta}) < \infty$. Then, there exists $\delta > 0$ such that:

$$0 \leq L(\hat{\theta}) - L(\bar{\theta}) \leq \frac{T}{M} \left(L_k C d T N_T + 4(\kappa C + B) d N_T \right) + \frac{2 \log 2}{\omega} \left(dT + \frac{N_T}{\delta} \right).$$

Proposition 5 (Approximation quality for least-squares estimation)**.** *Consider Problem (P1) as a least-squares problem and let*

$$\hat{\theta} \in \arg\min_{\theta \in \Omega} J_{M,\omega}(\theta) \quad \textit{and} \quad \bar{\theta} \in \arg\min_{\theta \in \Omega} J(\theta).$$

Assume that the kernel k is bounded (by some $\kappa > 0$) and L_k-Lipschitz continuous ($L_k > 0$), and let $H = 2(B + (\kappa C + B) N_T + \log 2)$. Then, for all $\omega \geq 1$:

$$0 \leq J(\hat{\theta}) - J(\bar{\theta}) \leq \frac{HT}{M} \left(L_k C d T N_T + 4(\kappa C + B) d N_T \right) + \frac{4 \log 2}{\omega} \left(H d T + N_T \right).$$

Both propositions tell that the difference between the proposed estimator $\hat{\theta}$ (based on approximations) and the one coming from the original criterion $\bar{\theta}$ is controlled by two error terms. The first one depends on the Riemann approximation and vanishes as soon as the number of bins M grows. The second one is related to the link upper bound $\tilde{\varphi}$, and disappears when this approximation function comes closer to the ReLU function φ (*i.e.* ω becomes larger).

Proposition 4 is based on two mild assumptions. The first one, which also appears in Proposition 5, is using a bounded and Lipschitz continuous kernel k. This is the case for instance with the Gaussian kernel, $k : (x, x') \in \mathbb{R}^2 \mapsto \mathrm{e}^{-\gamma(x-x')^2}$ (where $\gamma > 0$), with $\kappa = 1$ and $L_k = \sqrt{\gamma}$. The second assumption says that the proposed estimator should have a finite log-likelihood, *i.e.* that the estimated conditional intensities $\lambda^{(j)}$ ($j \in [\![1, d]\!]$) are not null at their times $T_n^{(j)}$ ($n \in [\![1, N_T^{(j)}]\!]$). This is a rational requirement since a process cannot jump if its intensity is zero.

4 Numerical study

4.1 Synthetic data

This section aims first at assessing the impact of the approximation parameters ω and M, then at comparing our approach to the most related ones from the literature. For this purpose, we consider synthetic data coming from a 3-variate Hawkes process with baseline intensities $\mu_1 = \mu_2 = \mu_3 = 0.05$ and triggering kernels depicted in blue in Figure 2 and defined below for all $t \in \mathbb{R}_+$. Auto-interactions (g_{11}, g_{22} and g_{33}) reflect the refractory phenomenon:

$$\begin{aligned} g_{11}(t) &= (8t^2 - 1)\mathbf{1}_{t \leq 0.5} + \mathrm{e}^{-2.5(t-0.5)}\, \mathbf{1}_{t > 0.5} \\ g_{22}(t) = g_{33}(t) &= (8t^2 - 1)\mathbf{1}_{t \leq 0.5} + \mathrm{e}^{-(t-0.5)}\, \mathbf{1}_{t > 0.5}, \end{aligned}$$

while inter-interactions are either exciting or inhibiting:

$$\begin{aligned} g_{12}(t) &= \mathrm{e}^{-10(t-1)^2} \\ g_{13}(t) &= -0.6\, \mathrm{e}^{-3t^2} - 0.4\, \mathrm{e}^{-3(t-1)^2} \\ g_{21}(t) &= 2^{-5t} \\ g_{23}(t) &= -\, \mathrm{e}^{-2(t-3)^2} \\ g_{31}(t) &= -\, \mathrm{e}^{-5(t-2)^2} \\ g_{32}(t) &= (1 + \cos(\pi t))\, \mathrm{e}^{-t} /2. \end{aligned}$$

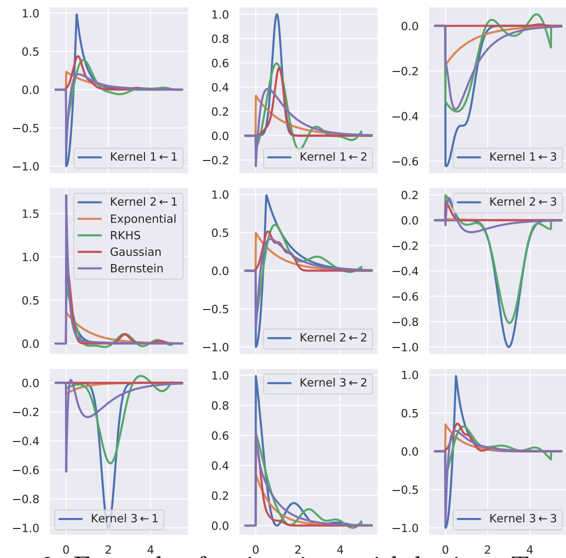

Figure 2: Example of estimations with horizon $T = 2000$.

The process is simulated thanks to the thinning method [Ogata, 1981] with a burn-in period ensuring stationarity and a sufficiently large time window, then estimations are performed successively for a growing horizon $T \in [250, 500, 1000, 2000]$ of the same observed trajectory (see Figures 4 and 5).

Methods compared are (see estimation examples in Figure 2):

Exponential: the mainly used parametric model, for which triggering kernels are supposed to be exponential functions [Bonnet et al., 2023]. It allows for estimating exciting or inhibiting interactions;

RKHS: our proposition (with a regularization coefficient η) based on the Gaussian kernel (parametrized by γ) and a support bound $A = 5$;

Bernstein: nonparametric estimation, for which triggering kernels are represented as a sum of exponential functions [Lemonnier and Vayatis, 2014]: $g_{j\ell}(t) = \sum_{u=1}^{U} \alpha_u^{(j\ell)} \,\mathrm{e}^{-\gamma u t}$ (with $U = 10$) and a quadratic penalty on the coefficients $\alpha_u^{(j\ell)}$ controlled by the parameter η ;

Gaussian: nonparametric estimation, for which triggering kernels are represented as a sum of Gaussian functions [Xu et al., 2016]: $g_{j\ell}(t) = \sum_{u=1}^{U} \alpha_u^{(j\ell)} \,\mathrm{e}^{-\gamma(t-t_u)^2}$ (with $U = 10$ and a regular grid on $[0, A]$) and a quadratic penalty on the coefficients $\alpha_u^{(j\ell)}$ controlled by the parameter η.

Since the closest methods to ours are based on minimizing the negative log-likelihood, we only include our approach with the criterion L. Numerical optimization is performed thanks to the L-BFGS-B method implemented in SciPy [Virtanen et al., 2020] on a personal computer. In addition, let us remark that the method by Lemonnier and Vayatis [2014] minimizes an approximation of the negative log-likelihood and that the one by Xu et al. [2016] is restricted to exciting interactions ($\alpha_u^{(j\ell)} \geq 0$).

All three nonparametric methods are tuned by parameters γ and η, respectively chosen on the grids $[1, 10, 100]$ and $[0.1, 1, 10, 100]$ to maximize the log-likelihood computed on an independent validation trajectory (with same distribution as the training one). The numerical results presented below are computed on an independent test trajectory, and this whole procedure is repeated 10 times to produce the statistics (mean and 95% confidence interval) reported in Figures 3 to 5.

As a first numerical experiment, Figure 3 depicts the the smallest (over the grids of parameters η and γ) sum of L^1-errors between the true triggering kernels and the estimation provided by our method **RKHS**, for a horizon $T = 1000$ and for different values of approximation parameters ω

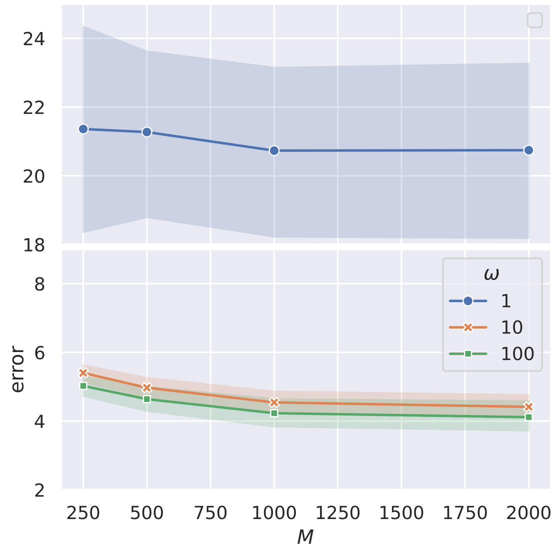

Figure 3: Approximation error of true kernels with respect to the hyperparameters ω and M.

and M. As suggested by Proposition 4, the bigger ω and the bigger M, the lower the estimation error. This shows that the considered approximations are not statistically harmful when parameters are well chosen. For the subsequent numerical analysis, the values chosen are $\omega = 100$ and $M = \max(1000, 2\max(N_T^{(1)}, N_T^{(2)}, N_T^{(3)}))$.

Now, Figures 4 and 5 display, for the considered methods, respectively the sum of L^1-errors and the test log-likelihood, for the best model according to the validation log-likelihood. It appears clearly that competing methods perform poorly in estimating the chosen triggering kernels. Indeed, while able to infer inhibiting effects, **Exponential** is based on a misspecified model in view of the interaction functions to estimate, exhibiting non-exponential shapes. The same reasoning is true for **Gaussian** since it is designed only for exciting interactions. At last, **Bernstein** suffers from the approximated log-likelihood it optimizes, as well as the kind of functions it is able to approximate. In particular, Figure 2 illustrates that it is able to recover auto-interactions with a refractory period but not cross-interactions. Let us remark that for **Gaussian** and **Bernstein**, augmenting the size U of the linear combination does not help because methods are limited by intrinsic misspecifications. Overall, an RKHS based method able to handle complex kernel functions, in particular that combine exciting and inhibiting interactions, with immediate or delayed effects, seems to be the best option.

4.2 Neuronal data

We illustrate our procedure on a neuronal dataset described in [Petersen and Berg, 2016, Radosevic et al., 2019], then analyzed in [Bonnet et al., 2023] with **Exponential** to study the neuronal activity of a red-eared turtle. The full dataset contains 10 recordings of 40ms for 250 neurons, that we preprocess as follows. In order to work with a process that is almost stationary, we only consider the events that took place on the interval $[11, 24]$ms, using a similar setting as in [Bonnet et al., 2023]. This procedure allows in particular to remove the effects of external stimuli that are performed before each recording. For the sake of clarity when displaying interaction functions, we focus on a small subnetwork of five neurons, chosen randomly among the neurons with a large enough number of spikes (at least 500 spikes in the concatenation of all recordings). Parameters γ and η are chosen to maximize the likelihood on a validation trajectory obtained by a randomized concatenation of half of the recordings.

The five estimated auto-interaction functions are displayed in Figure 6, with **Exponential** as a

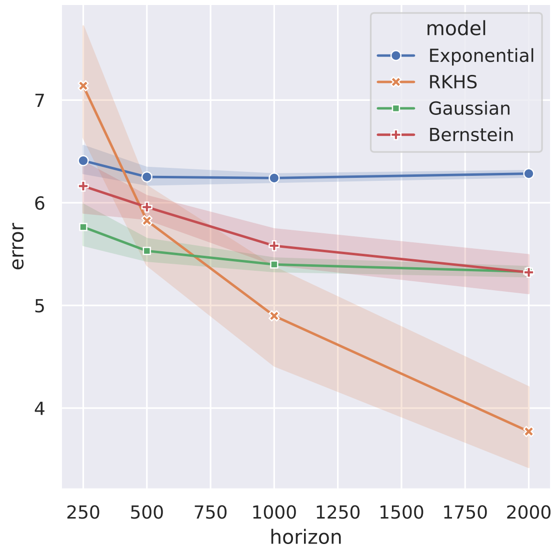

Figure 4: Approximation error of true kernels with respect to the horizon T.

reference. The estimation provided by our method is consistent with current knowledge of neurons' behaviour, exhibiting first a large negative peak corresponding to the refractory period followed by a self-exciting effect. As expected, **Exponential** cannot detect such a structure and only shows the self-exciting effect, which is probably underestimated due to the lack of refractory period modeling (see for instance the second graph where the exponential method only estimates a small self-interaction, which corresponds to an averaged version of sharp negative and positive peaks). We provide in Figure 8 of Appendix F all 25 interaction functions obtained with the four benchmarked methods. We see that the two methods that are able to detect accurately the refractory period are **Bernstein** and **RKHS**, which is consistent with the results obtained on synthetic data. Interestingly, many cross-interactions appear to be low compared to auto-interactions. Besides this qualitative assessment, Table 1 of Appendix F gives the log-likelihood scores computed on a test trajectory obtained by a randomized concatenation of half of the recordings. It shows that **RKHS** has the highest score, which leads to believe that it is more suited than competitors.

5 Discussion

In this paper, we filled a gap in nonparametric inference of Hawkes processes by introducing a batch learning method based on RKHSs, which is able to handle combinations of exciting and inhibiting effects.

On the one hand, as a kernel method, the proposed approach has for drawback its computational complexity, which, even if lower than that of neural networks based techniques, is higher than parametric or equivalent models. This limitation could be lifted thanks to kernel approximation techniques (such as the Nyström method or random Fourier features [Yang et al., 2012]) or discretization schemes [Staerman et al., 2023].

On the other hand, we believe that the proposed method can serve as a stepping stone for many future developments, including sparse learning of interaction functions, the use of operator-valued kernels [Micchelli and Pontil, 2005, Paulsen and Raghupathi, 2016, Brault et al., 2016] and extension to inference of spatiotemporal Hawkes processes [Li and Cui, 2024, Siviero et al., 2024].

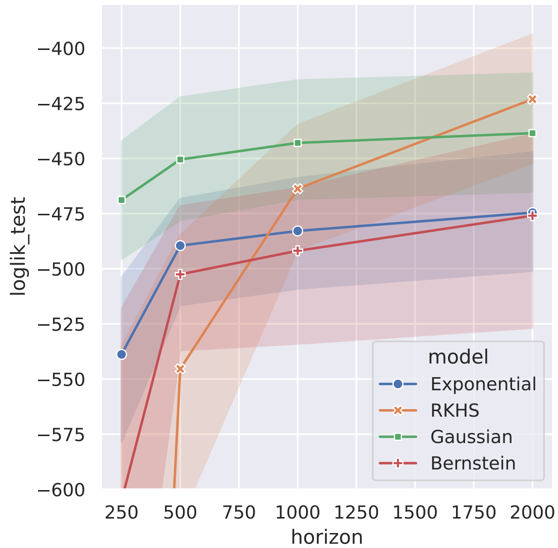

Figure 5: Test log-likelihood with respect to the horizon T.

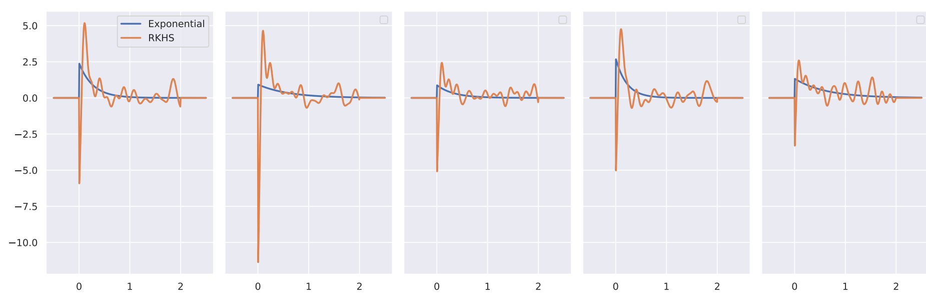

Figure 6: Auto-interaction functions learned on the neuronal subnetwork.

Acknowledgement

The authors are grateful to the anonymous referees for their very relevant comments.

References

R. P. Adams, I. Murray, and D. J. C. MacKay. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 9–16, New York, NY, USA, 2009. Association for Computing Machinery.

E. Bacry and J.-F. Muzy. First- and Second-Order Statistics Characterization of Hawkes Processes and Non-Parametric Estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016.

E. Bacry, K. Dayri, and J. F. Muzy. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *Eur. Phys. J. B*, 85(5):157, 2012.

E. Bacry, M. Bompaire, S. Gaïffas, and J.-F. Muzy. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32, 2020.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

R. W. Berg and S. Ditlevsen. Synaptic inhibition and excitation estimated via the time constant of membrane potential fluctuations. *J Neurophysiol*, 110(4):1021–1034, 2013.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, Boston, MA, 2004.

A. Bonnet, M. Martinez Herrera, and M. Sangnier. Inference of multivariate exponential Hawkes processes with inhibition and application to neuronal activity. *Stat Comput*, 33(4):91, 2023.

R. Brault, M. Heinonen, and F. Buc. Random Fourier Features For Operator-Valued Kernels. In *Proceedings of The 8th Asian Conference on Machine Learning*, pages 110–125. PMLR, 2016.

R. Browning, J. Rousseau, and K. Mengersen. A flexible, random histogram kernel for discrete-time Hawkes processes, 2022.

P. Brémaud and L. Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, 24 (3):1563–1588, 1996.

O. Chapelle. Training a Support Vector Machine in the Primal. *Neural Computation*, 19(5):1155–1178, 2007.

F. Chen and P. Hall. Nonparametric Estimation for Self-Exciting Point Processes—A Parsimonious Approach. *Journal of Computational and Graphical Statistics*, 25(1):209–224, 2016.

M. Costa, C. Graham, L. Marsalle, and V. C. Tran. Renewal in Hawkes processes with self-excitation and inhibition. *Advances in Applied Probability*, 52(3):879–915, 2020.

J. P. Cunningham, K. V. Shenoy, and M. Sahani. Fast Gaussian process methods for point process intensity estimation. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 192–199, New York, NY, USA, 2008. Association for Computing Machinery.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Elementary Theory and Methods*. Springer-Verlag, New York, 2003.

I. Deutsch and G. J. Ross. Estimating Product Cannibalisation in Wholesale using Multivariate Hawkes Processes with Inhibition, 2022.

S. Donnet, V. Rivoirard, and J. Rousseau. Nonparametric Bayesian estimation for multivariate Hawkes processes. *The Annals of Statistics*, 48(5):2698–2727, 2020.

N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song. 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*, KDD '16, pages 1555–1564, New York, NY, USA, 2016. Association for Computing Machinery.

C. Duval, E. Luçon, and C. Pouzat. Interacting Hawkes processes with multiplicative inhibition. *Stochastic Processes and their Applications*, 148:180–226, 2022.

M. Eichler, R. Dahlhaus, and J. Dueck. Graphical Modeling for Multivariate Hawkes Processes with Nonparametric Link Functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

S. Flaxman, Y. W. Teh, and D. Sejdinovic. Poisson intensity estimation with reproducing kernels. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 270–279. PMLR, 2017a.

S. Flaxman, Y. W. Teh, and D. Sejdinovic. Poisson intensity estimation with reproducing kernels. *Electronic Journal of Statistics*, 11(2):5081–5104, 2017b.

N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

A. G. Hawkes. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58 (1):83–90, 1971.

A. G. Hawkes and D. Oakes. A Cluster Process Representation of a Self-Exciting Process. *Journal of Applied Probability*, 11(3):493–503, 1974.

S. Joseph and S. Jain. A neural network based model for multi-dimensional non-linear Hawkes processes. *Journal of Computational and Applied Mathematics*, 447:115889, 2024a.

S. Joseph and S. Jain. Non-Parametric Estimation of Multi-dimensional Marked Hawkes Processes, 2024b.

S. Joseph, L. D. Kashyap, and S. Jain. Shallow Neural Hawkes: Non-parametric kernel estimation for Hawkes processes. *Journal of Computational Science*, 63:101754, 2022.

M. Kirchner. An estimation procedure for the Hawkes process. *Quantitative Finance*, 17(4):571–595, 2017.

M. Kirchner and A. Bercher. A nonparametric estimation procedure for the Hawkes process: Comparison with maximum likelihood estimation. *Journal of Statistical Computation and Simulation*, 88(6):1106–1116, 2018.

D. Kurisu. Discretization of Self-Exciting Peaks Over Threshold Models, 2017.

R. Lemonnier and N. Vayatis. Nonparametric Markovian Learning of Triggering Kernels for Mutually Exciting and Mutually Inhibiting Multivariate Hawkes Processes. In T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 161–176, Berlin, Heidelberg, 2014. Springer.

E. Lewis and G. Mohler. A Nonparametric EM Algorithm for Multiscale Hawkes Processes. *Journal of Nonparametric Statistics*, 2011.

C. Li and K. Cui. Multivariate Hawkes processes with spatial covariates for spatiotemporal event data analysis. *Ann Inst Stat Math*, 76(4):535–578, 2024.

W. Lian, R. Henao, V. Rao, J. Lucas, and L. Carin. A multitask point process predictive model. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2030–2038, Lille, France, 2015. JMLR.org.

S. Linderman and R. Adams. Discovering Latent Network Structure in Point Process Data. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1413–1421. PMLR, 2014.

S. W. Linderman and R. P. Adams. Scalable Bayesian Inference for Excitatory Point Process Networks, 2015.

C. Lloyd, T. Gunter, M. Osborne, and S. Roberts. Variational Inference for Gaussian Process Modulated Poisson Processes. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1814–1822. PMLR, 2015.

R. E. Lovelace and S. J. Myers. Chapter 10 - Nerve Conduction and Neuromuscular Transmission. In J. A. Downey, S. J. Myers, E. G. Gonzalez, and J. S. Lieberman, editors, *The Physiological Basis of Rehabilitation Medicine (Second Edition)*, pages 215–242. Butterworth-Heinemann, 1994.

J. Luo, H. Qiao, and B. Zhang. Learning with smooth Hinge losses. *Neurocomputing*, 463:379–387, 2021.

D. Marsan and O. Lengliné. Extending earthquakes' reach through cascading. *Science*, 319(5866): 1076–1079, 2008.

H. Mei and J. Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6757–6767, Red Hook, NY, USA, 2017. Curran Associates Inc.

H. Mei, C. Yang, and J. Eisner. Transformer Embeddings of Irregularly Spaced Events and Their Participants. In *International Conference on Learning Representations*, 2022.

Z. Meng, K. Wan, Y. Huang, Z. Li, Y. Wang, and F. Zhou. Interpretable Transformer Hawkes Processes: Unveiling Complex Interactions in Social Networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 2200–2211, New York, NY, USA, 2024. Association for Computing Machinery.

C. A. Micchelli and M. Pontil. On Learning Vector-Valued Functions. *Neural Computation*, 17(1): 177–204, 2005.

G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-Exciting Point Process Modeling of Crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

J. Møller, A. R. Syversveen, and R. P. Waagepetersen. Log Gaussian Cox Processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

Y. Ogata. Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

J. Olinde and M. B. Short. A Self-limiting Hawkes Process: Interpretation, Estimation, and Use in Crime Modeling. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3212–3219, 2020.

T. Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Ann Inst Stat Math*, 31(1):145–155, 1979.

Z. Pan, Z. Wang, J. M. Phillips, and S. Zhe. Self-Adaptable Point Processes with Nonparametric Time Decays. In *Advances in Neural Information Processing Systems*, volume 34, pages 4594–4606. Curran Associates, Inc., 2021.

V. I. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2016.

P. C. Petersen and R. W. Berg. Lognormal firing rate distribution reveals prominent fluctuation-driven regime in spinal motor networks. *Elife*, 5:e18805, 2016.

M. Radosevic, A. Willumsen, P. C. Petersen, H. Lindén, M. Vestergaard, and R. W. Berg. Decoupling of timescales reveals sparse convergent CPG network in the adult spinal cord. *Nat Commun*, 10(1): 2937, 2019.

P. Reynaud-Bouret and S. Schbath. Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822, 2010.

P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Inference of functional connectivity in Neurosciences via Hawkes processes. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 317–320, 2013.

P. Reynaud-Bouret, V. Rivoirard, F. Grammont, and C. Tuleau-Malot. Goodness-of-Fit Tests and Nonparametric Adaptive Estimation for Spike Train Analysis. *J Math Neurosci*, 4(1):3, 2014.

Y.-L. K. Samo and S. Roberts. Scalable nonparametric Bayesian inference on point processes with Gaussian processes. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2227–2236, Lille, France, 2015. JMLR.org.

O. Shchur, M. Biloš, and S. Günnemann. Intensity-Free Learning of Temporal Point Processes. In *International Conference on Learning Representations*, 2019.

E. Siviero, G. Staerman, S. Clémençon, and T. Moreau. Flexible Parametric Inference for Space-Time Hawkes Processes, 2024.

G. Staerman, C. Allain, A. Gramfort, and T. Moreau. FaDIn: Fast discretized inference for Hawkes processes with general parametric kernels. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 32575–32597, Honolulu, Hawaii, 2023. PMLR.

D. Sulem, V. Rivoirard, and J. Rousseau. Scalable and adaptive variational Bayes methods for Hawkes processes, 2023.

D. Sulem, V. Rivoirard, and J. Rousseau. Bayesian estimation of nonlinear Hawkes processes. *Bernoulli*, 30(2):1257–1286, 2024.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods*, 17(3):261–272, 2020.

X. Wang, Z. Yang, X. Chen, and W. Liu. Distributed Inference for Linear Support Vector Machine. *Journal of Machine Learning Research*, 20(113):1–41, 2019.

D. Wu, T. Ide, G. Kollias, J. Navratil, A. Lozano, N. Abe, Y. Ma, and R. Yu. Learning Granger Causality from Instance-wise Self-attentive Hawkes Processes. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 415–423. PMLR, 2024.

H. Xu, M. Farajtabar, and H. Zha. Learning Granger Causality for Hawkes Processes. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1717–1726. PMLR, 2016.

T. Yang, Y.-f. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou. Nyström Method vs Random Fourier Features: A Theoretical and Empirical Comparison. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Y. Yang, J. Etesami, N. He, and N. Kiyavash. Online Learning for Multivariate Hawkes Processes. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Q. Zhang, A. Lipani, O. Kirnap, and E. Yilmaz. Self-Attentive Hawkes Process. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11183–11193. PMLR, 2020a.

R. Zhang, C. Walder, M.-A. Rizoiu, and L. Xie. Efficient Non-parametric Bayesian Hawkes Processes. pages 4299–4305, 2019.

R. Zhang, C. Walder, and M.-A. Rizoiu. Variational Inference for Sparse Gaussian Process Modulated Hawkes Process. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6803–6810, 2020b.

F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, and F. Chen. Scalable Inference for Nonparametric Hawkes Process Using P\'{o}lya-Gamma Augmentation, 2019.

F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, and F. Chen. Efficient Inference for Nonparametric Hawkes Processes Using Auxiliary Latent Variables. *Journal of Machine Learning Research*, 21(241): 1–31, 2020.

F. Zhou, S. Luo, Z. Li, X. Fan, Y. Wang, A. Sowmya, and F. Chen. Efficient EM-variational inference for nonparametric Hawkes process. *Statistics and Computing*, 31(4):46, 2021.

K. Zhou, H. Zha, and L. Song. Learning Triggering Kernels for Multi-dimensional Hawkes Processes. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1301–1309. PMLR, 2013.

S. Zuo, H. Jiang, Z. Li, T. Zhao, and H. Zha. Transformer Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pages 11692–11702. JMLR.org, 2020.

A Proof of Proposition 1

The optimization problem of interest can be written:

$$\begin{aligned}
\underset{\substack{\theta\in\Theta,\\ (\xi_j)_{1\le j\le d}\in(\mathbb{R}^M)^d}}{\text{minimize}} \quad & \sum_{j=1}^{d}\left[\frac{T}{M}\|\xi_j\|^2 - 2N_T^{(j)}\mu_j - 2\sum_{\ell=1}^{d}\sum_{\substack{1\le n\le N_T^{(j)}\\ 1\le i\le N_T^{(\ell)}}}\left(h_{j\ell}\left(T_n^{(j)}-T_i^{(\ell)}\right)+b_{j\ell}\right)\mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A}\right] \\
& +\frac{\eta}{2}\sum_{1\le j,\ell\le d}\|h_{j\ell}\|_{\mathcal{H}}^2 \\
\text{s. t.} \quad & \begin{cases}
\forall (j,\ell)\in[\![1,d]\!]^2 & \\
\mu_j \ge 0 & : \delta_j \ge 0\\
\forall n\in[\![1,P]\!],\ h_{j\ell}(x_n)+b_{j\ell}\ge 0 & : \beta_n^{(j\ell)}\ge 0\\
\forall n\in[\![1,M]\!],\ \mu_j+\displaystyle\sum_{\ell=1}^{d}\sum_{i=1}^{N_T^{(\ell)}}\left(h_{j\ell}\left(\tau_n-T_i^{(\ell)}\right)+b_{j\ell}\right)\mathbf{1}_{0<\tau_n-T_i^{(\ell)}\le A}=\xi_{jn} & : \alpha_n^{(j)}\in\mathbb{R},
\end{cases}
\end{aligned}$$

where Lagrangian variables are indicated on the right.

This problem is convex (in that it has a quadratic objective with affine constraints) and statisfies Slater's constraint qualification. As a consequence, the dual problem is attained for some dual variables $\delta\in\mathbb{R}_+^d$, $\beta^{(j\ell)}\in\mathbb{R}_+^P$ and $\alpha^{(j)}\in\mathbb{R}^M$, and Karush-Kuhn-Tucker conditions hold for solutions θ and $(\xi_j)_{1\le j\le d}$, and δ, $\beta^{(j\ell)}$ and $\alpha^{(j)}$. In particular, by stationarity, for all $(j,\ell)\in[\![1,d]\!]^2$:

$$h_{j\ell}=\eta^{-1}\left[2q_{j\ell}+\sum_{n=1}^{P}\beta_n^{(j\ell)}k(\cdot,x_n)-\sum_{n=1}^{M}\alpha_n^{(j)}r_{\ell n}\right],$$

where

$$\begin{cases} q_{j\ell} = \sum_{\substack{1\le n\le N_T^{(j)}\\ 1\le i\le N_T^{(\ell)}}} k\left(\cdot, T_n^{(j)} - T_i^{(\ell)}\right) \mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A} \\ r_{\ell n} = \sum_{i=1}^{N_T^{(\ell)}} k\left(\cdot, \tau_n - T_i^{(\ell)}\right) \mathbf{1}_{0<\tau_n - T_i^{(\ell)}\le A}. \end{cases}$$

As a byproduct, the dual problem reads:

$$\begin{aligned} \underset{(\alpha^{(j)})_{1\le j\le d}\in(\mathbb{R}^M)^d,\,(\beta^{(j\ell)})_{1\le j\le d}\in(\mathbb{R}^P)^{d\times d}}{\text{maximize}} \quad & -\sum_{j=1}^d \left[{\alpha^{(j)}}^\top \tilde{K}\alpha^{(j)} - \frac{2}{\eta}{\alpha^{(j)}}^\top \sum_{\ell=1}^d v_{j\ell} \right] \\ & -\frac{1}{2\eta}\sum_{1\le j,\ell\le d} \left[{\beta^{(j\ell)}}^\top R\beta^{(j\ell)} + 2{\beta^{(j\ell)}}^\top \left(2u_{j\ell} - W_\ell \alpha^{(j)}\right) \right] \\ \text{s. t.} \quad & \begin{cases} \forall (j,\ell)\in[\![1,d]\!]^2, \beta^{(j\ell)} \succcurlyeq 0 \\ \forall j\in[\![1,d]\!], 2N_T^{(j)} \le \mathbb{1}^\top \alpha^{(j)} \\ \forall (j,\ell)\in[\![1,d]\!]^2, {\alpha^{(j)}}^\top z_\ell - \mathbb{1}^\top \beta^{(j\ell)} = 2\sum_{\substack{1\le n\le N_T^{(j)}\\ 1\le i\le N_T^{(\ell)}}} \mathbb{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A}, \end{cases} \end{aligned}$$

where $R = (k(x_n, x_{n'}))_{1\le n,n'\le P}$, $\tilde{K} = \frac{M}{4T} I_M + \frac{1}{2\eta}\sum_{\ell=1}^d K_\ell$ (with I_M the identity matrix of size M), and for all $\ell\in[\![1,d]\!]$, $K_\ell = (\langle r_{\ell n}, r_{\ell n'}\rangle_{\mathcal{H}})_{1\le n,n'\le M}$, $W_\ell = (r_{\ell n'}(x_n))_{1\le n,n'\le M}$, $z_\ell = \left(\sum_{i=1}^{N_T^{(\ell)}} \mathbb{1}_{0<\tau_n - T_i^{(\ell)}\le A}\right)_{1\le n\le M}$, and for all $j\in[\![1,d]\!]$, $u_{j\ell} = (q_{j\ell}(x_n))_{1\le n\le P}$, and $v_{j\ell} = \left(\langle q_{j\ell}, r_{\ell n}\rangle_{\mathcal{H}}\right)_{1\le n\le M}$.

B Proof of Proposition 2

First of all, conditional intensities read:

$$\begin{aligned} \forall j\in[\![1,d]\!], \forall t\in\mathbb{R}_+: \quad \lambda^{(j)}(t) &= \varphi\left(\mu_j + \sum_{\ell=1}^d \sum_{i=1}^{N_t^{(\ell)}} g_{j\ell}\left(t - T_i^{(\ell)}\right)\right) \\ &= \varphi\left(\mu_j + \sum_{\ell=1}^d \sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(t - T_i^{(\ell)}\right) + b_{j\ell}\right) \mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right). \end{aligned}$$

Then, the approximation L_0 is [Lemonnier and Vayatis, 2014]:

$$\forall \theta\in\Theta^+, \quad L_0(\theta) = \sum_{j=1}^d \left[\int_0^T \left\{ \mu_j + \sum_{\ell=1}^d \sum_{i=1}^{N_T^{(\ell)}} g_{j\ell}\left(t - T_i^{(\ell)}\right) \right\} \mathrm{d}t - \sum_{n=1}^{N_T^{(j)}} \log\left(\lambda^{(j)}\left(T_n^{(j)}\right)\right) \right],$$

which is a lower bound of L when φ is the ReLU function. Now, let $F:\Theta\to\mathbb{R}\cup\{\infty\}$ denote the objective function: for all $\theta\in\Theta^+$,

$$\begin{aligned} F(\theta) &= L_0(\theta) + \frac{\eta}{2}\sum_{1\le j,\ell\le d} \|h_{j\ell}\|_{\mathcal{H}}^2 \\ &= \sum_{j=1}^d \left[\mu_j T + \sum_{\ell=1}^d \sum_{i=1}^{N_T^{(\ell)}} \int_0^T h_{j\ell}\left(t - T_i^{(\ell)}\right) \mathbf{1}_{0<t-T_i^{(\ell)}\le A}\,\mathrm{d}t + \sum_{\ell=1}^d b_{j\ell} B^{(\ell)} \right. \\ &\left. - \sum_{n=1}^{N_T^{(j)}} \log\left(\mu_j + \sum_{\ell=1}^d \sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(T_n^{(j)} - T_i^{(\ell)}\right) + b_{j\ell}\right) \mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A}\right) \right] + \frac{\eta}{2}\sum_{1\le j,\ell\le d}\|h_{j\ell}\|_{\mathcal{H}}^2, \end{aligned}$$

where $B^{(\ell)} = \sum_{i=1}^{N_T^{(\ell)}} \min\left(T - T_i^{(\ell)}, A\right)$.

Let θ be a minimizer of F. By the reproducing property, we have:

$$
\begin{aligned}
F(\theta) = &\sum_{j=1}^{d}\Bigg[\mu_j T + \sum_{\ell=1}^{d}\sum_{i=1}^{N_T^{(\ell)}} \int_0^T \left\langle h_{j\ell}, k\left(\cdot, t - T_i^{(\ell)}\right) \mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right\rangle_{\mathcal{H}} \mathrm{d}t + \sum_{\ell=1}^{d} b_{j\ell} B^{(\ell)} \\
&- \sum_{n=1}^{N_T^{(j)}} \log\left(\mu_j + \sum_{\ell=1}^{d}\sum_{i=1}^{N_T^{(\ell)}}\left(\left\langle h_{j\ell}, k\left(\cdot, T_n^{(j)} - T_i^{(\ell)}\right)\mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A}\right\rangle_{\mathcal{H}} + b_{j\ell}\mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A}\right)\right)\Bigg] \\
&+ \frac{\eta}{2}\sum_{1\le j,\ell\le d}\|h_{j\ell}\|_{\mathcal{H}}^2 \\
= &\sum_{j=1}^{d}\Bigg[\mu_j T + \sum_{\ell=1}^{d}\int_0^T \left\langle h_{j\ell}, \sum_{i=1}^{N_T^{(\ell)}} k\left(\cdot, t - T_i^{(\ell)}\right)\mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right\rangle_{\mathcal{H}} \mathrm{d}t + \sum_{\ell=1}^{d} b_{j\ell}B^{(\ell)} \\
&- \sum_{n=1}^{N_T^{(j)}} \log\left(\mu_j + \sum_{\ell=1}^{d}\left(\left\langle h_{j\ell}, \sum_{i=1}^{N_T^{(\ell)}} k\left(\cdot, T_n^{(j)} - T_i^{(\ell)}\right)\mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A}\right\rangle_{\mathcal{H}} + b_{j\ell}\sum_{i=1}^{N_T^{(\ell)}}\mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A}\right)\right)\Bigg] \\
&+ \frac{\eta}{2}\sum_{1\le j,\ell\le d}\|h_{j\ell}\|_{\mathcal{H}}^2
\end{aligned}
$$

Now, for all $\ell \in [\![1, d]\!]$, $L_\ell : h \in \mathcal{H} \mapsto \int_0^T \left\langle h, \sum_{i=1}^{N_T^{(\ell)}} k\left(\cdot, t - T_i^{(\ell)}\right)\mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right\rangle_{\mathcal{H}} \mathrm{d}t$ is a continuous linear operator from $\mathcal{H}$ to $\mathbb{R}$ so by the Riesz representation theorem, there exists $r_\ell \in \mathcal{H}$ such that for all $h \in \mathcal{H}$, $L_\ell(h) = \langle h, r_\ell\rangle_{\mathcal{H}}$. Moreover, by the reproducing property, for all $x \in \mathbb{R}$:

$$
\begin{aligned}
r_\ell(x) = \langle r_\ell, k(\cdot, x)\rangle_{\mathcal{H}} = L_\ell\left(k(\cdot, x)\right) &= \int_0^T \left\langle k(\cdot, x), \sum_{i=1}^{N_T^{(\ell)}} k\left(\cdot, t - T_i^{(\ell)}\right)\mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right\rangle_{\mathcal{H}} \mathrm{d}t \\
&= \int_0^T \sum_{i=1}^{N_T^{(\ell)}} k\left(x, t - T_i^{(\ell)}\right)\mathbf{1}_{0<t-T_i^{(\ell)}\le A}\,\mathrm{d}t.
\end{aligned}
$$

Then, denoting for all $n \in [\![1, N_T^{(j)}]\!]$:

$$
q_{nj\ell} = \sum_{i=1}^{N_T^{(\ell)}} k\left(\cdot, T_n^{(j)} - T_i^{(\ell)}\right)\mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A},
$$

the objective function reads:

$$
\begin{aligned}
F(\theta) = &\sum_{j=1}^{d}\Bigg[\mu_j T + \sum_{\ell=1}^{d}\langle h_{j\ell}, r_\ell\rangle_{\mathcal{H}} + \sum_{\ell=1}^{d} b_{j\ell}B^{(\ell)} \\
&- \sum_{n=1}^{N_T^{(j)}} \log\left(\mu_j + \sum_{\ell=1}^{d}\left(\langle h_{j\ell}, q_{nj\ell}\rangle_{\mathcal{H}} + b_{j\ell}\sum_{i=1}^{N_T^{(\ell)}}\mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A}\right)\right)\Bigg] + \frac{\eta}{2}\sum_{1\le j,\ell\le d}\|h_{j\ell}\|_{\mathcal{H}}^2.
\end{aligned}
$$

Thus, still for all $(j, \ell) \in [\![1, d]\!]^2$, decomposing $\mathcal{H}$ as the direct sum of the vector space $\mathcal{V}_{j\ell}$ spanned by $\{r_\ell\} \cup \left\{q_{uj\ell}, u \in [\![1, N_T^{(j)}]\!]\right\}$ and its orthogonal subspace $\mathcal{V}_{j\ell}^{\perp}$, we can write $h_{j\ell} = h_{j\ell}^{\parallel} + h_{j\ell}^{\perp}$ with

$h^{\|}_{j\ell} \in \mathcal{V}_{j\ell}$ and $h_{j\ell} \in \mathcal{V}^{\perp}_{j\ell}$. Then, for $\theta^{\|} = \Big((\mu_j)_{1\le j\le d}, (h^{\|}_{j\ell})_{1\le j,\ell\le d}, (b_{j\ell})_{1\le j,\ell\le d}\Big)$, we have $\theta^{\|} \in \Theta$ and

$$\begin{aligned}\min_{\theta'\in\Theta} F(\theta') = F(\theta) = L_0(\theta) + \frac{\eta}{2}\sum_{1\le j,\ell\le d}\|h_{j\ell}\|^2_{\mathcal{H}} = L_0(\theta^{\|}) + \frac{\eta}{2}\sum_{1\le j,\ell\le d}\|h_{j\ell}\|^2_{\mathcal{H}} \\ \ge L_0(\theta^{\|}) + \frac{\eta}{2}\sum_{1\le j,\ell\le d}\|h^{\|}_{j\ell}\|^2_{\mathcal{H}} = F(\theta^{\|}),\end{aligned}$$

by the Pythagorean theorem. This proves that $\theta^{\|}$ is a minimizer of F, which has the desired form:

$$\forall (j,\ell)\in [\![1,d]\!]^2 : \quad h^{\|}_{j\ell} = \alpha_0^{(j\ell)} r_\ell + \sum_{u=1}^{N_T^{(j)}} \alpha_u^{(j\ell)} q_{uj\ell},$$

for some $(N_T^{(j)}+1)$ real values $\alpha_0^{(j\ell)}, \dots, \alpha^{(j\ell)}_{N_T^{(j)}}$.

C Proof of Proposition 3

Following Lemonnier and Vayatis [2014], we consider the approximation J_0:

$$\forall\theta\in\Theta^+, \quad J_0(\theta) = \sum_{j=1}^d \left[\int_0^T \left\{\mu_j + \sum_{\ell=1}^d \sum_{i=1}^{N_T^{(\ell)}} g_{j\ell}\left(t - T_i^{(\ell)}\right)\right\}^2 \mathrm{d}t - 2\sum_{n=1}^{N_T^{(j)}} \lambda^{(j)}\left(T_n^{(j)}\right)\right],$$

which is an upper bound of J when φ is the ReLU function. Now, it is enough to remark that for any candidate $\theta\in\Theta^+$ and all $j\in[\![1,d]\!]$,

$$\begin{aligned}
&\int_0^T \left\{\mu_j + \sum_{\ell=1}^d \sum_{i=1}^{N_T^{(\ell)}} g_{j\ell}\left(t - T_i^{(\ell)}\right)\right\}^2 \mathrm{d}t \\
&= \int_0^T \left\{\mu_j + \sum_{\ell=1}^d \sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(t - T_i^{(\ell)}\right) + b_{j\ell}\right)\mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right\}^2 \mathrm{d}t \\
&= \mu_j^2 T + \int_0^T \left\{\sum_{\ell=1}^d \left(\left\langle h_{j\ell}, \sum_{i=1}^{N_T^{(\ell)}} k(\cdot, t - T_i^{(\ell)})\mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right\rangle_{\mathcal{H}} + \sum_{i=1}^{N_T^{(\ell)}} b_{j\ell}\mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right)\right\}^2 \mathrm{d}t \\
&+ 2\mu_j\left(\sum_{\ell=1}^d \int_0^T \left\langle h_{j\ell}, \sum_{i=1}^{N_T^{(\ell)}} k\left(\cdot, t - T_i^{(\ell)}\right)\mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right\rangle_{\mathcal{H}} \mathrm{d}t + \sum_{\ell=1}^d b_{j\ell} B^{(\ell)}\right) \\
&\le \mu_j^2 T + \sum_{1\le\ell,\ell'\le d} C_{\ell\ell'}\|h_{j\ell}\|_{\mathcal{H}}\|h_{j\ell'}\|_{\mathcal{H}} + \int_0^T \left(\sum_{\ell=1}^d b_{j\ell}\sum_{i=1}^{N_T^{(\ell)}} \mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right)^2 \mathrm{d}t \\
&+ 2\sum_{1\le\ell,\ell'\le d} D_{\ell\ell'}|b_{j\ell'}|\|h_{j\ell}\|_{\mathcal{H}} + 2\mu_j\sum_{\ell=1}^d \int_0^T \left\langle h_{j\ell}, \sum_{i=1}^{N_T^{(\ell)}} k(\cdot, t - T_i^{(\ell)})\mathbf{1}_{0<t-T_i^{(\ell)}\le A}\right\rangle_{\mathcal{H}} \mathrm{d}t \\
&+ 2\mu_j\sum_{\ell=1}^d b_{j\ell}B^{(\ell)}, \qquad (1)
\end{aligned}$$

by Cauchy–Schwarz inequality, where

$$C_{\ell\ell'} = \int_0^T \left\| \sum_{i=1}^{N_T^{(\ell)}} k(\cdot, t - T_i^{(\ell)}) \mathbf{1}_{0<t-T_i^{(\ell)}\le A} \right\|_{\mathcal{H}} \left\| \sum_{i=1}^{N_T^{(\ell)}} k(\cdot, t - T_i^{(\ell')}) \mathbf{1}_{0<t-T_i^{(\ell')}\le A} \right\|_{\mathcal{H}} \mathrm{d}t \ge 0,$$

and

$$D_{\ell\ell'} = \int_0^T \left\| \left(\sum_{i=1}^{N_T^{(\ell)}} k(\cdot, t - T_i^{(\ell)}) \mathbf{1}_{0<t-T_i^{(\ell)}\le A} \right) \left(\sum_{i=1}^{N_T^{(\ell')}} \mathbf{1}_{0<t-T_i^{(\ell')}\le A} \right) \right\|_{\mathcal{H}} \mathrm{d}t \ge 0.$$

The upper bound $J^+(\theta)$ consists in replacing in $J_0(\theta)$ the left-hand side of Equation (1) by its right-hand side. The objective function then becomes: for all $\theta \in \Theta^+$,

$$\begin{aligned}
F(\theta) &= J^+(\theta) + \frac{\eta}{2} \sum_{1\le j,\ell\le d} \|h_{j\ell}\|_{\mathcal{H}}^2 \\
&= \sum_{j=1}^d \Bigg[\mu_j^2 T + 2\mu_j \sum_{\ell=1}^d b_{j\ell} B^{(\ell)} + \int_0^T \left(\sum_{\ell=1}^d b_{j\ell} \sum_{i=1}^{N_T^{(\ell)}} \mathbf{1}_{0<t-T_i^{(\ell)}\le A} \right)^2 \mathrm{d}t \\
&+ 2\mu_j \sum_{\ell=1}^d \int_0^T \left\langle h_{j\ell}, \sum_{i=1}^{N_T^{(\ell)}} k(\cdot, t - T_i^{(\ell)}) \mathbf{1}_{0<t-T_i^{(\ell)}\le A} \right\rangle_{\mathcal{H}} \mathrm{d}t + \sum_{1\le \ell,\ell'\le d} \left(C_{\ell\ell'} \|h_{j\ell'}\|_{\mathcal{H}} + 2D_{\ell\ell'} |b_{j\ell'}| \right) \|h_{j\ell}\|_{\mathcal{H}} \\
&- 2 \sum_{n=1}^{N_T^{(j)}} \varphi \left(\mu_j + \sum_{\ell=1}^d \left(\left\langle h_{j\ell}, \sum_{i=1}^{N_T^{(\ell)}} k\left(\cdot, T_n^{(j)} - T_i^{(\ell)}\right) \mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A} \right\rangle_{\mathcal{H}} + b_{j\ell} \sum_{i=1}^{N_T^{(\ell)}} \mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A} \right) \right) \Bigg] \\
&+ \frac{\eta}{2} \sum_{1\le j,\ell\le d} \|h_{j\ell}\|_{\mathcal{H}}^2 \\
&= \sum_{j=1}^d \Bigg[\mu_j^2 T + 2\mu_j \sum_{\ell=1}^d b_{j\ell} B^{(\ell)} + \int_0^T \left(\sum_{\ell=1}^d b_{j\ell} \sum_{i=1}^{N_T^{(\ell)}} \mathbf{1}_{0<t-T_i^{(\ell)}\le A} \right)^2 \mathrm{d}t \\
&+ 2\mu_j \sum_{\ell=1}^d \langle h_{j\ell}, r_\ell \rangle_{\mathcal{H}} + \sum_{1\le \ell,\ell'\le d} \left(C_{\ell\ell'} \|h_{j\ell'}\|_{\mathcal{H}} + 2D_{\ell\ell'} |b_{j\ell'}| \right) \|h_{j\ell}\|_{\mathcal{H}} \\
&- 2 \sum_{n=1}^{N_T^{(j)}} \varphi \left(\mu_j + \sum_{\ell=1}^d \left(\langle h_{j\ell}, q_{nj\ell} \rangle_{\mathcal{H}} + b_{j\ell} \sum_{i=1}^{N_T^{(\ell)}} \mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\le A} \right) \right) \Bigg] + \frac{\eta}{2} \sum_{1\le j,\ell\le d} \|h_{j\ell}\|_{\mathcal{H}}^2.
\end{aligned}$$

Now, the rest of the proof is similar to that of Proposition 2: $h_{j\ell}$ appears through $\|h_{j\ell}\|_{\mathcal{H}}$ and the same linear terms as in the proof of Proposition 2, so the Pythagorean theorem still applies and the final form of $h_{j\ell}$ is the same.

D Implementation details

From Section 3.3, for all $\theta \in \Theta_{\|}^+$, for each $(j, \ell) \in [\![1, d]\!]^2$, there exist parameters $\left\{\alpha_u^{(j\ell)}\right\}_{0\le u\le N_T^{(j)}}$ such that:

$$h_{j\ell} = \sum_{u=0}^{N_T^{(j)}} \alpha_u^{(j\ell)} q_{uj\ell},$$

where $q_{0j\ell} = r_\ell$. Then

$$
\begin{aligned}
F_M(\theta) &= \sum_{j=1}^{d} \left[\frac{T}{M} \sum_{n=1}^{M} \varphi_1 \left(\mu_j + \sum_{\ell=1}^{d} \sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell} \left(\tau_n - T_i^{(\ell)} \right) + b_{j\ell} \right) \mathbf{1}_{0<\tau_n - T_i^{(\ell)} \leq A} \right) \right. \\
&\quad \left. - \sum_{n=1}^{N_T^{(j)}} \varphi_2 \left(\mu_j + \sum_{\ell=1}^{d} \sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell} \left(T_n^{(j)} - T_i^{(\ell)} \right) + b_{j\ell} \right) \mathbf{1}_{0<T_n^{(j)} - T_i^{(\ell)} \leq A} \right) \right] + \frac{\eta}{2} \sum_{1 \leq j, \ell \leq d} \|h_{j\ell}\|_{\mathcal{H}}^2 \\
&= \sum_{j=1}^{d} \left[\frac{T}{M} \sum_{n=1}^{M} \varphi_1 \left(\mu_j + \sum_{\ell=1}^{d} \sum_{u=0}^{N_T^{(j)}} \alpha_u^{(j\ell)} \left\langle q_{uj\ell}, \sum_{i=1}^{N_T^{(\ell)}} k \left(\cdot, \tau_n - T_i^{(\ell)} \right) \mathbf{1}_{0<\tau_n - T_i^{(\ell)} \leq A} \right\rangle_{\mathcal{H}} + \sum_{\ell=1}^{d} b_{j\ell} \sum_{i=1}^{N_T^{(\ell)}} \mathbf{1}_{0<\tau_n - T_i^{(\ell)} \leq A} \right) \right. \\
&\quad \left. - \sum_{n=1}^{N_T^{(j)}} \varphi_2 \left(\mu_j + \sum_{\ell=1}^{d} \sum_{u=0}^{N_T^{(j)}} \alpha_u^{(j\ell)} \left\langle q_{uj\ell}, q_{nj\ell} \right\rangle_{\mathcal{H}} + \sum_{\ell=1}^{d} b_{j\ell} \sum_{i=1}^{N_T^{(\ell)}} \mathbf{1}_{0<T_n^{(j)} - T_i^{(\ell)} \leq A} \right) \right] \\
&\quad + \frac{\eta}{2} \sum_{1 \leq j, \ell \leq d} \sum_{u=0}^{N_T^{(j)}} \sum_{n=0}^{N_T^{(j)}} \alpha_u^{(j\ell)} \alpha_n^{(j\ell)} \left\langle q_{uj\ell}, q_{nj\ell} \right\rangle_{\mathcal{H}} \\
&= \sum_{j=1}^{d} \left[\frac{T}{M} \mathbb{1}^\top \varphi_1 \left(\mu_j \mathbb{1} + Q^{(j)} \alpha^{(j)} + B b^{(j)} \right) - \mathbb{1}^\top \varphi_2 \left(\mu_j \mathbb{1} + K^{(j)} \alpha^{(j)} + E^{(j)} b^{(j)} \right) \right] \\
&\quad + \frac{\eta}{2} \sum_{1 \leq j, \ell \leq d} {\alpha^{(j\ell)}}^\top K^{(j\ell)} \alpha^{(j\ell)},
\end{aligned}
$$

where

$$
B = \left(\sum_{i=1}^{N_T^{(\ell)}} \mathbf{1}_{0<\tau_n - T_i^{(\ell)} \leq A} \right)_{\substack{1 \leq n \leq M \\ 1 \leq \ell \leq d}} \in \mathbb{R}^{M \times d},
$$

and for all $j \in [\![1, d]\!]$,

$$
E^{(j)} = \left(\sum_{i=1}^{N_T^{(\ell)}} \mathbf{1}_{0<T_n^{(j)} - T_i^{(\ell)} \leq A} \right)_{\substack{1 \leq n \leq N_T^{(j)} \\ 1 \leq \ell \leq d}} \in \mathbb{R}^{N_T^{(j)} \times d},
$$

and

$$
\alpha^{(j)} = \begin{bmatrix} \alpha^{(j1)} \\ \vdots \\ \alpha^{(jd)} \end{bmatrix} \in \mathbb{R}^{d(N_T^{(j)}+1)} \quad \text{and} \quad b^{(j)} = \begin{bmatrix} b_{j1} \\ \vdots \\ b_{jd} \end{bmatrix} \in \mathbb{R}^d,
$$

are concatenated vectors,

$$
Q^{(j)} = \left[Q^{(j1)} \mid \cdots \mid Q^{(jd)} \right] \in \mathbb{R}^{M \times d(N_T^{(j)}+1)} \quad \text{and} \quad K^{(j)} = \left[K_1^{(j1)} \mid \cdots \mid K_1^{(jd)} \right] \in \mathbb{R}^{N_T^{(j)} \times d(N_T^{(j)}+1)},
$$

are concatenated matrices with, for all $\ell \in [\![1, d]\!]$,

$$
Q^{(j\ell)} = \left(\left\langle q_{uj\ell}, \sum_{i=1}^{N_T^{(\ell)}} k \left(\cdot, \tau_n - T_i^{(\ell)} \right) \mathbf{1}_{0<\tau_n - T_i^{(\ell)} \leq A} \right\rangle_{\mathcal{H}} \right)_{\substack{1 \leq n \leq M \\ 0 \leq u \leq N_T^{(j)}}}, \quad K^{(j\ell)} = \left(\left\langle q_{uj\ell}, q_{nj\ell} \right\rangle_{\mathcal{H}} \right)_{\substack{0 \leq n \leq N_T^{(j)} \\ 0 \leq u \leq N_T^{(j)}}},
$$

and $K_1^{(j\ell)}$ is the submatrix composed of the last n rows of $K^{(j\ell)}$.

Now, for all $u \in [\![0, N_T^{(j)}]\!]$,

$$q_{uj\ell} = \left\{ \sum_{v=1}^{N_T^{(\ell)}} \int_0^T k\left(\cdot, t - T_v^{(\ell)}\right) \mathbf{1}_{0 < t - T_v^{(\ell)} \leq A} \,\mathrm{d}t \right\} \mathbf{1}_{u=0} + \left\{ \sum_{v=1}^{N_T^{(\ell)}} k\left(\cdot, T_u^{(j)} - T_v^{(\ell)}\right) \mathbf{1}_{0 < T_u^{(j)} - T_v^{(\ell)} \leq A} \right\} \mathbf{1}_{u \geq 1},$$

so for all $n \in [\![1, M]\!]$,

$$Q_{nu}^{(j\ell)} = \left\{ \int_0^T s_\ell\left(\tau_n, t\right) \,\mathrm{d}t \right\} \mathbf{1}_{u=0} + s_\ell\left(\tau_n, T_u^{(j)}\right) \mathbf{1}_{u \geq 1},$$

and for all $n \in [\![0, N_T^{(j)}]\!]$,

$$\begin{aligned} K_{nu}^{(j\ell)} = &\left\{ \int_0^T \int_0^T s_\ell\left(t, t'\right) \,\mathrm{d}t \,\mathrm{d}t' \right\} \mathbf{1}_{u=0} \mathbf{1}_{n=0} + \left\{ \int_0^T s_\ell\left(t, T_u^{(j)}\right) \,\mathrm{d}t \right\} \mathbf{1}_{u \geq 1} \mathbf{1}_{n=0} \\ &+ \left\{ \int_0^T s_\ell\left(T_n^{(j)}, t\right) \,\mathrm{d}t \right\} \mathbf{1}_{u=0} \mathbf{1}_{n \geq 1} + s_\ell\left(T_n^{(j)}, T_u^{(j)}\right) \mathbf{1}_{u \geq 1} \mathbf{1}_{n \geq 1}, \end{aligned}$$

where

$$s_\ell : (x, x') \in \mathbb{R}^2 \mapsto \sum_{1 \leq i, v \leq N_T^{(\ell)}} k\left(x - T_i^{(\ell)}, x' - T_v^{(\ell)}\right) \mathbf{1}_{0 < x - T_i^{(\ell)} \leq A} \mathbf{1}_{0 < x' - T_v^{(\ell)} \leq A}.$$

Remark D.1. *In the particular case of the Gaussian kernel, $k : (x, x') \in \mathbb{R}^2 \mapsto \mathrm{e}^{-\gamma(x - x')^2}$ (where $\gamma > 0$), for all $\ell \in [\![1, d]\!]^2$,*

$$\begin{aligned} \forall x \in \mathbb{R} : \quad \int_0^T s_\ell(x, t) \,\mathrm{d}t = \frac{\sqrt{\pi}}{2} \sum_{1 \leq i, v \leq N_T^{(\ell)}} &\Big[\mathrm{erf}_\gamma\left(\min\left(T - T_v^{(\ell)}, A\right) - \left(x - T_i^{(\ell)}\right)\right) \\ &+ \mathrm{erf}_\gamma\left(x - T_i^{(\ell)}\right)\Big] \mathbf{1}_{0 < x - T_i^{(\ell)} \leq A}, \end{aligned}$$

where $\mathrm{erf}_\gamma : x \mapsto \gamma^{-1/2} \,\mathrm{erf}(\gamma^{1/2} x)$ and erf is the Gauss error function. Moreover,

$$\begin{aligned} \int_0^T \int_0^T s_\ell\left(t, t'\right) \,\mathrm{d}t \,\mathrm{d}t' = \frac{\sqrt{\pi}}{2} \sum_{1 \leq i, v \leq N_T^{(\ell)}} &\Big[2G_\gamma\left(\min\left(T - T_v^{(\ell)}, A\right)\right) \\ &- G_\gamma\left(\min\left(T - T_v^{(\ell)}, A\right) - \min\left(T - T_i^{(\ell)}, A\right)\right)\Big], \end{aligned}$$

where $G_\gamma : x \in \mathbb{R} \mapsto x \,\mathrm{erf}_\gamma(x) + \frac{\gamma^{-1}}{\sqrt{\pi}}\left(\mathrm{e}^{-\gamma x^2} - 1\right)$ is an antiderivative of erf_γ.

If φ is differentiable, this also holds true for φ_1, φ_2 and F_M, and for all $j \in [\![1, d]\!]$ and all $\theta \in \Theta_{\parallel}^{+}$, gradients read:

$$\frac{\partial F_M}{\partial \mu_j}(\theta) = \frac{T}{M} \mathbb{1}^\top \varphi_1'\left(\mu_j \mathbb{1} + Q^{(j)} \alpha^{(j)} + B b^{(j)}\right) - \mathbb{1}^\top \varphi_2'\left(\mu_j \mathbb{1} + K^{(j)} \alpha^{(j)} + E^{(j)} b^{(j)}\right),$$

and

$$\nabla_{\alpha^{(j)}} F_M(\theta) = \frac{T}{M} Q^{(j)^\top} \varphi_1'\left(\mu_j \mathbb{1} + Q^{(j)} \alpha^{(j)} + B b^{(j)}\right) - K^{(j)^\top} \varphi_2'\left(\mu_j \mathbb{1} + K^{(j)} \alpha^{(j)} + E^{(j)} b^{(j)}\right) + \eta \begin{bmatrix} K^{(j1)} \alpha^{(j1)} \\ \vdots \\ K^{(jd)} \alpha^{(jd)} \end{bmatrix},$$

and

$$\nabla_{b^{(j)}} F_M(\theta) = \frac{T}{M} B^\top \varphi_1' \left(\mu_j \mathbb{1} + Q^{(j)} \alpha^{(j)} + B b^{(j)}\right) - E^{(j)\top} \varphi_2' \left(\mu_j \mathbb{1} + K^{(j)} \alpha^{(j)} + E^{(j)} b^{(j)}\right).$$

E Proofs of Propositions 4 and 5

Throughout this section, we assume that the kernel assotiated to the RKHS $\mathcal{H}$ is bounded:

$$\exists \kappa > 0 : \quad \forall x \in \mathbb{R}, k(x,x) \leq \kappa^2,$$

and L_k-Lipschitz continuous:

$$\forall x \in \mathbb{R} : \quad \forall (y, y') \in \mathbb{R}^2, |k(x,y) - k(x,y')| \leq L_k |y - y'|.$$

We also assume that φ is the ReLU function.

Lemma 6. *Let $\delta > 0$. Then,*

$$\forall x \in \mathbb{R}, \quad |\varphi(x) - \tilde{\varphi}(x)| \leq \frac{\log 2}{\omega},$$

and

$$\forall x > \delta, \quad |(\log \circ \varphi)(x) - (\log \circ \tilde{\varphi})(x)| \leq \frac{\log 2}{\delta \omega}.$$

Proof. Let $x \in \mathbb{R}$. If $x < 0$, then

$$|\varphi(x) - \tilde{\varphi}(x)| = \tilde{\varphi}(x) \leq \tilde{\varphi}(0) = \frac{\log 2}{\omega}.$$

If $x \geq 0$, then

$$|\varphi(x) - \tilde{\varphi}(x)| = \frac{\log(1 + \mathrm{e}^{-\omega x})}{\omega} \leq \frac{\log 2}{\omega}.$$

Let $\delta > 0$ and $x > \delta$. Then

$$|(\log \circ \varphi)(x) - (\log \circ \tilde{\varphi})(x)| = \log\left(\frac{\tilde{\varphi}(x)}{x}\right) = \log\left(1 + \frac{\log(1 + \mathrm{e}^{-\omega x})}{\omega x}\right) \leq \frac{\log(1 + \mathrm{e}^{-\omega x})}{\omega x} \leq \frac{\log 2}{\delta \omega}.$$

□

Lemma 7. *Let $\theta \in \Omega$. Then, for all $j \in [\![1, d]\!]$ and for every interval I without a discountinuity of $\lambda^{(j)}$, $\lambda^{(j)}$ is L_λ-Lipschitz continuous, where $L_\lambda = L_k N_T C$.*

Proof. Let $\theta \in \Theta^+$ and $j \in [\![1, d]\!]$, and I an interval without a jump time. For all $(t, t') \in I^2$,

$$
\begin{aligned}
|\lambda^{(j)}(t) - \lambda^{(j)}(t')| &= \left|\varphi\left(\mu_j + \sum_{\ell=1}^{d}\sum_{i=1}^{N_t^{(\ell)}} g_{j\ell}\left(t - T_i^{(\ell)}\right)\right) - \varphi\left(\mu_j + \sum_{\ell=1}^{d}\sum_{i=1}^{N_t^{(\ell)}} g_{j\ell}\left(t' - T_i^{(\ell)}\right)\right)\right| \\
&\leq \left|\sum_{\ell=1}^{d}\sum_{i=1}^{N_t^{(\ell)}} g_{j\ell}\left(t - T_i^{(\ell)}\right) - \sum_{\ell=1}^{d}\sum_{i=1}^{N_t^{(\ell)}} g_{j\ell}\left(t' - T_i^{(\ell)}\right)\right| \\
&\leq \sum_{\ell=1}^{d}\sum_{i=1}^{N_t^{(\ell)}} \left|\left(h_{j\ell}\left(t - T_i^{(\ell)}\right) + b_{j\ell}\right)\mathbf{1}_{0<t-T_i^{(\ell)}\leq A} - \left(h_{j\ell}\left(t' - T_i^{(\ell)}\right) + b_{j\ell}\right)\mathbf{1}_{0<t'-T_i^{(\ell)}\leq A}\right| \\
&= \sum_{\ell=1}^{d}\sum_{i=1}^{N_t^{(\ell)}} \left|h_{j\ell}\left(t - T_i^{(\ell)}\right) - h_{j\ell}\left(t' - T_i^{(\ell)}\right)\right|\mathbf{1}_{0<t-T_i^{(\ell)}\leq A} \\
&\leq \sum_{\ell=1}^{d}\sum_{i=1}^{N_t^{(\ell)}} \|h_{j\ell}\|_{\mathcal{H}} \left\|k\left(\cdot, t - T_i^{(\ell)}\right) - k\left(\cdot, t' - T_i^{(\ell)}\right)\right\|_{\mathcal{H}} \mathbf{1}_{0<t-T_i^{(\ell)}\leq A} \\
&\leq \sum_{\ell=1}^{d}\sum_{i=1}^{N_t^{(\ell)}} L_k \|h_{j\ell}\|_{\mathcal{H}} |t - t'| \mathbf{1}_{0<t-T_i^{(\ell)}\leq A} \\
&\leq L_k \sum_{\ell=1}^{d} N_T^{(\ell)} \|h_{j\ell}\|_{\mathcal{H}} |t - t'|.
\end{aligned}
$$

Thus, $\lambda^{(j)}$ is $\left(L_k \sum_{\ell=1}^{d} N_T^{(\ell)} \|h_{j\ell}\|_{\mathcal{H}}\right)$-Lipschitz continuous. Now, for $\theta \in \Omega$,

$$
L_k \sum_{\ell=1}^{d} N_T^{(\ell)} \|h_{j\ell}\|_{\mathcal{H}} \leq L_k N_T C = L_\lambda.
$$

□

Lemma 8. *Let $\theta \in \Omega$, $j \in [\![1, d]\!]$ and I an interval (of length $|I|$) with S_I discountinuities of $\lambda^{(j)}$. Then, for any $\tau \in I$,*

$$
\int_I \left|\lambda^{(j)}(t) - \lambda^{(j)}(\tau)\right| \mathrm{d}t \leq \frac{L_\lambda |I|^2}{2} + |I| S_I G,
$$

where $L_\lambda = L_k N_T C$ and $G = \kappa C + B$.

Proof. Let $\theta \in \Omega$. Since discountinuities of $\lambda^{(j)}$ come from discountinuities of the interactions functions $g_{j\ell}$ ($\ell \in [\![1, d]\!]$) at the boundaries of their support $[0, A]$ (the functions $h_{j\ell}$, $\ell \in [\![1, d]\!]$, are continuous by Lipschitz-continuity of the kernel k), their amplitudes are bounded by

$$
\max_{1\leq \ell\leq d} \max(|g_{j\ell}(0), |g_{j\ell}(A)|) \leq \max_{1\leq \ell\leq d} \|g_{j\ell}\|_\infty.
$$

But for all $\ell \in [\![1, d]\!]$ and $x \in \mathbb{R}$,

$$
|g_{j\ell}(x)| \leq |\langle h_{j\ell}, k(\cdot, x)\rangle_{\mathcal{H}}| + |b_{j\ell}| \leq \|h_{j\ell}\|_{\mathcal{H}} \sqrt{k(x, x)} + |b_{j\ell}| \leq \kappa C + B = G.
$$

So discountinuities have jumps bounded by G.

Let $\tau \in I$. Then, for all $t \in I$, by Lemma 7:

$$
\left|\lambda^{(j)}(t) - \lambda^{(j)}(\tau)\right| \leq L_\lambda |t - \tau| + S_I G.
$$

By integration, it comes

$$\int_I \left|\lambda^{(j)}(t) - \lambda^{(j)}(\tau)\right| \,\mathrm{d}t \leq L_\lambda \int_0^{|I|} t \,\mathrm{d}t + |I| S_I G = \frac{L_\lambda |I|^2}{2} + |I| S_I G.$$

□

Proof of Proposition 4

Denoting, for all $j \in [\![1, d]\!]$, $\tilde{\lambda}^{(j)}(t) = \tilde{\varphi}\left(\mu_j + \sum_{\ell=1}^{d} \sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(t - T_i^{(\ell)}\right) + b_{j\ell}\right) \mathbf{1}_{0<t-T_i^{(\ell)}\leq A}\right)$, we have:

$$\begin{aligned}
0 &\leq L(\hat{\theta}) - L(\bar{\theta}) \\
&= L(\hat{\theta}) - L_{M,\omega}(\hat{\theta}) + L_{M,\omega}(\hat{\theta}) - L_{M,\omega}(\bar{\theta}) + L_{M,\omega}(\bar{\theta}) - L(\bar{\theta}) \\
&\leq 2 \max_{\theta\in\{\hat{\theta},\bar{\theta}\}} |L(\theta) - L_{M,\omega}(\theta)| \\
&\leq 2 \max_{\theta\in\{\hat{\theta},\bar{\theta}\}} \sum_{j=1}^{d} \left| \left[\int_0^T \lambda^{(j)}(t)\,\mathrm{d}t - \frac{T}{M}\sum_{n=1}^{M} \tilde{\lambda}^{(j)}(\tau_n)\right] - \left[\sum_{n=1}^{N_T^{(j)}} \log\left(\lambda^{(j)}\left(T_n^{(j)}\right)\right) - \sum_{n=1}^{N_T^{(j)}} \log\left(\tilde{\lambda}^{(j)}\left(T_n^{(j)}\right)\right)\right]\right|.
\end{aligned}$$

Let θ be either $\hat{\theta}$ or $\bar{\theta}$. Regarding the first term,

$$\begin{aligned}
\left|\int_0^T \lambda^{(j)}(t)\,\mathrm{d}t - \frac{T}{M}\sum_{n=1}^{M}\tilde{\lambda}^{(j)}(\tau_n)\right| &= \left|\sum_{n=1}^{M}\int_{\tau_n}^{\tau_n+\frac{T}{M}} \left(\lambda^{(j)}(t) - \tilde{\lambda}^{(j)}(\tau_n)\right)\,\mathrm{d}t\right| \\
&= \left|\sum_{n=1}^{M}\left(\int_{\tau_n}^{\tau_n+\frac{T}{M}} \left(\lambda^{(j)}(t) - \lambda^{(j)}(\tau_n)\right)\,\mathrm{d}t + \frac{T}{M}\left(\lambda^{(j)}(\tau_n) - \tilde{\lambda}^{(j)}(\tau_n)\right)\right)\right| \\
&\leq \sum_{n=1}^{M}\left|\int_{\tau_n}^{\tau_n+\frac{T}{M}} \left(\lambda^{(j)}(t) - \lambda^{(j)}(\tau_n)\right)\,\mathrm{d}t\right| + \frac{T\log 2}{\omega},
\end{aligned}$$

by Lemma 6. Now, by Lemma 8, denoting $L_\lambda = L_k N_T C$ and $G = \kappa C + B$:

$$\begin{aligned}
\left|\int_0^T \lambda^{(j)}(t)\,\mathrm{d}t - \frac{T}{M}\sum_{n=1}^{M}\tilde{\lambda}^{(j)}(\tau_n)\right| &\leq \sum_{n=1}^{M}\left(\frac{L_\lambda T^2}{2M^2} + \frac{T}{M} S_{[\tau_n,\tau_n+\frac{T}{M})} G\right) + \frac{T\log 2}{\omega} \\
&= \frac{L_\lambda T^2}{2M} + \frac{T}{M} S_{[0,T)} G + \frac{T\log 2}{\omega} \\
&\leq \frac{L_\lambda T^2}{2M} + \frac{2N_T G T}{M} + \frac{T\log 2}{\omega},
\end{aligned}$$

since $\lambda^{(j)}$ has at most $2N_T$ discountinuities on $[0, T]$ (since the functions $h_{j\ell}$, $\ell \in [\![1, d]\!]$, are continuous –by Lipschitz-continuity of the kernel k–, discountinuities of $\lambda^{(j)}$ come from discountinuities of the interactions functions $g_{j\ell}$, $\ell \in [\![1, d]\!]$, at the boundaries of their support $[0, A]$).

In addition, regarding the second term, since $L(\theta) < \infty$ (by assumption $L(\hat{\theta}) < \infty$ and $L(\bar{\theta}) < \infty$, since F is proper), there exists $\delta > 0$ such that

$$\min_{\theta\in\{\hat{\theta},\bar{\theta}\},\, j\in[\![1,d]\!],\, n\in[\![1,N_T^{(j)}]\!]} \left\{\mu_j + \sum_{\ell=1}^{d}\sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(T_n^{(j)} - T_i^{(\ell)}\right) + b_{j\ell}\right)\mathbf{1}_{0<t-T_i^{(\ell)}\leq A}\right\} > \delta.$$

It results that:

$$
\begin{aligned}
&\left|\sum_{n=1}^{N_T^{(j)}} \log\left(\lambda^{(j)}\left(T_n^{(j)}\right)\right) - \sum_{n=1}^{N_T^{(j)}} \log\left(\tilde{\lambda}^{(j)}\left(T_n^{(j)}\right)\right)\right| \\
&\leq \sum_{n=1}^{N_T^{(j)}} \left|(\log\circ\varphi)\left(\mu_j + \sum_{\ell=1}^{d}\sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(T_n^{(j)} - T_i^{(\ell)}\right) + b_{j\ell}\right)\mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\leq A}\right)\right. \\
&\left. - (\log\circ\tilde{\varphi})\left(\mu_j + \sum_{\ell=1}^{d}\sum_{i=1}^{N_T^{(\ell)}} \left(h_{j\ell}\left(T_n^{(j)} - T_i^{(\ell)}\right) + b_{j\ell}\right)\mathbf{1}_{0<T_n^{(j)}-T_i^{(\ell)}\leq A}\right)\right| \\
&\leq N_T^{(j)}\frac{\log 2}{\delta\omega},
\end{aligned}
$$

by Lemma 6.

Combining both bounds:

$$
\begin{aligned}
L(\hat{\theta}) - L(\bar{\theta}) &\leq 2\sum_{j=1}^{d}\left\{\frac{L_\lambda T^2}{2M} + \frac{2N_T G T}{M} + \frac{T\log 2}{\omega} + N_T^{(j)}\frac{\log 2}{\delta\omega}\right\} \\
&= \frac{T}{M}\left(dL_\lambda T + 4dN_T G\right) + \frac{2\log 2}{\omega}\left(dT + \frac{N_T}{\delta}\right) \\
&= \frac{T}{M}\left(L_k C d T N_T + 4(\kappa C + B)dN_T\right) + \frac{2\log 2}{\omega}\left(dT + \frac{N_T}{\delta}\right).
\end{aligned}
$$

Proof of Proposition 5

Let $j \in [\![1, d]\!]$ and θ be either $\hat{\theta}$ or $\bar{\theta}$. Then, for all $\omega \geq 1$ and $t \in [0, T]$,

$$
\begin{aligned}
0 \leq \lambda^{(j)}(t) \leq \tilde{\lambda}^{(j)}(t) \leq \lambda^{(j)}(t) + \frac{\log 2}{\omega} &= \varphi\left(\mu_j + \sum_{\ell=1}^{d}\sum_{i=1}^{N_T^{(\ell)}} g_{j\ell}\left(t - T_i^{(\ell)}\right)\right) + \frac{\log 2}{\omega} \\
&\leq \mu_j + \sum_{\ell=1}^{d}\sum_{i=1}^{N_T^{(\ell)}} \|g_{j\ell}\|_\infty + \frac{\log 2}{\omega} \\
&\leq B + N_T G + \frac{\log 2}{\omega} \\
&\leq \frac{H}{2},
\end{aligned}
$$

with $G = \kappa C + B$ and $H = 2(B + N_T G + \log 2)$. Thus, both $\lambda^{(j)}$ and $\tilde{\lambda}^{(j)}$ have values in $[0, H/2]$, interval on which the square function is H-Lipschitz continuous.

Now, following the proof of Proposition 4, we have:

$$
\begin{aligned}
0 &\leq J(\hat{\theta}) - J(\bar{\theta}) \\
&= J(\hat{\theta}) - J_{M,\omega}(\hat{\theta}) + J_{M,\omega}(\hat{\theta}) - J_{M,\omega}(\bar{\theta}) + J_{M,\omega}(\bar{\theta}) - J(\bar{\theta}) \\
&\leq 2\max_{\theta\in\{\hat{\theta},\bar{\theta}\}} |J(\theta) - J_{M,\omega}(\theta)| \\
&\leq 2\max_{\theta\in\{\hat{\theta},\bar{\theta}\}} \sum_{j=1}^{d}\left|\left[\int_0^T \lambda^{(j)}(t)^2\,\mathrm{d}t - \frac{T}{M}\sum_{n=1}^{M}\tilde{\lambda}^{(j)}(\tau_n)^2\right] - 2\left[\sum_{n=1}^{N_T^{(j)}}\lambda^{(j)}\left(T_n^{(j)}\right) - \sum_{n=1}^{N_T^{(j)}}\tilde{\lambda}^{(j)}\left(T_n^{(j)}\right)\right]\right|.
\end{aligned}
$$

Let θ be either $\hat{\theta}$ or $\bar{\theta}$. Regarding the first term,

$$\begin{aligned}\left|\int_0^T \lambda^{(j)}(t)^2 \,\mathrm{d}t - \frac{T}{M}\sum_{n=1}^{M} \tilde{\lambda}^{(j)}(\tau_n)^2\right| &= \left|\sum_{n=1}^{M}\int_{\tau_n}^{\tau_n+\frac{T}{M}} \left(\lambda^{(j)}(t)^2 - \tilde{\lambda}^{(j)}(\tau_n)^2\right) \mathrm{d}t\right| \\ &\leq H\sum_{n=1}^{M}\int_{\tau_n}^{\tau_n+\frac{T}{M}} \left|\lambda^{(j)}(t) - \tilde{\lambda}^{(j)}(\tau_n)\right| \mathrm{d}t \\ &\leq H\left(\frac{L_\lambda T^2}{2M} + \frac{2N_T G T}{M} + \frac{T\log 2}{\omega}\right),\end{aligned}$$

where $L_\lambda = L_k N_T C$, by a derivation similar to previously.

In addition, regarding the second term, we immediately have, by Lemma 6, that:

$$\left|\sum_{n=1}^{N_T^{(j)}} \lambda^{(j)}\left(T_n^{(j)}\right) - \sum_{n=1}^{N_T^{(j)}} \tilde{\lambda}^{(j)}\left(T_n^{(j)}\right)\right| \leq N_T^{(j)}\frac{\log 2}{\omega}.$$

Combining both bounds:

$$\begin{aligned}J(\hat{\theta}) - J(\bar{\theta}) &\leq \frac{HT}{M}\left(LT + 4N_T dG\right) + \frac{2\log 2}{\omega}\left(HdT + N_T\right). \\ &= \frac{HT}{M}\left(L_k C dT N_T + 4(\kappa C + B) dN_T\right) + \frac{4\log 2}{\omega}\left(HdT + N_T\right).\end{aligned}$$

F Additional numerical results

F.1 Synthetic data

Figure 7 depicts the learning time (average single training time) on a personal computer of the several methods compared on the synthetic data. The methods **Gaussian** and **Bernstein** are very fast to train because they have only $d(1+dU)$ parameters (with $U = 10$ here), while **RKHS** has $d(1+N_T+2d)$ parameters. That being said, it is interesting to observe that **RKHS** is only two times slower than **Exponential**, which is a model with $d(2+d)$ parameters but that necessitates a specific treatment to compute exactly the compensator and that involves a non-convex objective function. Let us remark that, on the considered dataset, the most accurate approach is the proposed **RKHS**, the training time being the price to pay for the accuracy.

F.2 Neuronal data

Figure 8 presents all estimated interaction functions for the several methods considered. The proposed **RKHS**, as well as **Bernstein** [Lemonnier and Vayatis, 2014], recover complex auto-interactions, including the refractory period. In return, cross-interactions are estimated close to 0, except for minor inhibiting effects by kernels $1 \leftarrow 3$, $2 \leftarrow 1$, $5 \leftarrow 1$, and exciting effects by kernels $2 \leftarrow 4$, $3 \leftarrow 4$, $4 \leftarrow 1$, $4 \leftarrow 3$.

As a quantitative assessment, Table 1 gives the log-likelihood scores computed on a test trajectory obtained by a randomized concatenation of half of the recordings. In agreement with Figure 8, **RKHS** and **Bernstein** seem to fit the underlying process better than **Exponential** and **Gaussian**. Moreover, the highest score is obtained by **RKHS**, leading to believe that it is more suited than competitors.

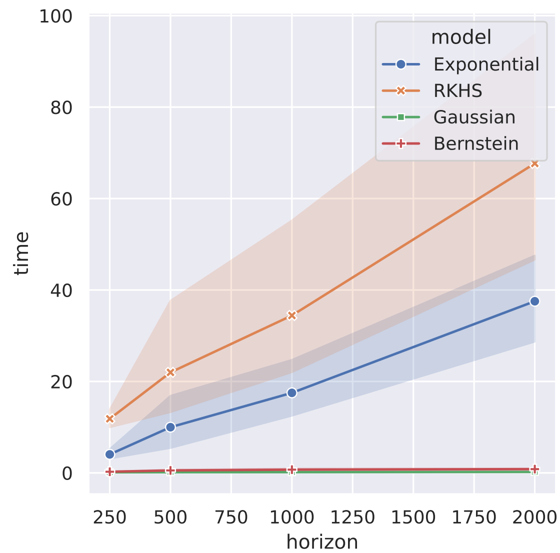

Figure 7: Learning time (in seconds) with respect to the horizon T.

Table 1: Log-likelihood scores on a test trajectory (the higher, the better).

Model	Log-Likelihood
Exponential	2152
RKHS	2485
Gaussian	2178
Bernstein	2334

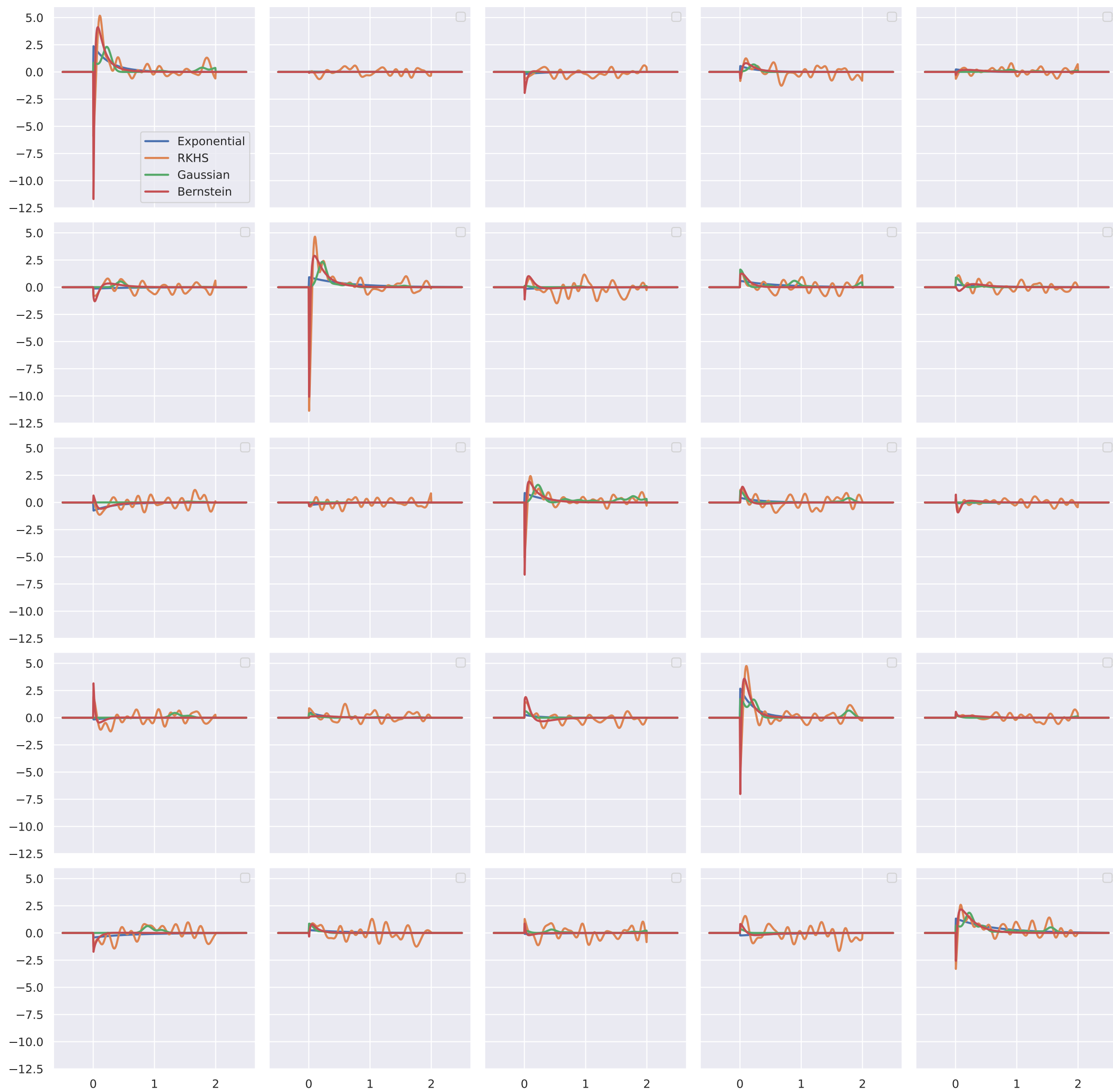

Figure 8: Interaction functions learned on the neuronal subnetwork.