Flexible Temporal Point Processes Modeling with Nonlinear Hawkes Processes with Gaussian Processes Excitations and Inhibitions

Anonymous Authors[1]

Abstract

We propose an extended Hawkes process model where the self–effects are of both excitatory and inhibitory type and follow a Gaussian Process. Whereas previous work either relies on a less flexible parameterization of the model, or requires a large amount of data, our formulation allows for both a flexible model and learning when data are scarce. Efficient approximate Bayesian inference is achieved via data augmentation, and we describe a mean–field variational inference approach to learn the model parameters. To demonstrate the flexibility of the model we apply our methodology on data from two different domains and compare it to previously reported results.

1. Introduction

Sequences of self exciting, or inhibiting, temporal events are frequent footmarks of natural phenomena: Earthquakes are known to be temporally clustered as aftershocks are commonly triggered following the occurrence of a main event (Ogata, 1988); in social networks, the propagation of news can be modeled in terms of information cascades over the edges of a graph (Zhao et al., 2015); and in neuronal activity, the occurrence of one spike may increase or decrease the probability of the occurrence of the next spike over some time period (Dayan & Abbott, 2001).

A common model for time series of events with history dependence is the Hawkes process. Originally, the dependence on the history in the Hawkes process is assumed to be self excitatory Assuming only excitatory relation between the events, does not hold for some of the phenomena we wish to model. For example, inhibitory effects between neurons (Maffei et al., 2004), and even self–inhibition (Smith & Jahr, 2002), are crucial for regulating the neuronal activity. Thus, the memory kernel should also include inhibitory relations between the events and by doing so the intensity may become negative. To ensure that the intensity function is non–negative, a nonlinear link function is applied on the memory kernel, and the resulting model is often referred to as a Nonlinear Hawkes process (Brémaud & Massoulié, 1996; Zhu, 2013; Truccolo, 2016).

In this work we present a *Nonlinear Hawkes process with Gaussian Process Self–effects* (NH-GPS) which extends the class of Nonlinear Hawkes processes. We choose a non–parametric approach which avoids the limiting parameterization of the memory kernel and the background rate. We assume a Gaussian Process (GP) prior on the exogenous events intensity and on the memory kernel, which allows also for an inhibitory effect between the events, and use the Sigmoid link function over the linear intensity. This modeling approach is not only descriptive, but also allows us to obtain a fast inference procedure. The history of self–effects defines an aggregated Gaussian process, and we perform the inference directly on this aggregation rather than obtaining a posterior over each self effect.

2. Related Work

A highly flexible approach to estimating the intensity function of the linear Hawkes process relies on GP priors (Zhang et al., 2018; Zhou et al., 2019; Zhang et al., 2020; Zhou et al., 2020). Differently to our work, Zhou et al. (2020) remain in the linear Hawkes process regime and assume that the effects of past events are assumed to be only excitatory, whereas our approach allows both excitatory and inhibitory effects.

A recent variation of the nonlinear Hawkes process is the Mutually Regressive Point Process (Apostolopoulou et al., 2019), which was designed to model neuronal spike trains. In this work, the classical self–excitatory Hawkes Process intensity function is augmented by a probability term. This term induces inhibition when it is close to zero. In a sense, this model includes two memory kernels – one excitatory only which appears in the intensity function and another which can also induce inhibition in the augmenting probability term. In the current work, we achieve such flexibility of the self–effects in a simpler fashion by assuming the GP

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

prior on the self effects. As mentioned before this also allows for the type of effect to change over time, which does not appear in the work of Apostolopoulou et al. (2019).

3. Proposed Model

3.1. Classical Hawkes Process

Let $\mathcal{T}_T = [0, t] \in \mathbb{R}$. We define the counting measure $N(\mathcal{T}_t)$ as the number of arrivals in the sequence $\mathcal{H}_t = \{T_1, ..., T_{N(\mathcal{T}_t)} : T_i \in \mathcal{T}_t \wedge T_{i-1} < T_i\}$ where $\mathcal{H}_t$ defines the history of the process until time t, and T_i corresponds to the time of arrival i. For a temporal point process, the counting measure $N(\cdot)$ has an associated intensity defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}[N(\mathcal{T}_{t+\Delta t}) - N(\mathcal{T}_t) | \mathcal{H}_t]}{\Delta t}.$$

The intensity function may depend on the history of the process. An example of such a process is the Hawkes process, or self exciting point process, (Kingman, 1993) which defines self excitations (Daley & Vere-Jones, 2007) around *exogenous events*.

Following Hawkes & Oakes (1974), the intensity of the Hawkes process is defined by

$$\lambda(t|\mathcal{H}_t) = s\left(t\right) + \sum_{t_n < t} g\left(t - t_n\right), \tag{1}$$

where $s(t)$ is the base intensity of exogenous arrivals and $g\left(t - t_n\right)$ is the memory kernel defining the change in the excitation value for each arrival. In the classical Hawkes process, only excitations are allowed and the memory kernel is usually of the form $g(t - t_n) = \beta e^{-\beta(t - t_n)}$ for an exponentially decaying memory.

3.2. Nonlinear Hawkes process with Gaussian Process Self–Effects

In the classical Hawkes process, the memory kernel g in Equation 1 must be non–negative, to prevent the intensity function from being negative. As a result the history of the model has only excitatory effect on future events. We are interested in a model that includes inhibition between events, and we release the constraint over g so it can be negative, and define the following nonlinear intensity function

$$\lambda\left(t\right) = \lambda^* \sigma\left(\phi(t)\right) \tag{2}$$

$$\sigma\left(\phi(t)\right) = \frac{1}{1 + \exp\left(-\phi\left(t\right)\right)} \tag{3}$$

$$\phi(t) = s\left(t\right) + \sum_{t_n < t} g\left(t - t_n\right) \exp\left(-\beta\left(t - t_n\right)\right). \tag{4}$$

Here, we choose the sigmoid function to ensure that the intensity function $\lambda\left(\cdot\right)$ is non–negative. λ^* is the intensity bound and we refer to $\phi\left(\cdot\right)$ as the linear intensity function. We explicitly add the exponential decay to enforce the forgetting constraint which is essential for most realistic processes and β determines the forgetting rate. Although we choose here a specific parameterization of the memory decay, one can choose other forms of memory decay with minimal adaptation to the learning procedure of the model parameters.

To maximize the flexibility of the model we avoid any specific parameterization of the background rate or the memory kernel. Thus, rather than specifying a functional form for $s\left(t\right)$ and $g\left(t\right)$, we assume the following GP priors

$$s \sim GP\left(0, K^s\right) \tag{5}$$

$$g \sim GP\left(0, K^g\right) \tag{6}$$

$$K_{\text{RBF}}\left(t_1, t_2\right) = a \cdot \exp\left(-\frac{\|t_1 - t_2\|^2}{\sigma^2}\right). \tag{7}$$

In this work we use the Radial Basis Function (RBF) kernel for the GP priors. This choice is not a constraint of the model – one can choose any other kernel, and it will not effect the augmentation and inference processes described bellow.

Finally, we assume a prior distribution also on the upper intensity bound

$$\lambda^* \sim Gamma\left(\beta_0, \beta_0\right).$$

and we identify the hyperparameters of the model as $\{\sigma_g, a_g, \beta, \sigma_s, a_s\}$.

In this work we propose Bayesian inference for fitting the model to data. Due to the non–linearity over $\phi\left(\cdot\right)$ we are no longer able to easily utilize the branching structure of the Hawkes process which allowed for the estimation of $s\left(\cdot\right)$ and $g\left(\cdot\right)$ (Rasmussen, 2013; Zhou et al., 2020). Thus, a natural solution is to perform the inference directly on $\phi\left(\cdot\right)$.

Next, we identify the prior over the entire linear intensity $p\left(\phi\right)$. From Equation 4 we see that the linear intensity function ϕ is nothing but the sum of GPs, and as such it is also a GP

$$\phi \sim GP\left(0, \tilde{K}\right)$$

$$\tilde{K}_{lk} =$$

$$K^s_{lk} + \sum_{t_i < t_l} \sum_{t_j < t_k} K^g_{t_l - t_i, t_k - t_j} \exp\left(-\beta\left(t_l - t_i + t_k - t_j\right)\right).$$

4. Inference

Conditioned on the intensity function $\lambda\left(\cdot\right)$, the likelihood of observations $\{t_1, ...t_n\}$ from a Hawkes process is (Daley & Vere-Jones, 2003)

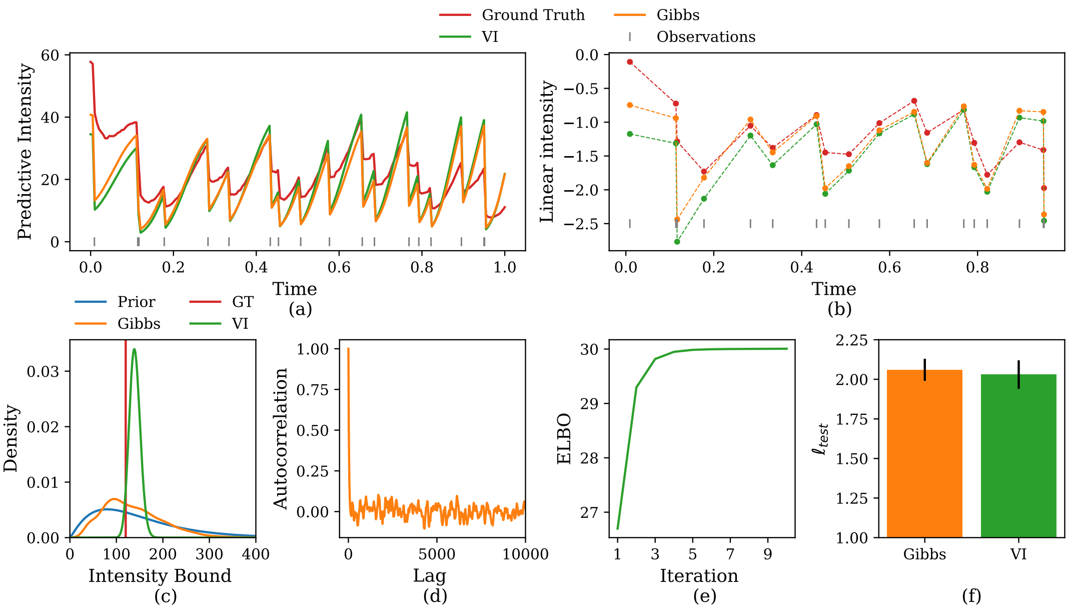

Figure 1. (a) Comparison of the ground truth predictive intensity and the one sampled from the VI and Gibbs inference. (b) Comparison of the ground truth linear intensity $\phi(\cdot)$ and the ones learned by the VI and Gibbs sampler. (c) Comparison of the Ground truth intensity bound and the one learned by the inference, and the prior distribution. (d) The autocorrelation of the intensity bound Gibbs samples. (e) The variational lower bound as a function of the algorithm iteration. (f) Comparison of the test log–likelihood of the Gibbs sampler and the VI.

$$\ell\left(\{t_1, ...t_n\}|\lambda(\cdot)\right) = \exp\left\{-\int_0^T \lambda(t')\,dt'\right\}\prod_{i=1}^{N}\lambda(t_i). \tag{8}$$

Looking at the likelihood defined above, Equations 4 and 8, implementing Bayesian inference for the model is not straightforward, due to the non-conjugate structure of the likelihood and prior. Similarly to previous work on Cox and Hawkes processes (Donner & Opper, 2018a; Apostolopoulou et al., 2019; Zhou et al., 2020), we augment the model with auxiliary variables, which leads to a conditionally conjugated model with closed form solutions for Gibbs sampler and variational inference. The details of the augmentation can be found in the supplementary material. The augmented posterior takes the following form

$$p\left(\{\hat{t}_m, \hat{w}_m\}, \{w_n\}, \phi, \lambda^*|\{t_n\}\right) \propto \exp\left(-\lambda^* T\right) \times \tag{9}$$
$$\prod_{m=1}^{M} \lambda^* e^{f\left(\hat{w}_m, -\phi\left(\hat{t}_m\right)\right)} PG\left(w_m; 1, 0\right) \times$$
$$\prod_{n=1}^{N} \lambda^* e^{f\left(w_n, \phi\left(t_n\right)\right)} PG\left(w_n; 1, 0\right) \times p\left(\phi\right) p\left(\lambda^*\right).$$

To summarize, we augment the model with two sets of variables – the Pólya-Gamma Polson et al. (2013) variables $\{w_n\}$ which augment the actual realizations and the tuples $\{\hat{t}_m, \hat{w}_m\}$ which are the realizations and marks of the auxiliary marked Poisson process.

As mentioned above, we intend to learn directly the linear intensity function $\phi(\cdot)$. This allows us to utilize the efficient mean–field variational inference previously introduced in Donner & Opper (2018a) and Donner & Opper (2018b). As a baseline for evaluating the mean–field variational inference algorithm we use a Gibbs sampler. Details of both the variational inference and the Gibbs sampler can be found in the supplementary material.

5. Experiments

5.1. Synthetic Data

To assess the performance of the inference algorithms presented in Section 4, we learn the parameters of data generated by the model, and compare the learned parameters to the ground truth. To generate data we start by sampling the memory GP and the background GP, based on Equations 6 and 7 and generate events using Poisson thinning (Lewis & Shedler, 1979).

The results for the synthetic data are included in Figure 1.

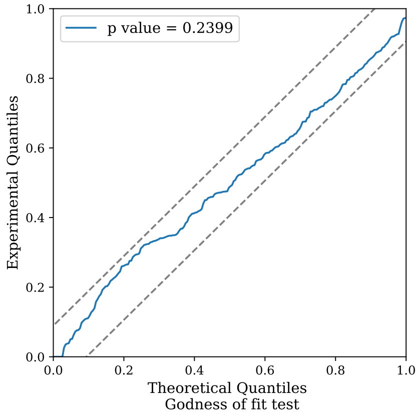

Figure 2. Results of the Kolmogorov-Smirnov test for the monkey cortex dataset. The model passes the goodness of fit test.

The time window used was one second, and the dataset includes 18 events. We use the test log–likelihood per data point, averaged over ten datasets, to quantify the performance of the two inference algorithms. The Gibbs sampler and the VI achieve very similar results. Thus, in the next section we present the results only from the VI.

5.2. Real Data

5.2.1. Crime Report Data

Our model assumes both inhibitory and excitatory self effects, but it should also be able to capture phenomena where only one of the two types of effects exist. To test this, we fit our model to crime report data, where it is assumed that past events have excitatory effect on future events (Mohler et al., 2011). We use the same two datasets described in Zhou et al. (2020), and follow their data processing procedure. Each dataset contains one type of crime and so we use the univariate version of the model. The work of Zhou et al. (2020) includes several inference methods and we compare our results to the results of their reported mean–field variational inference approach, as it is the closest to our inference procedure.

Table 1 compares the log of the mean of the test likelihood of our NH-GPS model to the one reported in Zhou et al. (2020). For details regarding the computation of the test likelihood can be found in the supplementary material. We perform the experiment five times and report the mean and variance, our model performs similarly to the non–parametric Hawkes process presented by Zhou et al. (2020).

5.2.2. Neuronal Activity Data

One of the motivating real world phenomena behind our work is the spiking activity of neurons, where it is known that the process has both self–excitatory and self–inhibitory effects. As an example for our model's ability to capture neuronal activity we use the datasets that were first presented in Gerhard et al. (2017) (Figure 2.c and 2.b there). These data were further analyzed in Apostolopoulou et al. (2019) (Figure 5 there) where the Mutually Regressive Point Process (MR-PP) is introduced. One dataset includes ten recordings from a single neuron in a monkey cortex, with the duration of one second each, and the other includes ten recordings from single neuron in a human cortex for a duration of ten seconds each.

Table 1. Crime Report Data Test Log–Likelihood.

Dataset	**Zhou et al. (2020)**	**NH–GPS**
Vancouver	453.11 ± 8.94	453.8 ± 12.2
NYPD	-200.7 ± 3.32	-202.8 ± 7.54

To quantify how suitable the model is to the data, we apply the random time change theorem (Daley & Vere-Jones, 2003) to the inferred intensity and the experimental data. The theorem states that realizations from a general point process can be transformed to realizations from a homogeneous Poisson process with unit rate. Similarly to the work of Apostolopoulou et al. (2019), we further transform the exponential realizations to those from a uniform distribution, following Brown et al. (2002). We then use the Kolmogorov-Smirnov test to compare the quantiles of the distribution of the transformed realizations to the quantiles of the uniform distribution. The model passes the goodness of fit test (p value > 0.05). The results of the KS test are shown in Figure 2. Further results can be found in the supplementary material.

6. Conclusion

In this work we presented the nonlinear Hawkes model with Gaussian process self–effects (NH-GPS). We motivated the development of the new model with the need for a flexible model that can capture both exciting and inhibiting interactions between events, while maintaining the ability to learn also when data are scarce.

Due to the structure of the model, we dispense with the branching structure that is commonly used for Bayesian inference in Hawkes processes. We propose an efficient mean–field variational inference algorithm which relies on a data augmenting scheme. We show that the results of the variational inference are comparable with those of a Gibbs sampler.

We demonstrate the performance of our model in two different real world applications. Due to the flexibility of our model, it achieves good results on data where events have only excitatory effects and on data where events have both excitatory and inhibitory effects.

References

Apostolopoulou, I., Linderman, S., Miller, K., and Dubrawski, A. Mutually regressive point processes. In *Advances in Neural Information Processing Systems*, pp. 5115–5126, 2019.

Bishop, C. Pattern recognition and machine learning. *Pattern Recognition and Machine Learning*, 2006.

Brémaud, P. and Massoulié, L. Stability of nonlinear hawkes processes. *The Annals of Probability*, pp. 1563–1588, 1996.

Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., and Frank, L. M. The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*, 14(2):325–346, 2002.

Csató, L., Opper, M., and Winther, O. Tap gibbs free energy, belief propagation and sparsity. In *NIPS*, pp. 657–663, 2001.

Daley, D. J. and Vere-Jones, D. *An introduction to the theory of point processes, volume 1: Elementary theory and methods*. Verlag New York Berlin Heidelberg: Springer, 2003.

Daley, D. J. and Vere-Jones, D. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

Dayan, P. and Abbott, L. F. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational Neuroscience Series, 2001.

Donner, C. and Opper, M. Efficient bayesian inference of sigmoidal gaussian cox processes. *The Journal of Machine Learning Research*, 19(1):2710–2743, 2018a.

Donner, C. and Opper, M. Efficient bayesian inference for a gaussian process density model. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2018b.

Gerhard, F., Deger, M., and Truccolo, W. On the stability and dynamics of stochastic spiking neuron models: Nonlinear hawkes process and point process glms. *PLoS computational biology*, 13(2):e1005390, 2017.

Hawkes, A. G. and Oakes, D. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kingman, J. F. C. *Poisson processes*. Wiley Online Library, 1993.

Lewis, P. W. and Shedler, G. S. Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.

Maffei, A., Nelson, S. B., and Turrigiano, G. G. Selective reconfiguration of layer 4 visual cortical circuitry by visual deprivation. *Nature neuroscience*, 7(12):1353–1359, 2004.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

Ogata, Y. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.

Polson, N. G., Scott, J. G., and Windle, J. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Rasmussen, J. G. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15 (3):623–642, 2013.

Smith, T. C. and Jahr, C. E. Self-inhibition of olfactory bulb neurons. *Nature neuroscience*, 5(8):760–766, 2002.

Titsias, M. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Truccolo, W. From point process observations to collective neural dynamics: Nonlinear hawkes process glms, low-dimensional dynamics and coarse graining. *Journal of Physiology-Paris*, 110(4):336–347, 2016.

Zhang, R., Walder, C., Rizoiu, M.-A., and Xie, L. Efficient non-parametric bayesian hawkes processes. *arXiv preprint arXiv:1810.03730*, 2018.

Zhang, R., Walder, C., and Rizoiu, M.-A. Variational inference for sparse gaussian process modulated hawkes process. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6803–6810, 2020.

Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., and Leskovec, J. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1513–1522. ACM, 2015.

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. Efficient em-variational inference for hawkes process. *arXiv preprint arXiv:1905.12251*, 2019.

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. Efficient inference for nonparametric hawkes processes using auxiliary latent variables. *Journal of Machine Learning Research*, 21(241):1–31, 2020.

Zhu, L. Central limit theorem for nonlinear hawkes processes. *Journal of Applied Probability*, 50(3):760–771, 2013.

Supplementary Material

1. Inference for NH–GPS model

1.1. Model Augmentation

The first step we take in treating the likelihood function is using the Pólya-Gamma (PG) augmentation scheme. Following Theorem 1 in Polson et al. (2013), we can rewrite the nonlinear intensity function as

$$\sigma\left(\phi\left(t\right)\right) = \int_0^\infty e^{f\left(w,\phi\left(t\right)\right)} PG\left(w;1,0\right) dw \qquad (10)$$

$$f\left(w,\phi\left(t\right)\right) = -\frac{\phi\left(t\right)^2 w}{2} + \frac{\phi\left(t\right)}{2} - \ln 2. \qquad (11)$$

As we augment each observation with a variable w_n from a PG distribution, the joint likelihood of the observed events $\{t_n\}$ and PG variables $\{w_n\}$ is

$$p\left(\{t_n\}_{n=1}^N, \{w_n\}_{n=1}^N | \phi, \lambda^*\right) = \qquad (12)$$
$$\exp\left(-\int_0^T \lambda^* \sigma\left(\phi\left(t\right)\right) dt\right) \cdot \prod_{n=1}^N \lambda^* e^{f\left(w_n,t_n\right)} PG\left(w_n;1,0\right)$$

with

$$\exp\left\{-\int_0^T \lambda^* \sigma\left(\phi\left(t\right)\right) dt\right\} = \qquad (13)$$
$$\exp\left(-\int_0^T \int_0^\infty \lambda^* PG\left(w;1,0\right)\left(1 - e^{f\left(w,-\phi\left(t\right)\right)}\right) dw dt\right).$$

Where we used $\sigma(t) = 1 - \sigma(-t)$.

Next, we utilize the Campbell's theorem (Kingman, 1993) which states that for a Poisson process Π with intensity φ

$$\mathbb{E}_\varphi\left(\prod_{x\in\Pi} \exp\left(h\left(x\right)\right)\right) =$$
$$\exp\left(-\int\left(1 - \exp\left(h\left(x\right)\right)\right)\varphi\left(x\right) dx\right).$$

Looking at Equation 13 we identify $x = (t, w)$ and $\varphi\left(t,w\right) = \lambda^* PG\left(w|1,0\right)$ is the intensity of a marked Poisson process in $\mathcal{T}$ with marks $w \sim PG\left(0,1\right)$. Further, we determine $h\left(x\right) = f\left(w, -\phi\left(t\right)\right)$. We can now rewrite the exponential in Equation 12 as

$$\exp\left\{-\int_0^T \lambda^* \sigma\left(\phi\left(t\right)\right) dt\right\} = \mathbb{E}_\varphi\left(\prod_{m=1}^M e^{f\left(\hat{w}_m,\hat{t}_m\right)}\right) \qquad (14)$$

for realizations $\{\hat{t}_m, \hat{w}_m\}_{m=1}^M$.

We substitute Equation 14 into Equation 12 which results in the full augmented likelihood. Given the prior distribution over ϕ and λ^*, we can now write the model's posterior distribution as

$$p\left(\{\hat{t}_m,\hat{w}_m\},\{w_n\},\phi,\lambda^*|\{t_n\}\right) \propto \exp\left(-\lambda^* T\right) \times \qquad (15)$$
$$\prod_{m=1}^M \lambda^* e^{f\left(\hat{w}_m,-\phi\left(\hat{t}_m\right)\right)} PG\left(w_m;1,0\right) \times$$
$$\prod_{n=1}^N \lambda^* e^{f\left(w_n,\phi\left(t_n\right)\right)} PG\left(w_n;1,0\right) \times p\left(\phi\right) p\left(\lambda^*\right).$$

To summarize, we augment the model with two sets of variables – the PG variables $\{w_n\}$ which augment the actual realizations and the tuples $\{\hat{t}_m, \hat{w}_m\}$ which are the realizations and marks of the auxiliary marked Poisson process.

1.2. Variational Inference for NH–GPS model

In variational inference (Jordan et al., 1999; Bishop, 2006) we define a tractable distribution family and adapt it to approximate the posterior by maximizing the lower bound $\mathcal{L}(Q)$ defined below. This procedure minimizes the Kullback–Leibler divergence between the unknown posterior and the proposed approximating distribution. The posterior density is approximated by

$$p\left(\{\hat{t}_m,\hat{w}_m\},\{w_n\},\phi,\lambda^*|\{t_n\}\right)$$
$$\approx q_1\left(\phi,\lambda^*\right) q_2\left(\{w_n\}_{n=1}^N,\{\hat{t}_m,\hat{w}_m\}_{m=1}^M\right).$$

This leads to the following lower bound on the evidence

$$\mathcal{L}(Q) = \mathbb{E}_Q\left[\log\left\{\frac{p\left(\{\hat{t}_m,\hat{w}_m\},\{w_n\},\phi,\lambda^*|\{t_n\}\right)}{q_1\left(\phi,\lambda^*\right) q_2\left(\{w_n\}_{n=1}^N,\{\hat{t}_m,\hat{w}_m\}_{m=1}^M\right)}\right\}\right].$$

Here Q refers to the probability measure of the variational posterior. We can maximize the bound by alternating the maximization over each of the factors (Bishop, 2006). The optimal solution for each factor is

$$\log q_1^*\left(\phi,\lambda^*\right) = \qquad (16)$$
$$\mathbb{E}_{q_2\left(\{w_n\}_{n=1}^N,\{\hat{t}_m,\hat{w}_m\}_{m=1}^M\right)}[\log P(\{\hat{t}_m,\hat{w}_m\},\{w_n\},\phi,\lambda^*,\{t_n\})]$$

$$\log q_2^*\left(\{w_n\}_{n=1}^N,\{\hat{t}_m,\hat{w}_m\}_{m=1}^M\right) = \qquad (17)$$
$$\mathbb{E}_{q_1(\phi,\lambda^*)}[\log P(\{\hat{t}_m,\hat{w}_m\},\{w_n\},\phi,\lambda^*,\{t_n\})].$$

Thus, to obtain the optimal distribution of one of the factors, one must calculate expectations of the logarithm of the joint distribution over the remaining factors in the approximation, resulting in an iterative algorithm.

In the following subsections, we explicitly express the functional form of the optimal distributions, and obtain the corresponding expectations required for updating the factors. The hyperparameters ($\{\sigma_g, a_g, \beta, \sigma_s, a_s\}$) are learned via gradients update of the lower bound, we present details in the Supplementary Material.

1.3. Optimal q_1

We find that the optimal q_1 is factorized as

$$q_1\left(\phi, \lambda^*\right) = q_1\left(\lambda^*\right) q_1\left(\phi\right)$$

The first factor is identified as a Gamma distribution

$$\begin{aligned} q_1\left(\lambda^*\right) &= Gamma\left(\beta, \beta\right) \\ \beta &= \beta_0 + N + \int_{\mathcal{T}x\mathcal{W}} \lambda_{q_2}\left(t, w\right) dt dw \\ \beta &= \beta_0 + T \end{aligned} \tag{18}$$

with known expectations.

The optimal distribution for the second factor is of the form

$$\begin{aligned} q_1^\star &\propto e^{-U(\phi)+\log p(\phi)} \\ U(\phi) &= \frac{1}{2}\int A(t)\phi^2(t)dt - \int b(t)\phi(t)dt \\ A(t) &= \sum_n \langle \omega_n \rangle_{q_2^\star} \delta(t-t_n) + \langle \omega(t) \rangle_{q_2^\star} \lambda_{q_2^\star}(t) \\ b(t) &= \sum_n \frac{1}{2}\delta(t-t_n) - \frac{1}{2}\lambda_{q_2^\star}\left(t\right). \end{aligned}$$

Generally, the integrals above cannot be evaluated analytically. Thus, we resort to another variational approximation, where we approximate the likelihood term, by a distribution that depends only on a finite set of inducing point $\{c\}$, $\tilde{q}\left(\phi_c, \phi\right) = p\left(\phi|\phi_c\right) q\left(\phi_c\right)$ and the ELBO is

$$\left\langle \log \frac{e^{-\log\langle U(\phi)\rangle_{p(\phi|\phi_c)}} p(\phi_c)}{\tilde{q}(\phi_c)} \right\rangle_{\tilde{q}}$$

and we use the notation $\langle p \rangle_q = \mathbb{E}_q\left(p\right)$. The optimal $\tilde{q}\left(\phi_c\right)$ is given by

$$\tilde{q}^\star(\boldsymbol{\phi}_c) \propto e^{-\log\langle U(\boldsymbol{\phi})\rangle_{p(\phi|\phi_c)}} p(\boldsymbol{\phi}_c).$$

From here, using known results of conditional GPs and sparse variational GPs (Csató et al., 2001; Titsias, 2009) we have

$$\begin{aligned} \tilde{q}^\star(\boldsymbol{\phi}_c) &= \mathcal{N}(\boldsymbol{\phi}_c|\boldsymbol{\mu}_c, \Sigma_c) \\ \Sigma_c &= \left[\int \boldsymbol{\kappa}(t)^\top A(t)\boldsymbol{\kappa}(t)dt + K_c^{-1}\right]^{-1} \\ \boldsymbol{\mu}_c &= \Sigma_c\left(\int b(t)\boldsymbol{\kappa}(t)dt\right) \end{aligned} \tag{19}$$

with K_c the covariance kernel between the inducing points, $\boldsymbol{\kappa}(t) = \boldsymbol{k}_c(t)^\top K_c^{-1}$ and $\boldsymbol{k}_c(t)$ is the kernel between the inducing points and another set of points (either the real data or the integration points). The mean and the variance of the sparse approximated GP are

$$\begin{aligned} \langle g(t) \rangle =& \boldsymbol{\kappa}(t)\boldsymbol{\mu}_c \\ \sigma^2(t) =& K(t,t) - \boldsymbol{\kappa}(t)^\top \boldsymbol{k}_c(t) + \boldsymbol{\kappa}(t)^\top \Sigma_c \boldsymbol{\kappa}(t) \end{aligned} \tag{20}$$

1.4. Optimal q_2

Similarly to the previous section, we find that the optimal q_2 is factorized as

$$q_2\left(\{w_n\}_{n=1}^N, \Pi\right) = q_2\left(\{w_n\}_{n=1}^N\right) q_2\left(\{\hat{t}_m, \hat{w}_m\}\right)$$

Given Equation 15, we define the first factor as

$$q_2^\star(w_n) \propto \exp\left(-\frac{\langle \phi_n^2 \rangle_{q_1^\star}}{2} w_n\right) PG(w_n|1, 0),$$

which corresponds to a tilted PG distribution

$$q_2^\star(w_n) = PG\left(w_n|1, \sqrt{\langle \phi_n^2 \rangle_{q_1^\star}}\right). \tag{21}$$

with known expectations (Polson et al., 2013).

The second factor takes the form

$$\begin{aligned} & q_2^\star(\{\hat{t}_m, \hat{w}_m\}_{m=1}^M) \\ & \propto \prod_{m=1}^M \exp\left(-\frac{\langle \phi_m \rangle_{q_1^\star}}{2} - \frac{\langle \phi_m^2 \rangle_{q_1^\star}}{2} w_m\right) \cdot \exp\left(\langle \ln \lambda^\star \rangle_{q_1^\star}\right). \end{aligned}$$

It can be shown that this distribution corresponds to a Poisson process with intensity function

$$\begin{aligned} & \lambda_{q_2}\left(\hat{t}, \hat{w}\right) \\ & = \exp\left(\langle \ln \lambda^* \rangle_{q_1^\star}\right) \frac{\exp\left(-\frac{\langle \phi \rangle_{q_1^\star}}{2}\right)}{2\cosh\left(\langle \phi^2 \rangle_{q_1^\star}\right)} PG\left(w_m|1, \sqrt{\langle \phi^2 \rangle_{q_1^\star}}\right) \end{aligned} \tag{22}$$

where to simplify the notation we write ϕ instead of $\phi\left(\hat{t}\right)$.

2. Results

2.1. Crime Data

Here, we describe the computation of the results reported in Table 1. Once we are done fitting the model to the training data, we have an approximation for the posterior distribution $\mathcal{Q}$. To estimate the likelihood on the test data, we sample the model parameters from $\mathcal{Q}$ multiple times and evaluate

$$\ell\left(D_{\text{test}}\right) \approx \ln \mathbb{E}_{\mathcal{Q}}\left(p\left(D_{\text{test}}|\phi, \lambda^*, w\right)\right). \tag{23}$$

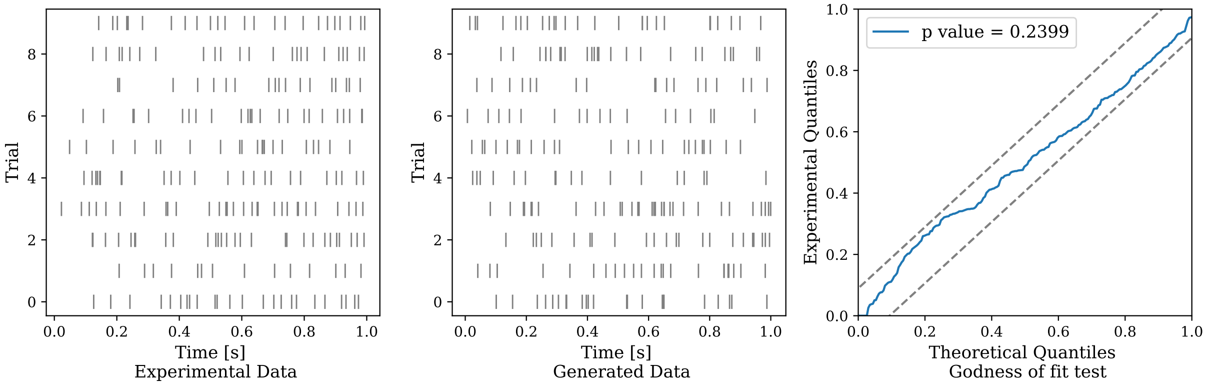

Figure 3. (a) raster plot of a neuron from a monkey cortex. (b) Data generated from the learned model. (c) Results of the Kolmogorov-Smirnov test. The NH–GPS generates data that resembles the real data, and passes the goodness of fit test.

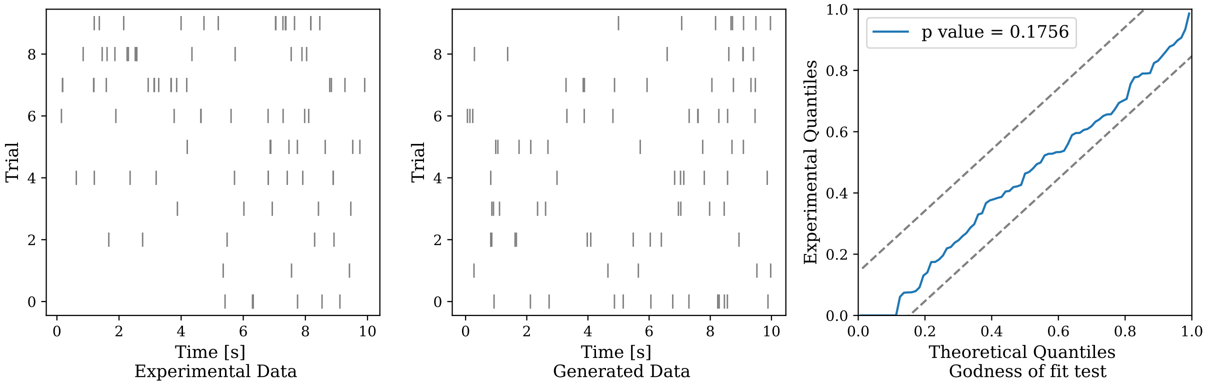

Figure 4. (a) raster plot of a neuron from a human cortex. (b) Data generated from the learned model. (c) Results of the Kolmogorov-Smirnov test. The NH–GPS generates data that resembles the real data, and passes the goodness of fit test.

2.2. Neuronal Data

In Figures 3 and 4 we assess the ability of the model to capture the data for the recordings from monkey cortex and human cortex. In both figures, panel a and b present the raster plot of the real data and the raster plot generated from the fitted model respectively. Similarly to the real data, the generated data displays both excitation, in the form of clustered events, and inhibition.

The results of the Kolmogorov-Smirnov test are displayed in the in Panel c. The comparison relies between the 95% confidence bounds, which are indicated by the dashed lines. The model passes the goodness of fit test (p value > 0.05).