Journal of Machine Learning Research 19 (2018) 1-34 Submitted 12/17; Revised 10/18; Published 11/18

Efficient Bayesian Inference of Sigmoidal Gaussian Cox Processes

Christian Donner CHRISTIAN.DONNER@BCCN-BERLIN.DE
Manfred Opper MANFRED.OPPER@TU-BERLIN.DE
Artificial Intelligence Group
Technische Universität Berlin
Berlin, Germany

Editor: Ryan Adams

Abstract

We present an approximate Bayesian inference approach for estimating the intensity of a inhomogeneous Poisson process, where the intensity function is modelled using a Gaussian process (GP) prior via a sigmoid link function. Augmenting the model using a latent marked Poisson process and Pólya–Gamma random variables we obtain a representation of the likelihood which is conjugate to the GP prior. We estimate the posterior using a variational free–form mean field optimisation together with the framework of sparse GPs. Furthermore, as alternative approximation we suggest a sparse Laplace's method for the posterior, for which an efficient expectation–maximisation algorithm is derived to find the posterior's mode. Both algorithms compare well against exact inference obtained by a Markov Chain Monte Carlo sampler and standard variational Gauss approach solving the same model, while being one order of magnitude faster. Furthermore, the performance and speed of our method is competitive with that of another recently proposed Poisson process model based on a quadratic link function, while not being limited to GPs with squared exponential kernels and rectangular domains.

Keywords: Poisson process; Cox process; Gaussian process; data augmentation; variational inference

1. Introduction

Estimating the intensity rate of discrete events over a continuous space is a common problem for real world applications such as modeling seismic activity (Ogata, 1998), neural data (Brillinger, 1988), forestry (Stoyan and Penttinen, 2000) and so forth. A particularly common approach is a Bayesian model based on a so–called Cox process (Cox, 1955). The observed events are assumed to be generated from a Poisson process, whose intensity function is modeled as another random process with a given prior probability measure. The problem of inference for such type of models has also attracted interest in the Bayesian machine learning community in recent years. Møller et al. (1998); Brix and Diggle (2001); Cunningham et al. (2008) assumed that the intensity function is sampled from a Gaussian Process (GP) prior (Rasmussen and Williams, 2006). However, to restrict the intensity function of the Poisson process to nonnegative values, a common strategy is to choose a nonlinear link function which takes the GP as its argument and returns a valid intensity. Based on the success of variational approximations to deal with complex Gaussian process models, the inference problem for such Poisson models has attracted considerable interest in the machine learning community.

While powerful black–box variational Gaussian inference algorithms are available which can be applied to arbitrary link–functions, the choice of link–functions is not only crucial for defining the prior over intensities but can also be important for the efficiency of variational inference. The 'standard' choice of Cox processes with an exponential link function was treated in (Hensman et al., 2015). However, variational Gaussian inference for this link function has the disadvantage that the posterior variance becomes decoupled from the observations (Lloyd et al., 2015).[1] An interesting choice is the quadratic link function of (Lloyd et al., 2015) for which integrations over the data domain, which are necessary for sparse GP inference, can be (for specific kernel) computed analytically.[2] For both models, the minimisation of the variational free energies is performed by gradient descent techniques.

In this paper we will deal with approximate inference for a model with a sigmoid link–function. This model was introduced by (Adams et al., 2009) together with a MCMC sampling algorithm which was further improved by (Gunter et al., 2014) and (Teh and Rao, 2011). Kirichenko and van Zanten (2015) have shown that the model has favourable (frequentist) theoretical properties provided priors and hyperparameters are chosen appropriately. In contrast to a direct variational Gaussian approximation for the posterior distribution of the latent function, we will introduce an alternative type of variational approximation which is specially designed for the *sigmoidal Gaussian Cox process*. We build on recent work on Bayesian logistic regression by data augmentation with Pólya–Gamma random variables (Polson et al., 2013). This approach was already used in combination with GPs (Linderman et al., 2015; Wenzel et al., 2017), for stochastic processes in discrete time (Linderman et al., 2017), and for jump processes (Donner and Opper, 2017). We extend this method to an augmentation by a latent, marked Poisson process, where the marks are distributed according to a Pólya–Gamma distribution.[3] In this way, the augmented likelihood becomes conjugate to a GP distribution. Using a combination of a mean–field variational approximation together with sparse GP approximations (Csató and Opper, 2002; Csató, 2002; Titsias, 2009) we obtain explicit analytical variational updates leading to fast inference. In addition, we show that the same augmentation can be used for the computation of the maximum a posteriori (MAP) estimate by an expectation–maximisation (EM) algorithm. With this we obtain a Laplace approximation to the non–augmented posterior.

The paper is organised as follows: In section 2, we introduce the sigmoidal Gaussian Cox process model and its transformation by the variable augmentation. In section 3, we derive a variational mean field method and an EM–algorithm to obtain the MAP estimate, followed by the Laplace approximation of the posterior. Both methods are based on a sparse GP approximation to make the infinite dimensional problem tractable. In section 4, we demonstrate the performance of our method on synthetic data sets and compare with the results of a Monte Carlo sampling method for the model and the variational approximation of Hensman et al. (2015), which we modify to solve the Cox–process model with the scaled sigmoid link function. Then we compare our method to the state-of-the-art inference

[1] Samo and Roberts (2015) propose an efficient approximate sampling scheme.

[2] For a frequentist nonparametric approach to this model, see (Flaxman et al., 2017). For a Bayesian extension see (Walder and Bishop, 2017).

[3] For a different application of marked Poisson processes, see (Lloyd et al., 2016).

algorithm (Lloyd et al., 2015) on artificial and real data sets with up to 10^4 observations. Section 5 presents a discussion and an outlook.

2. The Inference problem

We assume that N events $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$ are generated by a Poisson process. Each point $\boldsymbol{x}_n$ is a d–dimensional vector in the compact domain $\mathcal{X} \subset \mathbb{R}^d$. The goal is to infer the varying *intensity function* $\Lambda(\boldsymbol{x})$ (the mean measure of the process) for all $\boldsymbol{x} \in \mathcal{X}$ based on the likelihood

$$L(\mathcal{D}|\Lambda) = \exp\left(-\int_{\mathcal{X}} \Lambda(\boldsymbol{x})d\boldsymbol{x}\right)\prod_{n=1}^N \Lambda(\boldsymbol{x}_n),$$

which is equal (up to a constant) to the density of a Poisson process having intensity Λ (see Appendix C and (Konstantopoulos et al., 2011)) with respect to a Poisson process with unit intensity. In a Bayesian framework, a prior over the intensity makes Λ a random process. Such a doubly stochastic point process is called *Cox process* (Cox, 1955). Since one needs $\Lambda(\boldsymbol{x}) \geq 0$, Adams et al. (2009) suggested a reparametrization of the intensity function by $\Lambda(\boldsymbol{x}) = \lambda\sigma(g(\boldsymbol{x}))$, where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function and λ is the maximum intensity rate. Hence, the intensity $\Lambda(\boldsymbol{x})$ is positive everywhere, for any arbitrary function $g(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ and the inference problem is to determine this function. Throughout this work we assume that $g(\cdot)$ will be modelled as a GP (Rasmussen and Williams, 2006) and the resulting process is called *sigmoidal Gaussian Cox process*. The likelihood for g becomes

$$L(\mathcal{D}|g, \lambda) = \exp\left(-\int_{\mathcal{X}} \lambda\sigma(g(\boldsymbol{x}))d\boldsymbol{x}\right)\prod_{n=1}^N \lambda\sigma(g_n), \tag{1}$$

where $g_n \doteq g(\boldsymbol{x}_n)$. For Bayesian inference we define a GP prior measure P_{GP} with zero mean and covariance kernel $k(\boldsymbol{x}, \boldsymbol{x}') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. λ has as prior density (with respect to the ordinary Lebesgue measure) $p(\lambda)$ which we take to be a Gamma density with shape-, and rate parameter α_0 and β_0, respectively. Hence, for the prior we get the product measure $dP_{\text{prior}} = dP_{\text{GP}} \times p(\lambda)d\lambda$. The posterior density $\boldsymbol{p}$ (with respect to the prior measure) is given by

$$\boldsymbol{p}(g, \lambda|\mathcal{D}) \doteq \frac{dP_{\text{posterior}}}{dP_{\text{prior}}}(g, \lambda|\mathcal{D}) = \frac{L(\mathcal{D}|g, \lambda)}{\mathbb{E}_{P_{\text{prior}}}\left[L(\mathcal{D}|g, \lambda)\right]}. \tag{2}$$

The normalising expectation in the denominator on the right hand side is with respect to the probability measure P_{prior}. To deal with the infinite dimensionality of GPs and Poisson processes we require a minimum of extra notation. We introduce densities or *Radon–Nikodým derivatives* such as defined in Equation (2) (see Appendix C or de G. Matthews et al. (2016)) with respect to infinite dimensional measures by boldface symbols $\boldsymbol{p}(\boldsymbol{z})$. On the other hand, non–bold densities $p(\boldsymbol{z})$ denote densities in the 'classical' sense, which means they are with respect to Lebesgue measure $d\boldsymbol{z}$.

Bayesian inference for this model is known to be doubly intractable (Murray et al., 2006). The likelihood in Equation (1) contains the integral of g over the space $\mathcal{X}$ in the exponent and the normalisation of the posterior in Equation (2) requires calculating expectation of Equation (1). In addition inference is hampered by the fact, that likelihood (1) depends non–linearly on g (through sigmoid and exponent of sigmoid). In the following we tackle this by an augmentation scheme for the likelihood, such that it becomes conjugate to a GP prior and we subsequently can derive an analytic form of a variational posterior given one simple mean field assumption (Section 3).

2.1. Data augmentation I: Latent Poisson process

We will briefly introduce a data augmentation scheme by a latent Poisson process which forms the basis of the sampling algorithm of Adams et al. (2009). We will then extend this method further to an augmentation by a *marked* Poisson process. We focus on the exponential term in Equation (1). Utilizing the well known property of the sigmoid that $\sigma(x) = 1 - \sigma(-x)$ we can write

$$\exp\left(-\int_{\mathcal{X}} \lambda\sigma(g(\boldsymbol{x}))d\boldsymbol{x}\right) = \exp\left(-\int_{\mathcal{X}} \left(1 - \sigma(-g(\boldsymbol{x}))\right)\lambda d\boldsymbol{x}\right). \tag{3}$$

The left hand side has the form of a characteristic functional of a Poisson process. Generally, for a random set of points $\Pi_{\mathcal{Z}} = \{\boldsymbol{z}_m; \boldsymbol{z}_m \in \mathcal{Z}\}$ on a space $\mathcal{Z}$ and with a function $h(\boldsymbol{z})$, this is defined as

$$\mathbb{E}_{P_\Lambda}\left[\prod_{\boldsymbol{z}_m \in \Pi_{\mathcal{Z}}} e^{h(\boldsymbol{z}_m)}\right] = \exp\left(-\int_{\mathcal{Z}} \left(1 - e^{h(\boldsymbol{z})}\right)\Lambda(\boldsymbol{z})d\boldsymbol{z}\right), \tag{4}$$

where P_Λ is the probability measure of a Poisson process with intensity $\Lambda(\boldsymbol{z})$. Equation (4) can be derived by Campbell's theorem (see Appendix A and (Kingman, 1993, chap. 3)) and identifies a Poisson process uniquely.

Setting $h(\boldsymbol{z}) = \ln\sigma(-g(\boldsymbol{z}))$, and $\mathcal{Z} = \mathcal{X}$, and combining Equation (3) and (4) we obtain the likelihood used by Adams et al. (2009, Eq. 4). However, in this work we make use of another augmentation, before invoking Campbell's theorem. This will result in a likelihood which is conjugate to the model priors and further simplifies inference.

2.2. Data augmentation II: Pólya–Gamma variables and marked Poisson process

Following Polson et al. (2013) we represent the inverse of the hyperbolic cosine as a scaled Gaussian mixture model

$$\cosh^{-b}(z/2) = \int_0^\infty e^{-\frac{z^2}{2}\omega} p_{\mathrm{PG}}(\omega|b,0)d\omega, \tag{5}$$

where p_{PG} is a *Pólya–Gamma* density (Appendix B). We further define the *tilted* Pólya–Gamma density by

$$p_{\mathrm{PG}}(\omega|b,c) \propto e^{-\frac{c^2}{2}\omega} p_{\mathrm{PG}}(\omega|b,0), \tag{6}$$

where $b > 0$ and c are parameters. We will not need an explicit form of this density, since the subsequently derived inference algorithms will only require the first moments. Those can be obtained directly from the moment generating function, which can be calculated straightforwardly from Equation (5) and (6) (see Appendix B). Equation (5) allows us to rewrite the sigmoid function as

$$\sigma(z) = \frac{e^{\frac{z}{2}}}{2\cosh(\frac{z}{2})} = \int_0^\infty e^{f(\omega,z)} p_{\mathrm{PG}}(\omega|1,0) d\omega, \tag{7}$$

where we define

$$f(\omega, z) \doteq \frac{z}{2} - \frac{z^2}{2}\omega - \ln 2.$$

Setting $z = -g(\boldsymbol{x})$ in Equation (3) and substituting Equation (7) we get

$$\exp\left(-\int_{\mathcal{X}} \lambda\left(1 - \sigma(-g(\boldsymbol{x}))\right) d\boldsymbol{x}\right) = \exp\left(-\int_{\mathcal{X}\times\mathbb{R}^+} \left(1 - e^{f(\omega,-g(\boldsymbol{x}))}\right)\ p_{\mathrm{PG}}(\omega|1,0)\ \lambda d\omega d\boldsymbol{x}\right). \tag{8}$$

Finally, we apply Campbell's theorem (Equation (4)) to Equation (8). The space is a product space $\mathcal{Z} = \hat{\mathcal{X}} \doteq \mathcal{X} \times \mathbb{R}^+$ and the intensity $\Lambda(\boldsymbol{x}, \omega) = \lambda p_{\mathrm{PG}}(\omega|1,0)$. This results in the final representation of the exponential in Equation (8)

$$\exp\left(-\int_{\hat{\mathcal{X}}} \left(1 - e^{f(\omega,-g(\boldsymbol{x}))}\right) \Lambda(\boldsymbol{x},\omega)\ d\omega d\boldsymbol{x}\right) = \mathbb{E}_{P_\Lambda}\left[\prod_{(\boldsymbol{x},\omega)_m \in \Pi_{\hat{\mathcal{X}}}} e^{f(\omega_m,-g_m)}\right].$$

Interestingly, the new Poisson process $\Pi_{\hat{\mathcal{X}}}$ with measure P_Λ has the form of a *marked* Poisson process (Kingman, 1993, chap. 5), where the latent Pólya-Gamma variables ω_m denote the 'marks' being independent random variables at each location $\boldsymbol{x}_m$. It is straightforward to sample such processes by first sampling the inhomogeneous Poisson process on domain $\mathcal{X}$ (for example by 'thinning' a process with constant rate (Lewis and Shedler, 1979; Adams et al., 2009)) and then drawing a mark ω on each event independently from the density $p_{\mathrm{PG}}(\omega|1,0)$.

Finally, using the Pólya–Gamma augmentation also for the discrete likelihood factors corresponding to the observed events in Equation (1) we obtain the following joint likelihood of the model

$$\begin{aligned} L(\mathcal{D}, \boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}|g,\lambda) &\doteq \frac{dP_{\text{joint}}}{dP_{\text{aug}}}(\mathcal{D}, \boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}|g,\lambda) \\ &= \prod_{(\boldsymbol{x},\omega)_m \in \Pi_{\hat{\mathcal{X}}}} e^{f(\omega_m,-g_m)} \prod_{n=1}^N \lambda e^{f(\omega_n,g_n)}, \end{aligned} \tag{9}$$

where we define the prior measure of augmented variables as $P_{\text{aug}} = P_\Lambda \times P_{\boldsymbol{\omega}_N}$ and where $\boldsymbol{\omega}_N = \{\omega_n\}_{n=1}^N$ are the Pólya–Gamma variables for the observations $\mathcal{D}$ with the prior measure $dP_{\boldsymbol{\omega}_N} = \prod_{n=1}^N p(\omega_n|1,0)d\omega_n$. This augmented representation of the likelihood contains the function $g(\cdot)$ only linearly and quadratically in the exponents and is thus conjugate to the GP prior of $g(\cdot)$. Note that the original likelihood in Equation (1) can be recovered by $\mathbb{E}_{P_{\text{aug}}}\left[L(\mathcal{D}, \boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}|g,\lambda)\right] = L(\mathcal{D}|g,\lambda)$.

3. Inference in the augmented space

Based on the augmentation we define a posterior density for the joint model with respect to the product measure $P_{\text{prior}} \times P_{\text{aug}}$

$$\begin{aligned}\boldsymbol{p}(\boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}, g, \lambda|\mathcal{D}) &\doteq \frac{dP_{\text{posterior}}}{d(P_{\text{prior}} \times P_{\text{aug}})}(\boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}, g, \lambda|\mathcal{D}) \\ &= \frac{L(\mathcal{D}, \boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}|g, \lambda)}{L(\mathcal{D})},\end{aligned} \tag{10}$$

where the denominator is the marginal likelihood $L(\mathcal{D}) = \mathbb{E}_{P_{\text{prior}} \times P_{\text{aug}}}\left[L(\mathcal{D}, \boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}|g, \lambda)\right]$. The posterior density of Equation (10) could be sampled using Gibbs sampling with explicit, tractable conditional densities. Similar to the variational approximation in the next section, one can show that the conditional measure of the point sets $\Pi_{\hat{\mathcal{X}}}$ and the variables $\boldsymbol{\omega}_N$, given the function $g(\cdot)$ and maximal intensity λ is a product of a specific marked Poisson process and independent (tilted) Pólya–Gamma densities. On the other hand, the distribution over function $g(\cdot)$ conditioned on $\Pi_{\hat{\mathcal{X}}}$ and $\boldsymbol{\omega}_N$ is a Gaussian process. Note, however, one needs to sample this GP only at the finite points $\boldsymbol{x}_m$ in the random set $\Pi_{\hat{\mathcal{X}}}$ and the fixed set $\mathcal{D}$.

3.1. Variational mean–field approximation

For variational inference one assumes that the desired posterior probability measure belongs to a family of measures for which the inference problem is tractable. Here we make a simple structured mean field assumption in order to fully utilise its conjugate structure: We approximate the posterior measure by

$$P_{\text{posterior}}(\boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}, g, \lambda|\mathcal{D}) \approx Q_1(\boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}) \times Q_2(g, \lambda), \tag{11}$$

meaning that the dependencies between the Pólya–Gamma variables $\boldsymbol{\omega}_N$ and the marked Poisson process $\Pi_{\hat{\mathcal{X}}}$ on the one hand, and the function g and the maximal intensity λ on the other hand, are neglected. As we will see in the following, this simple mean–field assumption allows us to derive the posterior approximation analytically.

The variational approximation is optimised by minimising the Kullback–Leibler divergence between exact and approximated posteriors. This is equivalent to maximising the lower bound on the marginal likelihood of the observations

$$\mathcal{L}(\boldsymbol{q}) = \mathbb{E}_Q\left[\log\left\{\frac{L(\mathcal{D}, \boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}|g, \lambda)}{\boldsymbol{q}_1(\boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}})\boldsymbol{q}_2(g, \lambda)}\right\}\right] \leq \log L(\mathcal{D}), \tag{12}$$

where Q is the probability measure of the variational posterior in Equation (11) and we introduced approximate likelihoods

$$\boldsymbol{q}_1(\boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}) \doteq \frac{dQ_1}{dP_{\text{aug}}}(\boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}), \qquad \boldsymbol{q}_2(g, \lambda) \doteq \frac{dQ_2}{dP_{\text{prior}}}(g, \lambda).$$

Using standard arguments for mean field variational inference (Bishop, 2006, chap. 10) and Equation (11), one can then show that the optimal factors satisfy

$$\ln \boldsymbol{q}_1\left(\boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}\right) = \mathbb{E}_{Q_2}\left[\log L(\mathcal{D}, \boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}|g, \lambda)\right] + \text{const.} \tag{13}$$

and

$$\ln \boldsymbol{q}_2(g, \lambda) = \mathbb{E}_{Q_1}\left[\log L(\mathcal{D}, \boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}|g, \lambda)\right] + \text{const.}\,, \tag{14}$$

respectively. These results lead to an iterative scheme for optimising $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ in order to increase the lower bound in Equation (12) in every step. From the structure of the likelihood one derives two further factorisations:

$$\begin{aligned}\boldsymbol{q}_1(\boldsymbol{\omega}_N, \Pi_{\hat{\mathcal{X}}}) &= \boldsymbol{q}_1(\boldsymbol{\omega}_N)\boldsymbol{q}_1(\Pi_{\hat{\mathcal{X}}}), &(15)\\ \boldsymbol{q}_2(g, \lambda) &= \boldsymbol{q}_2(g)\boldsymbol{q}_2(\lambda), &(16)\end{aligned}$$

where the densities are defined with respect to the measures $dP(\boldsymbol{\omega}_N)$, dP_Λ, dP_{GP}, and $p(\lambda)d\lambda$, respectively. The subsequent section describes these updates explicitly.

Optimal Pólya–Gamma density Following Equation (13) and (15) we obtain

$$\boldsymbol{q}_1(\boldsymbol{\omega}_N) = \prod_{n=1}^{N} \frac{\exp\left(-\frac{c_1^{(n)}}{2}\omega_n\right)}{\cosh^{-1}\left(c_1^{(n)}/2\right)} = \prod_{n=1}^{N} \frac{p_{\mathrm{PG}}\left(\omega_n|1, c_1^{(n)}\right)}{p_{\mathrm{PG}}\left(\omega_n|1, 0\right)},$$

where the factors are *tilts* of the prior Pólya-Gamma densities (see Equation (6) and Appendix B) with $c_1^{(n)} = \sqrt{\mathbb{E}_{Q_2}[g_n^2]}$. By simple density transformation we obtain the density with respect to the Lebesgue measure as

$$q_1(\boldsymbol{\omega}_N) = \boldsymbol{q}_1(\boldsymbol{\omega}_N)\left|\frac{dP_{\boldsymbol{\omega}_N}}{d\boldsymbol{\omega}_N}\right| = \prod_{n=1}^{N} p_{\mathrm{PG}}\left(\omega_n|1, c_1^{(n)}\right), \tag{17}$$

being a product of *tilted* Pólya–Gamma densities.

Optimal Poisson process Using Equation (13) and (15) we obtain

$$\boldsymbol{q}_1(\Pi_{\hat{\mathcal{X}}}) = \frac{\prod_{(\boldsymbol{x},\omega)_m \in \Pi_{\hat{\mathcal{X}}}} e^{\mathbb{E}_{Q_2}[f(\omega_m, -g_m)]}\lambda_1}{\exp\left(\int_{\hat{\mathcal{X}}}\left(e^{\mathbb{E}_{Q_2}[f(\omega, -g(\boldsymbol{x}))]} - 1\right)\lambda_1 p_{\mathrm{PG}}(\omega|1,0)d\boldsymbol{x}d\omega\right)}, \tag{18}$$

with $\lambda_1 \doteq e^{\mathbb{E}_{Q_2}[\log \lambda^*]}$. Note, that $\mathbb{E}_{Q_2}[f(\omega_m, -g_m)]$ involves the expectations $\mathbb{E}_{Q_2}[g_m]$ and $\mathbb{E}_{Q_2}\left[(g_m)^2\right]$. One can show, that Equation (18) is again a marked Poisson process with intensity

$$\begin{aligned}\Lambda_1(\boldsymbol{x}, \omega) =& \lambda_1 \frac{\exp\left(-\frac{\mathbb{E}_{Q_2}[g(\boldsymbol{x})]}{2}\right)}{2\cosh\left(\frac{c_1(\boldsymbol{x})}{2}\right)} p_{\mathrm{PG}}\left(\omega|1, c_1(\boldsymbol{x})\right)\\ =& \lambda_1\sigma(-c_1(\boldsymbol{x}))\exp\left(\frac{c_1(\boldsymbol{x}) - \mathbb{E}_{Q_2}[g(\boldsymbol{x})]}{2}\right) p_{\mathrm{PG}}\left(\omega|1, c_1(\boldsymbol{x})\right)\end{aligned} \tag{19}$$

where $c_1(\boldsymbol{x}) = \sqrt{\mathbb{E}_{Q_2}[g(\boldsymbol{x})^2]}$ (for a proof see Appendix D).

Optimal Gaussian process From Equation (14) and (16) we obtain the optimal approximation of the posterior likelihood (note that this is defined relative to GP prior)

$$\boldsymbol{q}_2(g) \propto e^{U(g)},$$

where the effective log–likelihood is given by

$$U(g) = \mathbb{E}_{Q_1}\left[\sum_{(\boldsymbol{x},\omega)_m \in \Pi_{\hat{\mathcal{X}}}} f(\omega_m, -g_m)\right] + \sum_{n=1}^{N} \mathbb{E}_{Q_1}\left[f(\omega_n, g(\boldsymbol{x}_n))\right].$$

The first expectation is over the variational Poisson process $\Pi_{\hat{\mathcal{X}}}$ and the second one over the Pólya–Gamma variables $\boldsymbol{\omega}_N$. These can be easily evaluated (see Appendix A) and one finds

$$U(g) = -\frac{1}{2}\int_{\mathcal{X}} A(\boldsymbol{x})g(\boldsymbol{x})^2 d\boldsymbol{x} + \int_{\mathcal{X}} B(\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x}, \tag{20}$$

with

$$\begin{aligned} A(\boldsymbol{x}) =& \sum_{n=1}^{N} \mathbb{E}_{Q_1}\left[\omega_n\right]\delta(\boldsymbol{x}-\boldsymbol{x}_n) + \int_0^\infty \omega\Lambda_1(\boldsymbol{x},\omega)d\omega, \\ B(\boldsymbol{x}) =& \frac{1}{2}\sum_{n=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_n) - \frac{1}{2}\int_0^\infty \Lambda_1(\boldsymbol{x},\omega)d\omega, \end{aligned}$$

where $\delta(\cdot)$ is the Dirac delta function. The expectations and integrals over ω are

$$\begin{aligned} &\mathbb{E}_{Q_1}\left[\omega_n\right] = \frac{1}{2c_1^{(n)}}\tanh\left(\frac{c_1^{(n)}}{2}\right), \\ &\int_0^\infty \Lambda_1(\boldsymbol{x},\omega)d\omega = \lambda_1\sigma(-c_1(\boldsymbol{x}))\exp\left(\frac{c_1(\boldsymbol{x}) - \mathbb{E}_{Q_2}\left[g(\boldsymbol{x})\right]}{2}\right) \doteq \Lambda_1(\boldsymbol{x}), \\ &\int_0^\infty \omega\Lambda_1(\boldsymbol{x},\omega)d\omega = \frac{1}{2c_1(\boldsymbol{x})}\tanh\left(\frac{c_1(\boldsymbol{x})}{2}\right)\Lambda_1(\boldsymbol{x}). \end{aligned}$$

The resulting variational distribution defines a Gaussian process. Because of the mean–field assumption the integrals in Equation (20) do not require integration over random variables, but only solving two deterministic integrals over space $\mathcal{X}$. However, those integrals depend on function g over the entire space and it is not possible for a general kernel to compute the marginal posterior density at an input $\boldsymbol{x}$ in closed form. For specific GP kernel operators, which are the inverses of differential operators, a solution in terms of linear partial differential equations would be possible. This could be of special interest for one–dimensional problems where Matern kernels with integer parameters (Rasmussen and Williams, 2006) fulfill this condition. Here, the problem becomes equivalent to inference for a (continuous time) Gaussian hidden Markov model and could be solved by performing a forward–backward algorithm (Solin, 2016). This would reduce the computations to the solution of ordinary differential equations. We will discuss details of such an approach elsewhere. To deal with general kernels we will resort instead to a the well known variational sparse GP approximation with inducing points.

Optimal sparse Gaussian process The sparse variational Gaussian approximation follows the standard approach (Csató and Opper, 2002; Csató, 2002; Titsias, 2009) and its generalisation to a continuum likelihood (Batz et al., 2018; de G. Matthews et al., 2016). For completeness, we repeat the derivation here and more detailed in Appendix E. We approximate $\boldsymbol{q}_2(g)$ by a sparse likelihood GP $\boldsymbol{q}_2^s(g)$ with respect to the GP prior

$$\frac{dQ_2^s}{dP}(g) = \boldsymbol{q}_2^s(\boldsymbol{g}_s), \tag{21}$$

which depends only on a finite dimensional vector of function values $\boldsymbol{g}_s = (g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_L))^\top$ at a set of *inducing points* $\{\boldsymbol{x}_l\}_{l=1}^L$. With this approach it is again possible to marginalise out exactly all the infinitely many function values outside of the set of inducing points. The sparse likelihood $\boldsymbol{q}_2^s$ is optimised by minimising the Kullback–Leibler divergence

$$\mathrm{D}_{\mathrm{KL}}(Q_2^s \| Q_2) = \mathbb{E}_{Q_2^s}\left[\log \frac{\boldsymbol{q}_2^s(g)}{\boldsymbol{q}_2(g)}\right].$$

A short computation (Appendix E) shows that

$$q_2^s(\boldsymbol{g}_s) \propto e^{U^s(\boldsymbol{g}_s)} \qquad \text{with } U^s(\boldsymbol{g}_s) = \mathbb{E}_{P(g|\boldsymbol{g}_s)}\left[U(g)\right],$$

where the conditional expectation is with respect to the GP prior measure given the function $\boldsymbol{g}_s$ at the inducing points. The explicit calculation requires the conditional expectations of $g(\boldsymbol{x})$ and of $(g(\boldsymbol{x}))^2$. We get

$$\mathbb{E}_{P(g|\boldsymbol{g}_s)}\left[g(\boldsymbol{x})\right] = \boldsymbol{k}_s(\boldsymbol{x})^\top \, K_s^{-1} \boldsymbol{g}_s, \tag{22}$$

where $\boldsymbol{k}_s(\boldsymbol{x}) = (k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_L))^\top$ and K_s is the kernel matrix between inducing points. For the second expectation, we get

$$\mathbb{E}_{P(g|\boldsymbol{g}_s)}\left[g^2(\boldsymbol{x})\right] = \left(\mathbb{E}_{P(g|\boldsymbol{g}_s)}\left[g(\boldsymbol{x})\right]\right)^2 + \text{const.} \tag{23}$$

The constant equals the conditional variance of $g(\boldsymbol{x})$ which does not depend on the sparse set $\boldsymbol{g}_s$, but only on the locations of the sparse points. Because we are dealing now with a finite problem we can define the 'ordinary' posterior density of the GP at the inducing points with respect to the Lebesgue measure $d\boldsymbol{g}_s$. From Equation (20), (22), and (23), we conclude that the sparse posterior at the inducing variables is a multivariate Gaussian density

$$q_2^s(\boldsymbol{g}_s) = \mathcal{N}(\boldsymbol{\mu}_2^s, \Sigma_2^s), \tag{24}$$

with the covariance matrix given by

$$\Sigma_2^s = \left[K_s^{-1} \int_{\mathcal{X}} A(\boldsymbol{x}) \boldsymbol{k}_s(\boldsymbol{x}) \boldsymbol{k}_s(\boldsymbol{x})^\top d\boldsymbol{x} \; K_s^{-1} + K_s^{-1}\right]^{-1}, \tag{25}$$

and the mean

$$\boldsymbol{\mu}_2^s = \Sigma_2^s \left(K_s^{-1} \int_{\mathcal{X}} B(\boldsymbol{x}) \boldsymbol{k}_s(\boldsymbol{x}) d\boldsymbol{x}\right). \tag{26}$$

In contrast to other variational approximations (see for example (Lloyd et al., 2015; Hensman et al., 2015)) we obtain a closed analytic form of the variational posterior mean and covariance which holds for arbitrary GP kernels. However, these results depend on finite dimensional integrals over the space $\mathcal{X}$ which cannot be computed analytically. This is different to the sparse approximation for the Poisson model with square link function (Lloyd et al., 2015), where similar integrals in the case of the squared exponential kernel can be obtained analytically. Hence, we resort to a simple Monte–Carlo integration, where *integration points* are sampled uniformly on $\mathcal{X}$ as

$$I_F = \int_{\mathcal{X}} F(\boldsymbol{x}) d\boldsymbol{x} \approx \frac{|\mathcal{X}|}{R} \sum_{r=1}^{R} F(\boldsymbol{x}_r).$$

The set of integration points $\{\boldsymbol{x}_r\}_{r=1}^R$ is drawn uniformly from the space $\mathcal{X}$.

Finally, from Equation (21) and (24) we obtain the mean function and the variance of the sparse approximation for every point $\boldsymbol{x} \in \mathcal{X}$, which is

$$\mu_2(\boldsymbol{x}) = \mathbb{E}_{Q_2}\left[g(\boldsymbol{x})\right] = \boldsymbol{k}_s(\boldsymbol{x})^\top K_s^{-1} \boldsymbol{\mu}_2^s, \tag{27}$$

and variance

$$(s_2(\boldsymbol{x}))^2 = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}_s(\boldsymbol{x})^\top K_s^{-1} \left(\mathbf{I} - \Sigma_2^s K_s^{-1}\right) \boldsymbol{k}_s(\boldsymbol{x}), \tag{28}$$

where $\mathbf{I}$ is the identity matrix.

Optimal density for maximal intensity λ From Equation (14) we identify the optimal density as a Gamma density

$$q_2(\lambda) = \mathrm{Gamma}(\lambda|\alpha_2, \beta_2) = \frac{\beta_2^{\alpha_2} (\lambda)^{\alpha_2 - 1} \mathrm{e}^{-\beta_2 \lambda}}{\Gamma(\alpha_2)}, \tag{29}$$

where $\alpha_2 = N + \mathbb{E}_{Q_1}\left[\mathbf{1}_\Pi(\boldsymbol{x})\right] + \alpha_0$, $\beta_2 = \beta_0 + \int_{\mathcal{X}} d\boldsymbol{x}$ and $\Gamma(\cdot)$ is the gamma function. $\mathbf{1}_\Pi(\boldsymbol{x})$ denotes the indicator function being 1 if $\boldsymbol{x} \in \Pi$ and 0 otherwise and the integral is again solved by Monte Carlo integration. This defines the required expectations for updating q_1 by $\mathbb{E}_{Q_2}\left[\lambda\right] = \frac{\alpha_2}{\beta_2}$ and $\mathbb{E}_{Q_2}\left[\log \lambda\right] = \psi(\alpha_2) - \log \beta_2$, where $\psi(\cdot)$ is the digamma function.

Hyperparameters Hyperparameters of the model are (i) the covariance parameters $\boldsymbol{\theta}$ of the GP, (ii) the locations of the inducing points $\{\boldsymbol{x}_l\}_{l=1}^L$, and (iii) the prior parameters α_0, β_0 for the maximal intensity λ. The covariance parameters (i) $\boldsymbol{\theta}$ are optimised by gradient ascent following the gradient of the lower bound in Equation (12) with respect to $\boldsymbol{\theta}$ (Appendix F). As gradient ascent algorithm we employ the ADAM algorithm (Kingma and Ba, 2014). We perform always one step after the variational posterior q is updated as described before. (ii) The locations of the sparse GP $\{\boldsymbol{x}_l\}_{l=1}^L$ could in principle be optimised as well, but we keep them fixed and position them on a regular grid over the space $\mathcal{X}$. From this choice it follows that K_s is a Toeplitz matrix, when the kernel is translationally invariant. This could be inverted in $\mathcal{O}(L(\log L)^2)$ instead of $\mathcal{O}(L^3)$ operations (Press et al., 2007) but we do not employ this fact. Finally, (iii) the value for prior parameters α_0 and β_0 are chosen such that $p(\lambda)$ has a mean twice and standard deviation once the intensity one would expect for a homogeneous Poisson Process observing $\mathcal{D}$. The complete variational procedure is outlined in Algorithm 1.

Algorithm 1: Variational Bayes algorithm for sigmoidal Gaussian Cox process.

```
Init: $\mathbb{E}_Q\left[g(\boldsymbol{x})\right], \mathbb{E}_Q\left[(g(\boldsymbol{x}))^2\right]$ at $\mathcal{D}$ and integration points, and $\mathbb{E}_Q\left[\lambda\right], \mathbb{E}_Q\left[\log\lambda\right]$
while $\mathcal{L}$ not converged do
    Update $q_1$
        PG distributions at observations: $q_1(\boldsymbol{\omega}_N)$ with Eq. (17)
        Rate of latent process: $\Lambda_1(\boldsymbol{x}, \omega)$ at integration points with Eq. (19)
    Update $q_2$
        Sparse GP distribution: $\Sigma_2^s, \boldsymbol{\mu}_2^s$ with Eq. (25), (26)
        GP at $\mathcal{D}$ and integration points: $\mathbb{E}_{Q_2}\left[g(\boldsymbol{x})\right], \mathbb{E}_{Q_2}\left[(g(\boldsymbol{x}))^2\right]$ with
          Eq. (27), (28)
        Gamma-distribution of $\lambda$: $\alpha_2, \beta_2$ with Eq. (29)
    Update kernel parameters with gradient ascent
end
```

3.2. Laplace approximation

In this section we will show that our variable augmentation method is well suited for computing a Laplace approximation (Bishop, 2006, chap. 4) to the joint posterior of the GP function $g(\cdot)$ and the maximal intensity λ as an alternative to the previous variational scheme. To do so we need the maximum a posteriori (MAP) estimate (equal to the mode of the posterior distribution) and a second order Taylor expansion around this mode. The augmentation method will be used to compute the MAP estimator iteratively using an EM algorithm.

Obtaining the MAP estimate In general, a proper definition of the posterior mode would be necessary, because the GP posterior is over a space of functions, which is an infinite dimensional object and does not have a density with respect to Lebesgue measure. A possibility to avoid this problem would be to discretise the spatial integral in the likelihood and to approximate the posterior by a multivariate Gaussian density for which the mode can then be computed by setting the gradient equal to zero. In this paper, we will use a different approach which defines the mode directly in function space and allows us to utilise the sparse GP approximation developed previously for the computations. A mathematically proper way would be to derive the MAP estimator by maximising a properly penalised log–likelihood. As discussed e.g. in Rasmussen and Williams (2006, chap. 6) for GP models with likelihoods which depend on finitely many inputs only, this penalty is given by the squared reproducing kernel Hilbert space (RKHS) norm that corresponds to the GP kernel. Hence, we would have

$$(g^*, \lambda^*) = \operatorname{argmin}_{g\in\mathcal{H}_k,\lambda}\left\{-\ln L(\mathcal{D}|g,\lambda) - \ln p(\lambda) + \frac{1}{2}\|g\|^2_{\mathcal{H}_k}\right\},$$

where $\|g\|^2_{\mathcal{H}_k}$ is the RKHS norm for the kernel k. This penalty term can be understood as a proper generalisation of a Gaussian log–prior density to function space. We will not give a formal definition here but work on a more heuristic level in the following. Rather than attempting a direct optimisation, we will use an EM algorithm instead, applying the variable augmentation with the Poisson process and Pólya–Gamma variables introduced in the previous sections. In this case, the likelihood part of the resulting '$\mathcal{Q}$–function'

$$\mathcal{Q}((g,\lambda)|(g,\lambda)^{\rm old}) \doteq \mathbb{E}_{P(\boldsymbol{\omega}_N,\Pi_{\hat{\mathcal{X}}}|(g,\lambda)^{\rm old})}\left[\ln L(\mathcal{D},\boldsymbol{\omega}_N,\Pi_{\hat{\mathcal{X}}}|g,\lambda)\right] + \ln p(\lambda) - \frac{1}{2}\|g\|^2_{\mathcal{H}_k}, \qquad (30)$$

that needs to be maximised in the M–step becomes (as in the variational approach before) the likelihood of *a Gaussian model* in the GP function g. Hence, we can argue that the function g which maximises $\mathcal{Q}$ is equal to the *posterior mean* of the resulting Gaussian model and can be computed without discussing the explicit form of the RKHS norm.

The conditional probability measure $P(\boldsymbol{\omega}_N,\Pi_{\hat{\mathcal{X}}}|(g,\lambda)^{\rm old})$ is easily obtained similar to the optimal measure Q_1 by not averaging over g and λ. This gives us straightforwardly the density

$$\boldsymbol{p}(\boldsymbol{\omega}_N,\Pi_{\hat{\mathcal{X}}}|(g,\lambda)^{\rm old}) = \boldsymbol{p}(\boldsymbol{\omega}_N|(g,\lambda)^{\rm old})\boldsymbol{p}(\Pi_{\hat{\mathcal{X}}}|(g,\lambda)^{\rm old}).$$

The first factor is

$$p(\boldsymbol{\omega}_N|(g,\lambda)^{\rm old}) = \boldsymbol{p}(\boldsymbol{\omega}_N|(g,\lambda)^{\rm old})\left|\frac{dP_{\boldsymbol{\omega}_N}}{d\boldsymbol{\omega}_N}\right| = \prod_{n=1}^N p_{\rm PG}\left(\omega_n|1,\tilde{c}_n\right),$$

with $\tilde{c}_n = |g_n^{\rm old}|$. The latent point process $\Pi_{\hat{\mathcal{X}}}$ is again a Poisson process density

$$\boldsymbol{p}(\Pi_{\hat{\mathcal{X}}}|(g,\lambda)^{\rm old}) = \frac{dP_{\tilde{\Lambda}}}{dP_{\Lambda}}(\Pi_{\hat{\mathcal{X}}}|(g,\lambda)^{\rm old}),$$

where the intensity is

$$\tilde{\Lambda}(\boldsymbol{x},\omega) = \lambda^{\rm old}\sigma(-g^{\rm old}(\boldsymbol{x}))p_{\rm PG}\left(\omega|1,\tilde{c}(\boldsymbol{x})\right),$$

with $\tilde{c}(\boldsymbol{x}) = |g^{\rm old}(\boldsymbol{x})|$. The first term in the $\mathcal{Q}$–function is

$$\begin{aligned} U(g,\lambda) &\doteq \mathbb{E}_{P(\boldsymbol{\omega}_N,\Pi_{\hat{\mathcal{X}}}|(g,\lambda)^{\rm old})}\left[\ln L(\mathcal{D},\boldsymbol{\omega}_N,\Pi_{\hat{\mathcal{X}}}|g,\lambda)\right] \\ &= -\frac{1}{2}\int_{\mathcal{X}}\tilde{A}(\boldsymbol{x})g(\boldsymbol{x})^2 d\boldsymbol{x} + \int_{\mathcal{X}}\tilde{B}(\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x}, \end{aligned}$$

with

$$\begin{aligned} \tilde{A}(\boldsymbol{x}) &= \sum_{n=1}^N \mathbb{E}_{P(\omega_n|(g,\lambda)^{\rm old})}\left[\omega_n\right]\delta(\boldsymbol{x}-\boldsymbol{x}_n) + \int_0^\infty \mathbb{E}_{P(\omega|(g,\lambda)^{\rm old})}\left[\omega\right]\tilde{\Lambda}(\boldsymbol{x},\omega)d\omega, \\ \tilde{B}(\boldsymbol{x}) &= \frac{1}{2}\sum_{n=1}^N \delta(\boldsymbol{x}-\boldsymbol{x}_n) - \frac{1}{2}\int_0^\infty \tilde{\Lambda}(\boldsymbol{x},\omega)d\omega. \end{aligned}$$

We have already tackled almost identical log–likelihood expressions in Section 3.1 (see Equation (20)). While for specific priors (with precision kernels given by differential operators) an exact treatment in terms of solutions of ODEs or PDEs is possible, we will again resort to the sparse GP approximation instead. The sparse version $U^s(\boldsymbol{g}_s,\lambda)$ is obtained by replacing $g(\boldsymbol{x}) \to \mathbb{E}_{P(g|\boldsymbol{g}_s)}\left[g(\boldsymbol{x})\right]$ in $U(g,\lambda)$. From this we obtain the sparse $\mathcal{Q}$–function as

$$\mathcal{Q}^s((\boldsymbol{g}_s,\lambda)|(\boldsymbol{g}_s,\lambda)^{\rm old}) \doteq U^s(\boldsymbol{g}_s,\lambda) + \ln p(\lambda) - \frac{1}{2}\boldsymbol{g}_s^\top K_s^{-1}\boldsymbol{g}_s. \qquad (31)$$

The function values $\boldsymbol{g}_s$ and the maximal intensity λ that maximise Equation (31) can be found analytically by solving

$$\frac{\partial Q^s}{\partial \boldsymbol{g}_s} = \mathbf{0} \text{ and } \frac{\partial Q^s}{\partial \lambda} = 0.$$

The final MAP estimate is obtained after convergence of the EM algorithm and the desired sparse MAP solution for $g(x)$ is given by (see Equation (27))

$$g_{MAP}(\boldsymbol{x}) = \boldsymbol{k}_s(\boldsymbol{x})^\top K_s^{-1} \boldsymbol{g}^s$$

As for the variational scheme, integrals over the space $\mathcal{X}$ are approximated by Monte–Carlo integration. An alternative derivation of the sparse MAP solution can be based on restricting the minimisation of (30) to functions which are linear combinations of kernels centred at the inducing points and using the definition of the RKHS norm (see (Rasmussen and Williams, 2006, chap. 6)).

Sparse Laplace posterior To complete the computation of the Laplace approximation, we need to evaluate the quadratic fluctuations around the MAP solution. We will also do this with the previously obtained sparse approximation. The idea is that from the converged MAP solution, we define a sparse likelihood of the Poisson model via the replacement

$$L^s(\boldsymbol{g}_s, \lambda) \doteq L(\mathcal{D}|\mathbb{E}_{P(g|\boldsymbol{g}_s)}[g], \lambda)$$

For this sparse likelihood it is easy to compute the Laplace posterior using second derivatives. Here, the change of variables $\rho = \ln \lambda$ will be made to ensure that $\lambda > 0$. This results in an effective log–normal density over the maximal intensity rate λ. While we do not address hyperparameter selection for the Laplace posterior in this work, a straightforward approach, as suggested by Flaxman et al. (2017), could be to use cross validation to optimise the kernel parameters while finding the MAP estimate or to use the Laplace approximation to approximate the evidence. As in the variational case the inducing point locations $\{\boldsymbol{x}_l\}_{l=1}^L$ will be on a regular grid over space $\mathcal{X}$.

Note that for the Laplace approximation, the augmentation scheme is only used to compute the MAP estimate in an efficient way. There are no further mean–field approximations involved. This also implies, that dependencies between $\boldsymbol{g}_s$ and λ are retained.

3.3. Predictive density

Both variational and Laplace approximation yield a posterior distribution q over $\boldsymbol{g}_s$ and λ. The GP approximation at any given points in $\mathcal{X}$ is given by

$$q(g(\boldsymbol{x})) = \int\int p(g(\boldsymbol{x})|\boldsymbol{g}_s) q(\boldsymbol{g}_s, \lambda)\, d\boldsymbol{g}_s d\lambda,$$

which for both methods results in a normal density. To find the posterior mean of the intensity function at a point $\boldsymbol{x} \in \mathcal{X}$ one needs to compute

$$\mathbb{E}_Q[\Lambda(\boldsymbol{x})] = \mathbb{E}_Q\left[\lambda \int_{-\infty}^{\infty} \sigma(g(\boldsymbol{x}))\right].$$

For variational and Laplace posterior the expectation over λ can be computed analytically, leaving the expectation over $g(\boldsymbol{x})$, which is computed numerically via quadrature methods. To evaluate the performance of inference results we are interested in computing the likelihood on test data $\mathcal{D}_{\text{test}}$, generated from the ground truth. We will consider two methods:

Sampling GPs g from the posterior we calculate the (log) mean of the test likelihood

$$\begin{aligned}\ell(\mathcal{D}_{\text{test}}) &= \ln \mathbb{E}_P \left[L(\mathcal{D}_{\text{test}}|\Lambda)|\mathcal{D}\right] \approx \ln \mathbb{E}_Q \left[L(\mathcal{D}_{\text{test}}|\Lambda)\right] \\ &= \ln \mathbb{E}_Q \left[\exp\left(-\int_{\mathcal{X}} \lambda\sigma(g(\boldsymbol{x}))d\boldsymbol{x}\right) \prod_{\boldsymbol{x}_n \in \mathcal{D}_{\text{test}}} \lambda\sigma(g(\boldsymbol{x}_n))\right]\end{aligned} \tag{32}$$

where the integral in the exponent is approximated by Monte–Carlo integration. The expectation is approximated by averaging over 2×10^3 samples from the inferred posterior Q of λ and g at the observations of $\mathcal{D}_{\text{test}}$ and the integration points.

Instead of sampling one can also obtain an analytic approximation for the log test likelihood in Equation (32) by a second order Taylor expansion around the mean of the obtained posterior. Applying this idea to the variational mean field posterior we get

$$\begin{aligned}\ell(\mathcal{D}_{\text{test}}) \approx{}& \ln L(\mathcal{D}_{\text{test}}|\Lambda_Q) + \frac{1}{2}\mathbb{E}_Q \left[(\boldsymbol{g}_s - \boldsymbol{\mu}_2^s)^\top \mathbf{H}_{\boldsymbol{g}_s}\big|_{\Lambda_Q} (\boldsymbol{g}_s - \boldsymbol{\mu}_2^s)\right] \\ &+ \frac{1}{2} H_\lambda\big|_{\Lambda_Q} \mathrm{Var}_Q(\lambda),\end{aligned} \tag{33}$$

where $\Lambda_Q(\boldsymbol{x}) = \mathbb{E}_Q[\lambda]\,\sigma(\mathbb{E}_Q[g(\boldsymbol{x})])$ and $\mathbf{H}_{\boldsymbol{g}_s}\big|_{\Lambda_Q}$, $H_\lambda\big|_{\Lambda_Q}$ are the second order derivative of the likelihood in Equation (1) with respect to $\boldsymbol{g}_s$ and λ at Λ_Q. While an approximation only involving the first term would neglect the uncertainties in the posterior (as done by John and Hensman (2018)), the second and third term take these into account.

4. Results

Generating data from the model To evaluate the two newly developed algorithms we generate data according to the sigmoidal Gaussian Cox process model

$$\begin{aligned} g \sim{}& \boldsymbol{p}_{\text{GP}}(\cdot|0, k), \\ \mathcal{D} \sim{}& \boldsymbol{p}_\Lambda(\cdot),\end{aligned}$$

where $\boldsymbol{p}_\Lambda(\cdot)$ is the Poisson process density over sets of point with $\Lambda(\boldsymbol{x}) = \lambda\sigma(g(\boldsymbol{x}))$ and $\boldsymbol{p}_{\text{GP}}(\cdot|0, k)$ is a GP density with mean 0 and covariance function k. As kernel we choose a squared exponential function

$$k(\boldsymbol{x}, \boldsymbol{x}') = \theta \prod_{i=1}^{d} \exp\left(-\frac{(x_i - x_i')^2}{2\nu_i^2}\right),$$

where the hyperparameters are scalar θ and length scales $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)^\top$. Sampling of the inhomogeneous Poisson process is done via *thinning* (Lewis and Shedler, 1979; Adams et al., 2009). We assume that hyperparameters are known for subsequent experiments with data sampled from the generative model.

Benchmarks for sigmoidal Gaussian Cox process inference We compare the proposed algorithms to two alternative inference methods for the sigmoidal Gaussian Cox process model. As an exact inference method we use the sampling approach of Adams et al. (2009)[4]. In terms of speed, a competitor is a different variational approach given by Hensman et al. (2015) who proposed to discretise space $\mathcal{X}$ in several regular bins with size Δ. Then the likelihood in Equation (1) is approximated by

$$L(\mathcal{D}|\lambda\sigma(g(\boldsymbol{x}))) \approx \prod_i p_{\mathrm{po}}(n_i|\lambda\sigma(g(\boldsymbol{x}_i))\Delta),$$

where p_{po} is the Poisson distribution conditioned on the mean parameter, $\boldsymbol{x}_i$ is the centre of bin i, and n_i the number of observations within this bin. Using a (sparse) Gaussian variational approximation the corresponding Kullback–Leibler divergence is minimised by gradient ascent to find the optimal posterior over the GP g and a point estimate for λ. This method was originally proposed for the log Cox-process ($\Lambda(\boldsymbol{x}) = e^{g(\boldsymbol{x})}$), but with the elegant GPflow package (Matthews et al., 2017) implementation of the scaled sigmoid link function is straightforward. It should be noted, that this method requires numerical integration over the sigmoid link function to evaluate the variational lower bound at every spatial bin and every gradient step, since it does not make use of our augmentation scheme (see Section 5 for discussion, how the proposed augmentation can be used for this model). We refer to this inference algorithm as 'variational Gauss'. To have fair comparison between the different methods, the inducing points for all algorithms (except for the sampler) are equal and the number of bins used to discretise the domain $\mathcal{X}$ for the variational Gauss algorithm is set equal to the number of integration points used for the MC integration in the variational mean field and the Laplace method.

Experiments on data from generative model As an illustrative example we sample a one dimensional Poisson process with the generative model and perform inference with the sampler (2×10^3 samples after 10^3 burn-in iterations), the mean field algorithm, the Laplace approximation and the variational Gauss. In Figure 1 **(a)**–**(d)** the different posterior mean intensity functions with their standard deviations are shown. For **(b)**–**(d)** 50 regularly spaced inducing points are used. For **(b)**–**(c)** 2×10^3 random integration points are drawn uniformly over the space $\mathcal{X}$, while for **(d)** $\mathcal{X}$ is discretised into the same number of bins. All algorithms recover the true intensity well. The mean field and the Laplace algorithm show smaller posterior variance compared to the sampler. The fastest inference result is obtained by the Laplace algorithm in 0.02 s, followed by the mean field (0.09), variational Gauss (80) and the sampler (1.8×10^3). The fast convergence of the Laplace and the variational mean field algorithm is illustrated in Figure 1 **(e)**, where objective functions of our two algorithms (minus the maximum they converged to) is shown as a function of run time. Both algorithms reach a plateau in only a few (~ 6) iterations. To compare performance in terms of log expected test likelihood ℓ_{test} (test sets $\mathcal{D}_{\mathrm{test}}$ sampled from the ground truth), we averaged results over ten independent data sets. The posterior of the sampler yields the highest value with 875.5, while variational ($\ell_{\mathrm{test}} = 686.2$, approximation by Equation (33) yields 686.5), variational Gauss (686.7) and Laplace (686.1) yield all similar results (see also Figure 4 **(a)**). The posterior density of the maximal intensity λ is shown in Figure 1 **(f)**.

[4] To increase efficiency, the GP values g are sampled by elliptical slice sampling (Murray et al., 2010).

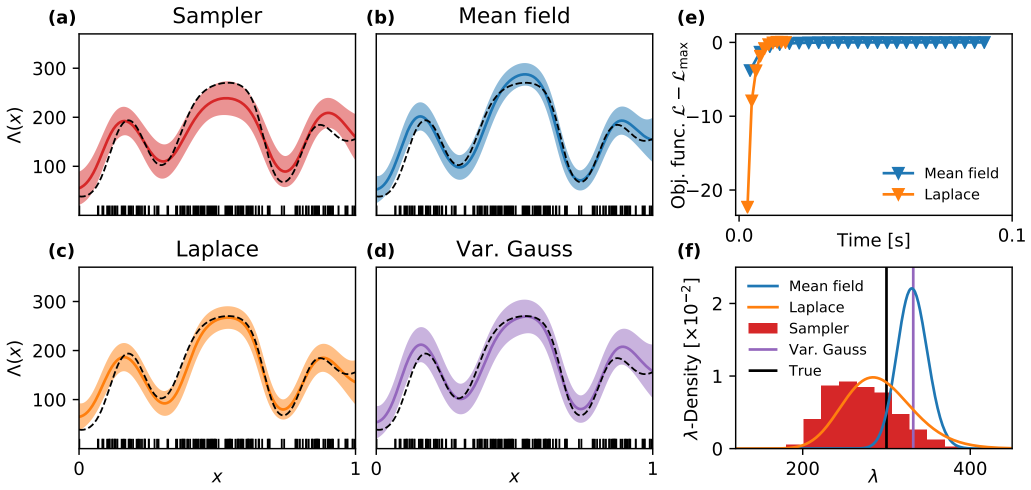

Figure 1: **Inference on 1D dataset. (a)–(d)** Inference result for sampler, mean field algorithm, Laplace approximation, and variational Gauss. Solid coloured lines denote the mean intensity function, shaded areas mean ± standard deviation, and dashed black lines the true rate functions. Vertical bars are observations $\mathcal{D}$. **(e)** Convergence of mean field and EM algorithm. Objective functions (Lower bound for mean–field and log likelihood for EM algorithm, shifted such that convergence is at 0) as function of run time (triangle marks one finished iteration of the respective algorithm). **(f)** Inferred posterior densities over the maximal intensity λ. Variational Gauss provides only a point estimate. Black vertical bar denotes the true λ.

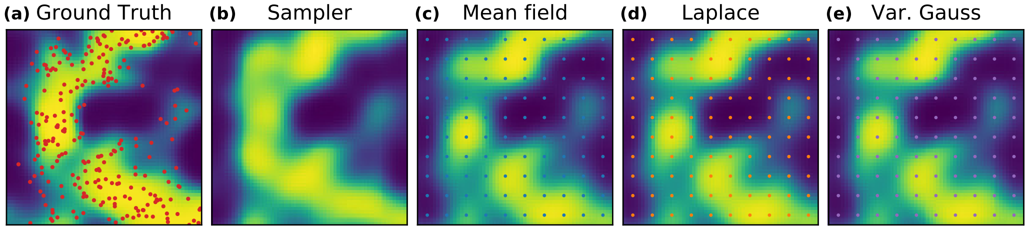

Figure 2: **Inference on 2D dataset.** **(a)** Ground truth intensity function $\Lambda(\boldsymbol{x})$ with observed dataset $\mathcal{D}$ (red dots).**(b)**–**(e)** Mean posterior intensity of the sampler, mean field algorithm, Laplace, and variational Gauss are shown. 100 inducing points on a regular grid (shown as coloured points) and 2500 integration points/bins are used.

In Figure 2 we show inference results for a two dimensional Cox process example. 10×10 inducing points and 2500 integration points/bins are used for mean field, Laplace and variational Gauss algorithm. The posterior mean of sampler **(b)**, of the mean field **(c)**, of the Laplace **(d)** and of the variational Gauss algorithm **(e)** recover the true intensity rate $\Lambda(\boldsymbol{x})$ **(a)** well.

To evaluate the role of the number of inducing points and number of integration points we generate 10 test sets $\mathcal{D}_{\text{test}}$ from a process with the same intensity as in Figure 2**(a)**. We evaluate the log expected likelihood (Equation (32)) on these test sets and compute the average. The result is shown for different numbers of inducing points (Figure 3**(a)** with 2500 integration points) and different numbers of integration points (Figure 3**(b)** with 10×10 inducing points). To account for randomness of integration points the fitting is repeated five times and the shaded area is between the minimum and maximum obtained by these fits. For all approximate algorithms the log predictive test likelihood saturates already for few inducing points (≈ 49 (7×7)) of the sparse GP. However, as expected, the inference approximations are slightly inferior to the sampler. The log expected test likelihood is hardly affected by the number of integration points as shown in Figure 3 **(b)**. Also the approximated test likelihood for the mean field algorithm in Equation (33) yields good estimates of the sampled value (dashed line in **(a)** and **(b)**). In terms of runtime (Figure 4 **(c)**–**(d)**) the mean field algorithm and the Laplace approximation are superior by more than one order of magnitude to the variational Gauss algorithm for this particular example. Difference increases with increasing number of inducing points.

In Figure 4 the four algorithms are compared on five different data sets sampled from the generative model. As we observed for the previous examples the three different approximating algorithms yield qualitatively similar performance in terms of log test likelihood ℓ_{test}, but the sampler is superior. Again the approximated test likelihood in Equation (33) (blue star) provides good estimate of the sampled value. In addition we provide the approximated root mean squared error (RMSE, evaluated on a fine grid and normalised by maximal intensity λ) between inferred mean and ground truth. In terms of run time the mean field and Laplace algorithm are by at least on order of magnitude faster than the vari-

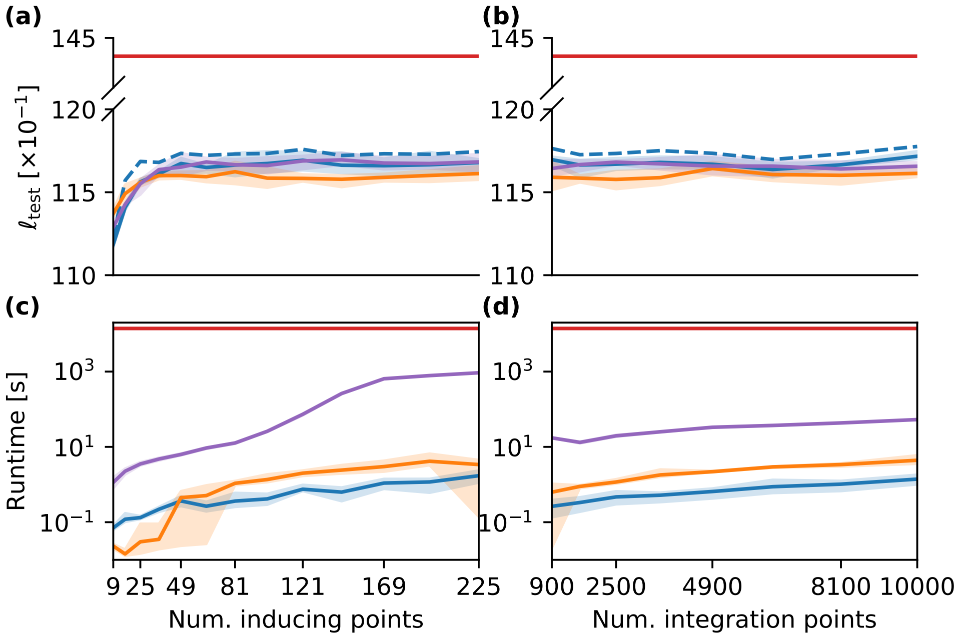

Figure 3: **Evaluation of inference. (a)** The log expected predictive likelihood averaged over ten test sets as a function of the number of inducing points. Number of integration points/bins is fixed to 2500. Results for sampler in (red), mean field (blue), Laplace (orange), and variational Gauss (purple) algorithm. Solid line denotes mean over five fits (same data), and shaded area denotes min. and max. result. Dashed blue line shows the approximated log expected predictive likelihood for the mean field algorithm. **(b)** Same as (a), but as function of number of integration points. Number of inducing points is fixed to 10 × 10. Below: Run time of the different algorithms as function of number of inducing points **(c)** and number of integration points **(d)**. Data are the same as in Figure 2.

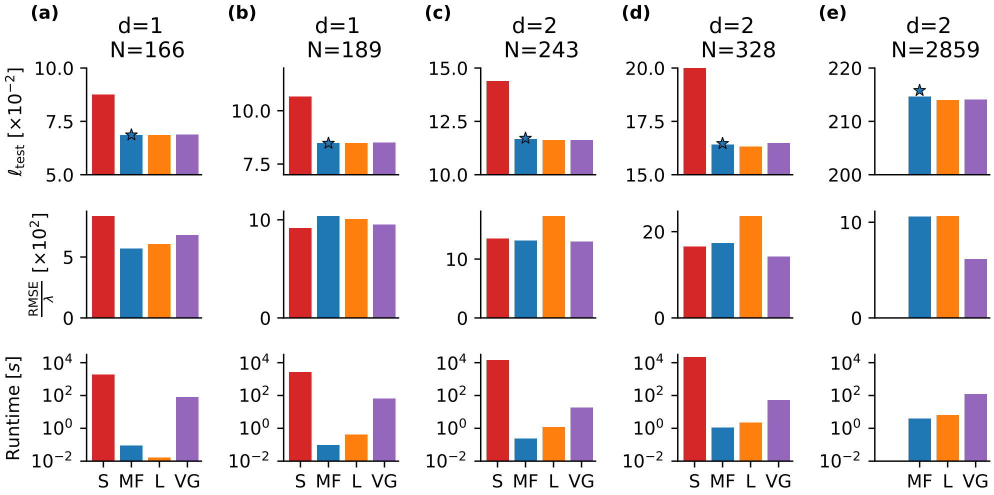

Figure 4: **Performance on different artificial datasets.** The sampler (S), the mean field algorithm (MF), the Laplace (L), and variational Gauss (VG) are compared on five different datasets with d–dimensions and N observations (one column corresponds to one dataset). Top row: Log expected test likelihood of the different inference results. The star denotes the approximated test likelihood of the variational algorithm. Center row: The approximated root mean squared error (normalised by true maximal intensity rate λ). Bottom row: Run time in seconds. The dataset **(e)** is intractable for the sampler due to the many observations. Data in Figure 1 and 2 correspond to **(a)** and **(c)**.

ational Gauss algorithm. In general, the mean–field algorithm seems to be slightly faster than the Laplace.

General data sets and comparison to the approach of Lloyd et al. Next, we test our variational mean field algorithm on data sets not coming from the generative model. On such data sets we do not know, whether our model provides a good prior. As discussed previously an alternative model was proposed by Lloyd et al. (2015) making use of the link function $\Lambda(\boldsymbol{x}) = g^2(\boldsymbol{x})$. While the sigmoidal Gaussian Cox process with the proposed augmentation scheme has analytic updates for the variational posterior, in case of the squared Gaussian Cox process the likelihood integral can be solved analytically and does not need to be sampled (if the kernel is a squared exponential and the domain is rectangular). Both algorithms rely on the sparse GP approximation. To compare the two methods empirically first we consider one dimensional data generated using a known intensity function. We choose $\Lambda(x) = 2\exp(-x/15) + \exp(-(x-25)^2/100)$ on an interval $[0, 50]$ already proposed by Adams et al. (2009). We generate three training and test sets, where we scale this rate function by factors of 1, 10, and 100 and fit the sigmoidal and squared Gaussian Cox

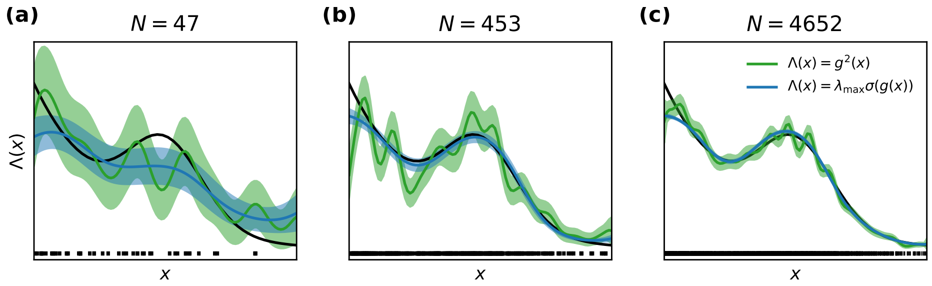

Figure 5: **1D example.** Observations (black bars) are sampled from the same function (black line) scaled by **(a)** 1, **(b)** 10, and **(c)** 100. Blue and green line show the mean posterior of the sigmoidal and squared Gaussian Cox process, respectively. Shaded area denotes mean ± standard deviation.

	$\Lambda(x) = \lambda_{\max}\sigma(g(x))$			$\Lambda(x) = g^2(x)$		
N	Runtime [s]	RMSE	ℓ_{test}	Runtime [s]	RMSE	ℓ_{test}
47	0.27 ± 0.30	0.24 ± 0.02	-43.43 ± 0.42	0.41 ± 0.05	0.24	-44.26 ± 0.09
453	0.50 ± 0.04	0.97 ± 0.13	720.81 ± 0.28	0.23 ± 0.05	2.11	710.43 ± 1.38
4652	0.41 ± 0.01	7.68 ± 0.75	17497.31 ± 2.13	0.79 ± 0.09	8.16	17496.75 ± 1.65

Table 1: **Benchmarks for Figure 5** The mean and standard deviation of runtime, RMSE, and log expected test likelihood for Figure 5**(a)**–**(c)** obtained from 5 fits. Note that the RMSE for $\Lambda(\boldsymbol{x}) = g^2(\boldsymbol{x})$ has no standard deviation, because the inference algorithm is deterministic.

process with their corresponding variational algorithm to each training set[5]. The number of inducing points is 40 in this example. For our variational mean field algorithm we used 5000 integration points. The posterior intensity $\Lambda(\boldsymbol{x})$ for the three data sets can be seen in Figure 5. The model with the sigmoidal link function infers smoother posterior functions with smaller variance compared to the posterior with the squared link function. For data sets shown in Figure 5 we run the fits five times and report mean and standard deviation of runtime, RMSE and log expected test likelihood ℓ_{test} in Table 1. Run times of the two algorithms are comparable, where for the intermediate data set the algorithm with the squared link function is faster while for the largest data set the one with the sigmoidal link function converges first. RMSE and ℓ_{test} are also comparable except for the intermediate data set, where the sigmoidal model is the superior one.

Next we deal with two real world two dimensional data sets for comparison. The first one is neuronal data, where spiking activity was recorded from a mouse, that was freely moving in an arena (For The Biology Of Memory and Sargolini, 2014; Sargolini et al., 2006). Here we consider as data $\mathcal{D}$ the position of the mouse when the recorded cell fired and the observations are randomly assigned to either training or test set. In Figure 6 **(a)**

[5] We thank Chris Lloyd and Tom Gunter for providing the code for inferring the variational posterior of the squared Gaussian Cox process.

the observations in the training set ($N = 583$) are shown. In Figure 6 **(b)** and **(c)** the variational posterior's mean intensity $\Lambda(\boldsymbol{x})$ is shown obtained for the sigmoidal and the squared link function, respectively, inferred with a regular grid of 20×20 inducing points. As in Figure 5 we see that the sigmoidal posterior is the smoother one. The major difference between the two algorithms (apart from the link function) is the fact that for the sigmoidal model we are required to sample an interval over the space. We investigate the effect of the number of integration points in terms of runtime[6] and log expected test likelihood in Figure 6 **(d)**. First, we observe regardless of the number of integration points that the variational posterior of the squared link function yields the superior expected test likelihood. For the sigmoidal model the test likelihood does not improve significantly with more integration points. Runtimes of both algorithms are comparable, when 5000 integration points are chosen. A speed up for our mean field algorithm is achieved by first fitting the model with 1000 integration points and once converged, redrawing the desired number of integration points and rerun the algorithm (dotted line in Figure 6**(d)**). This method allows for a significant speed up without loss in terms of test likelihood ℓ_{test}. The variational mean-field algorithm with the sigmoid link function is faster with up to 5000 integration points and equally fast with 10000 integration points.

As second data set we consider the Porto taxi data set (Moreira-Matias et al., 2013). This data contains trajectories of taxi travels from the years 2013/14 in the city of Porto. As John and Hensman (2018) we consider the pick-ups as observations of a Poisson process[7]. We consider 20000 taxi rides randomly split into training and test set ($N = 10017$ and $N = 9983$, respectively). The training set is shown in Figure 6**(e)**. Inducing points are positioned on a regular grid of 20×20. The variational posterior mean of the respective intensity is shown in Figure 6 **(f)** and **(g)**. With as many data points as in these data the differences between the two models are more subtle as compared to **(b)** and **(c)**. In terms of test likelihood ℓ_{test} the variational posterior of the sigmoidal model (with ≥ 2000 integration points) outperforms the model with squared link function (Figure 6 **(h)**). For similar test likelihoods ℓ_{test} our variational algorithm is $\sim 2\times$ faster than the variational posterior with squared link function. The results show that the choice of number of integration points reduces to the question of speed vs accuracy trade–off. As for the previous data set, the strategy of first fitting the posterior with 1000 integration points and then with the desired number of integration points (dotted line) proves that we can get a significant speed up without loosing predictive power.

5. Discussion and Outlook

Using a combination of two known variable augmentation methods, we derive a conjugate representation for the posterior measure of a sigmoidal Gaussian Cox process. The approximation of the augmented posterior by a simple mean field factorisation yields an efficient variational algorithm. The rationale behind this method is that the variational updates in the conjugate model are explicit and analytical and do not require (black–box) gradient

[6] Note, that - in contrast to Figures 3 and 4 - the runtime is displayed on linear scale, meaning both algorithms are of same order of magnitude.

[7] As John and Hensman (2018) report some regions to be highly peaked we consider only pickups happening within the coordinates $(41.147, -8.58)$ and $(41.18, -8.65)$ in order to exclude those regions.

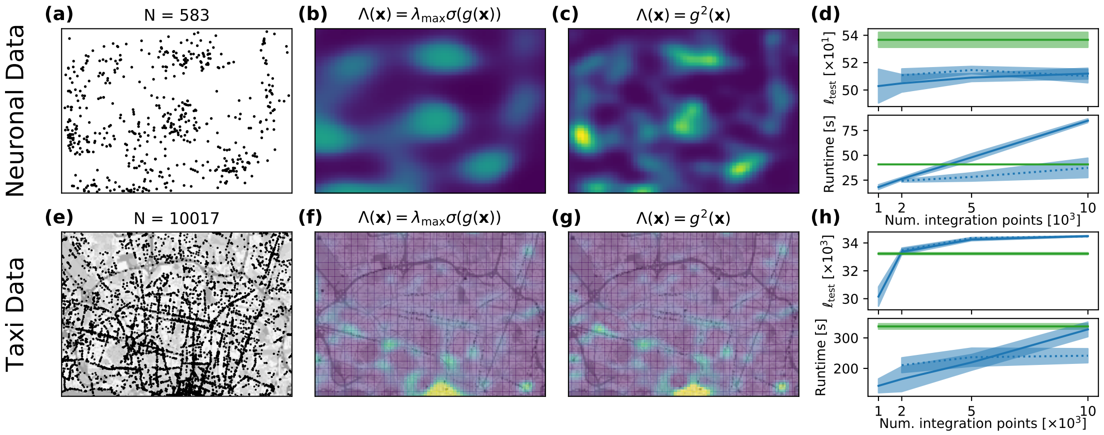

Figure 6: **Fits to real world data sets.** **(a)** Position of the mouse while the recorded neuron spiked. **(b)** Posterior mean obtained by the variational mean–field algorithm for the sigmoidal Gaussian Cox process. **(c)** Same as in (b) for the variational approximation of the squared Gaussian Cox process. **(d)** Log expected test–likelihood ℓ_{test} and runtime as function of number of integration points for both algorithms. The dotted line is obtained by first fitting the sigmoidal model with 1000 integration points and then with the number that is indicated on the x-axis. Shaded area is mean ± standard deviation obtained in 5 repeated fits. **(e)–(h)** Same as (a)–(d), but for a data set, where the observations are positions of taxi pick–ups in the city of Porto.

descent methods. In fact, a comparison with a different variational algorithm for the same model - not based on augmentation, but on direct approximation of the posterior with a Gaussian - shows that the qualities of inference for both approaches are similar, while the mean field algorithm is at least one order of magnitude faster. We use the same variable augmentation method for computation of the MAP estimate for the (unaugmented) posterior by a fast EM algorithm. This is finally applied to the calculation of Laplace's approximation. Both methods yield an explicit result for the approximate GP posterior. Since the corresponding effective likelihood contains a continuum of the GP latent variables, the exact computations of means and marginal variances would require the inversion of a linear operator instead of a simpler matrix inverse. While for specific priors, this problem could be solved by PDE or ODE methods, we resort to a well known sparse GP approach with inducing points in this paper. We can apply this to arbitrary kernels but need to solve spatial integrals over the domain. These can be (at least for moderate dimensionality) well approximated by simple Monte Carlo integration. Advantage of this approach is, that one is not limited to rectangular domains. The only requirement is that the volume $|\mathcal{X}|$ is known. An alternative Poisson model for which similar spatial integrals can be performed analytically (Lloyd et al., 2015) within the sparse GP approximation (limited to squared exponential kernels and rectangular domains) is based on a quadratic link function (Lloyd et al., 2015; Flaxman et al., 2017; John and Hensman, 2018). We compare our variational algorithm with the variational algorithm of Lloyd et al. (2015) on different data sets and observe that both algorithms act on the same order of magnitude in terms of runtime (with slight advantages for our variational mean field algorithm). As expected, we show that whether one or the other model is better in predictive power is highly data dependent.

As an alternative to the Monte Carlo integration in our approach we could avoid the infinite dimensionality of the latent GP from the beginning by working with a binning scheme for the Poisson observations as in Hensman et al. (2015). It would be straightforward to adopt our augmentation method to this case. The resulting Poisson likelihoods would then be augmented by pairs of Poisson and Pólya–Gamma variables (see Donner and Opper (2017)) for each bin. This approach could be favourable when the number of observed data points becomes very large, because the discretisation method does not scale with the number data points but with the resolution of discretisation. However, we do expect, that any approach based on either spatial discretisation or on the sparse, inducing point method would become problematic for large or high dimensional domains $\mathcal{X}$. Alternative methods based on spectral representations of kernels (Knollmüller et al., 2017; John and Hensman, 2018) are promising for tackling those problems.

It will be interesting to apply the variable augmentation method to other Bayesian models with the sigmoid link function. For example, the inherent boundedness of the resulting intensity can be crucial for point processes such as the nonlinear *Hawkes process* (Hawkes, 1971) which is widely used for modelling stock market data (Embrechts et al., 2011) or seismic activity (Ogata, 1998). For other point process models the sigmoid function appears naturally. We mention the kinetic Ising model, a Markov jump process (Donner and Opper, 2017) which was originally introduced to model the dynamics of classical spin systems in physics. More recently it was used to model the joint activity of neurons (Dunn et al., 2015). Finally, a Gaussian process density model introduced by (Murray et al., 2009) can be treated by the augmentations developed in this work (Donner and Opper, 2018).

Acknowledgments

CD was supported by the Deutsche Forschungsgemeinschaft (GRK1589/2) and partially funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1294 “Data Assimilation”, Project (A06) “Approximative Bayesian inference and model selection for stochastic differential equations (SDEs)”.

Appendix A. Poisson processes

In this paragraph we briefly summarise those properties of a Poisson process, which are relevant for this work. For a thorough and more complete description we recommend the concise book by Kingman (1993), particularly chapter 3 and 5.

We consider a general space $\mathcal{Z}$ and a countable subset $\Pi_{\mathcal{Z}} = \{\boldsymbol{z}; \boldsymbol{z} \in \mathcal{Z}\}$.

Definition of a Poisson process A random countable subset $\Pi_{\mathcal{Z}} \subset \mathcal{Z}$ is a Poisson process on $\mathcal{Z}$, if

i) for any sequence of disjoint subsets $\{\mathcal{Z}_k \subset \mathcal{Z}\}_{k=1}^K$ the cardinality of the union $N(\mathcal{Z}_k) \doteq |\{\Pi_{\mathcal{Z}} \cap \mathcal{Z}_k\}|$ is independent of $N(\mathcal{Z}_l)$ for all $l \neq k$.

ii) $N(\mathcal{Z}_k)$ is Poisson distributed with mean $\int_{\mathcal{Z}_k} \Lambda(\boldsymbol{z}) d\boldsymbol{z}$, and mean measure $\Lambda(\boldsymbol{z}) : \mathcal{X} \to \mathbb{R}^+$.

If the mean measure is constant ($\Lambda(\boldsymbol{z}) = \Lambda$) the Poisson process is *homogeneous*, and *inhomogeneous* otherwise.

Campbell's Theorem Let $\Pi_{\mathcal{Z}}$ be a Poisson process on $\mathcal{Z}$ with mean measure $\Lambda(\boldsymbol{z})$. Furthermore, we define a function $h(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ and the sum

$$H(\Pi_{\mathcal{Z}}) = \sum_{\boldsymbol{z} \in \Pi_{\mathcal{Z}}} h(\boldsymbol{z}).$$

If $\Lambda(\boldsymbol{z}) < \infty$ for $\boldsymbol{z} \ \in \mathcal{Z}$, then

$$\mathbb{E}_{P_\Lambda}\left[e^{\xi H(\Pi_{\mathcal{Z}})}\right] = \exp\left\{\int_{\mathcal{Z}} \left(e^{\xi h(\boldsymbol{z})} - 1\right) \Lambda(\boldsymbol{z}) d\boldsymbol{z}\right\}, \tag{36}$$

for any $\xi \in \mathbb{C}$, such that the integral converges. P_Λ is the probability measure of a Poisson process with intensity $\Lambda(\boldsymbol{z})$. Mean and variance are obtained as

$$\mathbb{E}_{P_\Lambda}\left[H(\Pi_{\mathcal{Z}})\right] = \int_{\mathcal{Z}} h(\boldsymbol{z}) \Lambda(\boldsymbol{z}) d\boldsymbol{z},$$
$$\mathrm{Var}_{P_\Lambda}\left[H(\Pi_{\mathcal{Z}})\right] = \int_{\mathcal{Z}} [h(\boldsymbol{z})]^2 \Lambda(\boldsymbol{z}) d\boldsymbol{z}.$$

Note, that Equation (36) defines the *characteristic functional* of a Poisson process.

Marked Poisson process Let $\Pi_{\mathcal{Z}} = \{\boldsymbol{z}_n\}_{n=1}^N$ a Poisson process on $\mathcal{Z}$ with intensity $\Lambda(\boldsymbol{z})$. Then $\Pi_{\hat{\mathcal{Z}}} = \{(\boldsymbol{z}_n, \boldsymbol{m}_n)\}_{n=1}^N$ is again a Poisson process on the product space $\hat{\mathcal{Z}} = \mathcal{Z} \times \mathcal{M}$, if $\boldsymbol{m}_n \sim p(\boldsymbol{m}_n|\boldsymbol{z}_n)$ is drawn independently at each $\boldsymbol{z}_n$. The $\boldsymbol{m}_n \in \mathcal{M}$ are the so–called 'marks', and the resulting Process is a *marked Poisson process* with intensity

$$\Lambda(\boldsymbol{z}, \boldsymbol{m}) = \Lambda(\boldsymbol{z}) p(\boldsymbol{m}|\boldsymbol{z}).$$

It is straightforward to extend Campbell's theorem and to show that the characteristic functional of such a process is

$$\mathbb{E}_{P_\Lambda}\left[e^{\xi H(\Pi_{\hat{\mathcal{Z}}})}\right] = \exp\left\{\int_{\hat{\mathcal{Z}}} \left(e^{\xi h(\boldsymbol{z}, \boldsymbol{m})} - 1\right) \Lambda(\boldsymbol{z}, \boldsymbol{m}) \, d\boldsymbol{m} d\boldsymbol{z}\right\}, \tag{37}$$

with $h(\boldsymbol{z}, \boldsymbol{m}) : \hat{\mathcal{Z}} \to \mathbb{R}$ and $H(\Pi_{\hat{\mathcal{Z}}}) = \sum_{(\boldsymbol{z}, \boldsymbol{m}) \in \Pi_{\hat{\mathcal{Z}}}} h(\boldsymbol{z}, \boldsymbol{m})$.

Appendix B. The Pólya-Gamma density

The Pólya-Gamma density (Polson et al., 2013) has the useful property, that it allows to represent the inverse hyperbolic cosine by an infinite Gaussian mixture as

$$\cosh^{-b}(c/2) = \int_0^\infty \exp\left(-\frac{c^2}{2}\omega\right) p_{\mathrm{PG}}(\omega|b,0)d\omega,$$

with parameter $b > 0$. Furthermore, one can define a *tilted Pólya-Gamma density* as

$$p_{\mathrm{PG}}(\omega|b,c) = \frac{\exp\left(-\frac{c^2}{2}\omega\right)}{\cosh^{-b}(c/2)} p_{\mathrm{PG}}(\omega|b,0).$$

From those two equations the moment generating function can be obtained from the basic definition, being

$$\int_0^\infty e^{\xi\omega} p_{\mathrm{PG}}(\omega|b,c)d\omega = \frac{\cosh^b(c/2)}{\cosh^b\left(\sqrt{\frac{c^2/2-\xi}{2}}\right)},$$

and differentiating with respect to ξ at $\xi = 0$ yields the first moment

$$\mathbb{E}_{p_{\mathrm{PG}}}[\omega] = \frac{b}{2c}\tanh(c/2).$$

Appendix C. Variational inference for stochastic processes

Densities for random processes A stochastic process X with probability measure $P(X)$ often has no density with respect to Lebesgue measure, since X can be an infinite dimensional object such as a function for the case of a Gaussian process. However, one can define densities with respect to another (reference) measure $R(X)$ written as

$$\boldsymbol{p}(X) = \frac{dP}{dR}(X), \tag{38}$$

if $R(X)$ is absolutely continuous with respect to $P(X)$ (if $R(X) = 0$ then $P(X) = 0$). Using such a density, expectations are

$$\mathbb{E}_P[f(X)] = \int f(X)dP(X) = \int f(x)\boldsymbol{p}(x)dR(X) = \mathbb{E}_R[f(x)\boldsymbol{p}(x)].$$

The density in Equation (38) is known as the *Radon–Nikodým derivative* of R with respect to P (Konstantopoulos et al., 2011).

Poisson process density As specific example consider the prior density of the Poisson process in Equation (9), which is defined with respect to a reference measure

$$\boldsymbol{p}_\Lambda(\Pi_{\mathcal{Z}}) = \frac{dP_\Lambda}{dP_{\Lambda_0}}(\Pi_{\mathcal{Z}}) = \exp\left(-\int_{\mathcal{Z}}(\Lambda(\boldsymbol{z}) - \Lambda_0(\boldsymbol{z}))d\boldsymbol{z}\right)\prod_{\boldsymbol{z}_n\in\Pi_{\mathcal{Z}}}\frac{\Lambda(\boldsymbol{z}_n)}{\Lambda_0(\boldsymbol{z}_n)},$$

where P_{Λ_0} is the probability measure with intensity Λ_0 and the expectation is defined as

$$\mathbb{E}_{P_\Lambda}\left[\sum_{\boldsymbol{z}_n\in\Pi_{\mathcal{Z}}} u(\boldsymbol{z}_n)\right] = \mathbb{E}_{P_{\Lambda_0}}\left[\boldsymbol{p}_\Lambda(\Pi_{\mathcal{Z}}) \sum_{\boldsymbol{z}_n\in\Pi_{\mathcal{Z}}} u(\boldsymbol{z}_n)\right]. \tag{39}$$

Calculating the expectation of $e^{\xi H(\Pi_{\mathcal{Z}})}$ with Equation (39) we identify the characteristic function of a Poisson process (see Equation (37)) with intensity $\Lambda(\boldsymbol{z})$.

Kullback-Leibler divergence Using these densities we can express the Kullback-Leibler divergence between two probability measures.

The KL–divergence between $\boldsymbol{q}(X)$ and $\boldsymbol{p}(X)$ is defined as

$$\mathrm{D_{KL}}(Q\|P) = \mathbb{E}_Q\left[\log\frac{dQ}{dP}(X)\right] = \int \log\frac{\boldsymbol{q}(X)}{\boldsymbol{p}(X)}dQ(X),$$

where

$$\boldsymbol{q}(X) = \frac{dQ}{dR}(X),$$

and where $R(X)$ also is absolutely continuous to $Q(X)$. The KL–divergence does not depend on the reference measure $R(X)$.

Appendix D. The posterior point process is a marked Poisson process

Here we prove that the optimal variational posterior point process in Equation (18) again is a Poisson process using Campbell's theorem. As posterior process in Equation (18) one gets

$$\boldsymbol{q}(\Pi_{\mathcal{Z}}) = \frac{dQ}{dP_\lambda}(\Pi_{\mathcal{Z}}) = \frac{\prod_{\boldsymbol{z}_m\in\Pi_{\mathcal{Z}}} e^{f(\boldsymbol{z}_m)}}{\mathbb{E}_{P_\lambda}\left[\prod_{\boldsymbol{z}_m\in\Pi_{\mathcal{Z}}} e^{f(\boldsymbol{z}_m)}\right]} = \frac{\prod_{\boldsymbol{z}_m\in\Pi_{\mathcal{Z}}} e^{f(\boldsymbol{z}_m)}}{\exp\left(\int_{\mathcal{Z}}(e^{f(\boldsymbol{z})}-1)\lambda(\boldsymbol{z})d\boldsymbol{z}\right)},$$

where $\Pi_{\mathcal{Z}}$ is some random set of points on space $\mathcal{Z}$ and P_λ is a random Poisson measure with intensity $\lambda(\boldsymbol{z})$. To proof, that the resulting point process density $\boldsymbol{q}(\Pi_{\mathcal{Z}})$ is again a Poisson process we calculate the characteristic functional for some arbitrary function $h : \mathcal{Z} \to \mathbb{R}$

$$\begin{aligned}
\mathbb{E}_Q\left[\prod_{\boldsymbol{z}_m\in\Pi_{\mathcal{Z}}} e^{h(\boldsymbol{z}_m)}\right] =& \frac{\mathbb{E}_{P_\lambda}\left[\prod_{\boldsymbol{z}_m\in\Pi_{\mathcal{Z}}} e^{h(\boldsymbol{z}_m)+f(\boldsymbol{z}_m)}\right]}{\exp\left(\int_{\mathcal{Z}}(e^{f(\boldsymbol{z})}-1)\lambda(\boldsymbol{z})d\boldsymbol{z}\right)}\\
=& \frac{\exp\left(\int_{\mathcal{Z}}(e^{h(\boldsymbol{z})+f(\boldsymbol{z})}-1)\lambda(\boldsymbol{z})d\boldsymbol{z}\right)}{\exp\left(\int_{\mathcal{Z}}(e^{f(\boldsymbol{z})}-1)\lambda(\boldsymbol{z})d\boldsymbol{z}\right)}\\
=& \exp\left(\int_{\mathcal{Z}}(e^{h(\boldsymbol{z})}-1)e^{f(\boldsymbol{z})}\lambda(\boldsymbol{z})d\boldsymbol{z}\right)\\
=& \exp\left(\int_{\mathcal{Z}}(e^{h(\boldsymbol{z})}-1)\Lambda_Q(\boldsymbol{z})d\boldsymbol{z}\right).
\end{aligned}$$

We identify the last row as the generating functional of a Poisson process (37) with $\xi = 1$. The intensity of the process is $\Lambda_Q(\boldsymbol{z}) = e^{f(\boldsymbol{z})}\lambda(\boldsymbol{z})$. With the fact that a Poisson process is uniquely characterised by its generating function (Kingman, 1993, chap. 3), the proof is complete.

Appendix E. Sparse Gaussian process approximation

To solve the inference problem for the function g, we define a sparse GP, using the same prior P, but by an effective likelihood which depends on a finite set of function values $\boldsymbol{g}_s = (g_1, \dots, g_L)^\top$ only. Hence, we get

$$\frac{dQ_2^s}{dP}(g) = \boldsymbol{q}_2^s(\boldsymbol{g}_s) \tag{40}$$

and the sparse posterior measure is

$$dQ_2^s(g) = \boldsymbol{q}_2^s(\boldsymbol{g}_s)dP(g) = dP(g|\boldsymbol{g}_s) \times \boldsymbol{q}_2^s(\boldsymbol{g}_s)dP(\boldsymbol{g}_s),$$

where the last equality holds true, since Equation (40) only depends on $\boldsymbol{g}_s$. The KL–divergence between the full posterior density

$$\boldsymbol{q}_2(g) = \frac{dQ_2}{dP}(g) = \frac{e^{U(g)}}{\mathbb{E}_P\left[e^{U(g)}\right]}$$

and the sparse one $\boldsymbol{q}_2^s(\boldsymbol{g}_s)$ is given by

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(Q_2^s \| Q_2) &= \mathbb{E}_{Q_2^s}\left[\log \frac{\boldsymbol{q}_2^s(\boldsymbol{g}_s)}{\boldsymbol{q}_2(g)}\right] = \mathbb{E}_{P(\boldsymbol{g}_s)}\left[\boldsymbol{q}_2^s(\boldsymbol{g}_s)\mathbb{E}_{P(g|\boldsymbol{g}_s)}\left[\log \frac{\boldsymbol{q}_2^s(\boldsymbol{g}_s)}{e^{U(g)}}\right]\right] + \text{const.} \\
&= \mathbb{E}_{P(\boldsymbol{g}_s)}\left[\boldsymbol{q}_2^s(\boldsymbol{g}_s) \log \frac{\boldsymbol{q}_2^s(\boldsymbol{g}_s)}{e^{\mathbb{E}_{P(g|\boldsymbol{g}_s)}[U(g)]}}\right] + \text{const.}
\end{aligned}$$

From this we derive directly the posterior density for the sparse GP

$$\boldsymbol{q}_2^s(g) \propto e^{U^s(\boldsymbol{g}_s)},$$

with the sparse log–likelihood

$$U^s(\boldsymbol{g}_s) = \mathbb{E}_{P(g|\boldsymbol{g}_s)}\left[U(g)\right] = \int U(g)dP(g|\boldsymbol{g}_s).$$

Appendix F. Lower bound & hyperparameter optimization

The lower bound in Equation (12) is given by

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{q}) =& \mathbb{E}_Q\left[\log \frac{L(\mathcal{D},\boldsymbol{\omega}_N,\Pi_{\hat{\mathcal{X}}}|g,\lambda)}{\boldsymbol{q}_1(\boldsymbol{\omega}_N)\boldsymbol{q}_1(\Pi_{\hat{\mathcal{X}}})\boldsymbol{q}_2^s(g)\boldsymbol{q}_2(\lambda)}\right] \\
=& \int_{\hat{\mathcal{X}}} \left(\mathbb{E}_Q\left[f(\omega,-g(\boldsymbol{x}))\right] - \mathbb{E}_Q\left[\log\Lambda_1\right] + \mathbb{E}_Q\left[\log\lambda\right] + 1\right)\Lambda_1(\boldsymbol{x},\omega)d\boldsymbol{x}d\omega \\
& - \int_{\hat{\mathcal{X}}} \Lambda_1(\boldsymbol{x},\omega)d\boldsymbol{x}d\omega \\
& + \sum_{n=1}^N \left(\mathbb{E}_Q\left[f(\omega_n,g_n)\right] + \mathbb{E}_Q\left[\log\lambda\right] - \cosh\left(\frac{c_1^{(n)}}{2}\right) + \frac{\left(c_1^{(n)}\right)^2}{2}\mathbb{E}_Q\left[\omega_n\right]\right) \\
& - \frac{1}{2} trace(K_s^{-1}(\Sigma_2^s + \boldsymbol{\mu}_2^s(\boldsymbol{\mu}_2^s)^\top)) - \frac{1}{2}\log\det(2\pi K_s) + \frac{1}{2}\log\det(2\pi e\Sigma_2^s) \\
& + \alpha_0\log\beta_0 - \log(\Gamma(\alpha_0)) + (\alpha_0 - 1)\mathbb{E}_Q\left[\log\lambda\right] - \beta_0\mathbb{E}_Q\left[\lambda\right] \\
& + \alpha_2 - \log\beta_2 + \log\Gamma(\alpha_2) + (1-\alpha_2)\psi(\alpha_2).
\end{aligned}
$$

To optimise the covariance kernel parameters $\boldsymbol{\theta}$ we differentiate the lower bound with respect to these parameters and perform then gradient ascent. The gradient for one specific parameter θ is given by

$$
\begin{aligned}
\frac{\partial\mathcal{L}(\boldsymbol{q})}{\partial\theta} =& \int_{\hat{\mathcal{X}}} \frac{\partial\mathbb{E}_Q\left[f(\omega,-g(\boldsymbol{x}))\right]}{\partial\theta}\Lambda_1(\boldsymbol{x},\omega)d\boldsymbol{x}d\omega + \sum_{n=1}^N \frac{\partial\mathbb{E}_Q\left[f(\omega_n,g(\boldsymbol{x}_n))\right]}{\partial\theta} \\
& - \frac{1}{2}\frac{trace(K_s^{-1}(\Sigma_2^s + \boldsymbol{\mu}_2^s(\boldsymbol{\mu}_2^s)^\top))}{\partial\theta} - \frac{1}{2}\frac{\partial\log\det(2\pi K_s)}{\partial\theta} \\
=& \int_{\hat{\mathcal{X}}} \frac{\partial\mathbb{E}_Q\left[f(\omega,-g(\boldsymbol{x}))\right]}{\partial\theta}\Lambda_1(\boldsymbol{x},\omega)d\boldsymbol{x}d\omega + \sum_{n=1}^N \frac{\partial\mathbb{E}_Q\left[f(\omega_n,g(\boldsymbol{x}_n))\right]}{\partial\theta} \\
& + \frac{1}{2} trace\left(K_s^{-1}\frac{\partial K_s}{\partial\theta}K_s^{-1}(\Sigma_2^s + \boldsymbol{\mu}_2^s(\boldsymbol{\mu}_2^s)^\top)\right) \\
& - \frac{1}{2} trace\left(K_s^{-1}\frac{\partial K_s}{\partial\theta}\right).
\end{aligned}
$$

The derivatives of function $\mathbb{E}_Q\left[f(\omega,g(\boldsymbol{x}))\right]$ are

$$
\frac{\partial\mathbb{E}_Q\left[f(\omega,g(\boldsymbol{x}))\right]}{\partial\theta} = \frac{1}{2}\left(\frac{\partial\mathbb{E}_Q\left[g(\boldsymbol{x})\right]}{\partial\theta} - \frac{\partial\mathbb{E}_Q\left[g(\boldsymbol{x})^2\right]}{\partial\theta}\mathbb{E}_Q\left[\omega\right]\right),
$$

with

$$
\begin{aligned}
\frac{\partial\mathbb{E}_Q\left[g(\boldsymbol{x})\right]}{\partial\theta} =& \frac{\partial\boldsymbol{\kappa}(\boldsymbol{x})}{\partial\theta}\boldsymbol{\mu}_2^s, \\
\frac{\partial\mathbb{E}_Q\left[g(\boldsymbol{x})^2\right]}{\partial\theta} =& \frac{\partial\tilde{k}(\boldsymbol{x},\boldsymbol{x})}{\partial\theta} + \frac{\partial\boldsymbol{\kappa}(\boldsymbol{x})}{\partial\theta}^\top\left(\Sigma_2^s + \boldsymbol{\mu}_2^s(\boldsymbol{\mu}_2^s)^\top\right)\boldsymbol{\kappa}(\boldsymbol{x}) + \boldsymbol{\kappa}(\boldsymbol{x})^\top\left(\Sigma_2^s + \boldsymbol{\mu}_2^s(\boldsymbol{\mu}_2^s)^\top\right)\frac{\partial\boldsymbol{\kappa}(\boldsymbol{x})}{\partial\theta},
\end{aligned}
$$

where $\boldsymbol{\kappa}(\boldsymbol{x}) = \boldsymbol{k}_s(\boldsymbol{x})^\top K_s^{-1}$ and $\tilde{k}(\boldsymbol{x},\boldsymbol{x}) = k(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}_s(\boldsymbol{x}) K_s^{-1} \boldsymbol{k}_s(\boldsymbol{x})^\top$. The remaining two terms are:

$$\begin{aligned}\frac{\partial \tilde{k}(\boldsymbol{x},\boldsymbol{x})}{\partial \theta} =& \frac{\partial k(\boldsymbol{x},\boldsymbol{x})}{\partial \theta} - \frac{\partial \boldsymbol{\kappa}(\boldsymbol{x})}{\partial \theta}\boldsymbol{k}_s(\boldsymbol{x}) - \boldsymbol{\kappa}(\boldsymbol{x})\frac{\partial \boldsymbol{k}_s(\boldsymbol{x})}{\partial \theta},\\ \frac{\partial \boldsymbol{\kappa}(\boldsymbol{x})}{\partial \theta} =& \frac{\partial \boldsymbol{k}_s(\boldsymbol{x})^\top}{\partial \theta} K_s^{-1} - \boldsymbol{k}_s(\boldsymbol{x}) K_s^{-1}\frac{\partial K_s}{\partial \theta} K_s^{-1}.\end{aligned}$$

After each variational step the hyperparameters are updated by

$$\boldsymbol{\theta}_{\text{new}} = \boldsymbol{\theta}_{\text{old}} + \varepsilon \frac{\partial \mathcal{L}(q)}{\partial \boldsymbol{\theta}},$$

where ε is the step size.

References

Ryan P. Adams, Iain Murray, and David 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*, pages 9–16, 2009. doi: 10.1145/1553374.1553376.

Philipp Batz, Andreas Ruttor, and Manfred Opper. Approximate Bayes learning of stochastic differential equations. *Phys. Rev.*, E98(2):022109, 2018. doi: 10.1103/PhysRevE.98.022109.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

David R. Brillinger. Maximum likelihood analysis of spike trains of interacting nerve cells. *Biological Cybernetics*, 59(3):189–200, 1988. doi: 10.1007/BF00318010.

Anders Brix and Peter J. Diggle. Spatiotemporal prediction for log-gaussian cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):823–841, 2001. doi: 10.1111/1467-9868.00315.

D. R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):129–164, 1955. ISSN 00359246.

Lehel Csató. Gaussian processes-iterative sparse approximations. 2002. URL `http://publications.aston.ac.uk/1327/`.

Lehel Csató and Manfred Opper. Sparse on-line gaussian processes. *Neural Computation*, 14(3):641–668, 2002. doi: 10.1162/089976602317250933.

John P Cunningham, Byron M Yu, Krishna V Shenoy, and Maneesh Sahani. Inferring neural firing rates from spike trains using gaussian processes. In *Advances in Neural Information Processing Systems 20*, pages 329–336. 2008. URL `http://papers.nips.cc/paper/3229-inferring-neural-firing-rates-from-spike-trains-using-gaussian-processes.pdf`.

Alexander G. de G. Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the kullback-leibler divergence between stochastic processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 231–239, 2016. URL `http://proceedings.mlr.press/v51/matthews16.html`.

Christian Donner and Manfred Opper. Inverse ising problem in continuous time: A latent variable approach. *Phys. Rev. E*, 96:062104, 2017. doi: 10.1103/PhysRevE.96.062104.

Christian Donner and Manfred Opper. Efficient bayesian inference for a gaussian process density model. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018. URL `http://auai.org/uai2018/proceedings/papers/34.pdf`.

Benjamin Dunn, Maria Mrreaunet, and Yasser Roudi. Correlations and functional connections in a population of grid cells. *PLOS Computational Biology*, 11(2):1–21, 2015. doi: 10.1371/journal.pcbi.1004052.

Paul Embrechts, Thomas Liniger, and Lu Lin. Multivariate hawkes processes: an application to financial data. *Journal of Applied Probability*, 48(A):367378, 2011. doi: 10.1239/jap/1318940477.

Seth Flaxman, Yee Whye Teh, and Dino Sejdinovic. Poisson intensity estimation with reproducing kernels. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 270–279. PMLR, 2017. URL `http://proceedings.mlr.press/v54/flaxman17a.html`.

Centre For The Biology Of Memory and Fransesca Sargolini. Grid cell data of sargolini et al 2006. 2014. doi: 10.11582/2014.00003.

Tom Gunter, Chris Lloyd, Michael A. Osborne, and Stephen J. Roberts. Efficient bayesian nonparametric modelling of structured point processes. In *Uncertainty in Artificial Intelligence (UAI)*, 2014. URL `https://arxiv.org/abs/1407.6949`.

Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971. ISSN 00063444.

James Hensman, Alexander G Matthews, Maurizio Filippone, and Zoubin Ghahramani. Mcmc for variationally sparse gaussian processes. In *Advances in Neural Information Processing Systems 28*, pages 1648–1656. 2015. URL `http://papers.nips.cc/paper/5875-mcmc-for-variationally-sparse-gaussian-processes.pdf`.

ST John and James Hensman. Large-scale Cox process inference using variational Fourier features. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2362–2370, 2018. URL `http://proceedings.mlr.press/v80/john18a.html`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *preprint arXiv*, abs/1412.6980, 2014. URL `http://arxiv.org/abs/1412.6980`.

John Frank Charles Kingman. *Poisson processes.* Oxford University Press, 1993. ISBN 9780198536932.

Alisa Kirichenko and Harry van Zanten. Optimality of poisson processes intensity learning with gaussian processes. *Journal of Machine Learning Research*, 16:2909–2919, 2015. URL http://jmlr.org/papers/v16/kirichenko15a.html.

J. Knollmüller, T. Steininger, and T. A. Enßlin. Inference of signals with unknown correlation structure from nonlinear measurements. *ArXiv e-prints*, 2017. URL https://arxiv.org/abs/1711.02955.

Takis Konstantopoulos, Zurab Zerakidze, and Grigol Sokhadze. *Radon–Nikodým Theorem*, pages 1161–1164. 2011. ISBN 978-3-642-04898-2.

P. A. W Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979. doi: 10.1002/nav.3800260304.

Scott Linderman, Matthew Johnson, and Ryan P Adams. Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. In *Advances in Neural Information Processing Systems 28*, pages 3456–3464. 2015. URL http://papers.nips.cc/paper/5660-dependent-multinomial-models-made-easy-stick-breaking-with-the-polya-gamma-augmentation.

Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 914–922, 2017. URL http://proceedings.mlr.press/v54/linderman17a.html.

Chris Lloyd, Tom Gunter, Michael Osborne, and Stephen Roberts. Variational inference for gaussian process modulated poisson processes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1814–1822, 2015. URL http://proceedings.mlr.press/v37/lloyd15.html.

Chris Lloyd, Tom Gunter, Michael Osborne, Stephen Roberts, and Tom Nickson. Latent point process allocation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 389–397, 2016. URL http://proceedings.mlr.press/v51/lloyd16.html.

Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke. Fujii, Alexis Boukouvalas, Pablo Leon-Villagra, Zoubin Ghahramani, and James Hensman. Gpflow: A gaussian process library using tensorflow. *Journal of Machine Learning Research*, 18 (40):1–6, 2017. URL http://jmlr.org/papers/v18/16-537.html.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998. doi: 10.1111/1467-9469.00115.

Luis Moreira-Matias, Joao Gama, Michel Ferreira, Joao Mendes-Moreira, and Luis Damas. Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1393–1402, 2013.

Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. Mcmc for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2006. ISBN 0-9749039-2-2.

Iain Murray, David MacKay, and Ryan P Adams. The gaussian process density sampler. In *Advances in Neural Information Processing Systems 21*, pages 9–16. 2009. URL `http://papers.nips.cc/paper/3410-the-gaussian-process-density-sampler.pdf`.

Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 541–548, 2010. URL `http://proceedings.mlr.press/v9/murray10a.html`.

Yosihiko Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998. doi: 10.1023/A:1003403601725.

Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistic models using plyagamma latent variables. *Journal of the American Statistical Association*, 108 (504):1339–1349, 2013. doi: 10.1080/01621459.2013.829001.

William H Press, Brian P Flannery, Saul A Teukolsky, William T Vetterling, et al. *Numerical recipes*, volume 3. Cambridge University Press, 2007. ISBN 978-0-521-88068-8.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, USA, 2006. ISBN 0-262-18253-X.

Yves-Laurent Kom Samo and Stephen Roberts. Scalable nonparametric bayesian inference on point processes with gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2227–2236, 2015. URL `http://proceedings.mlr.press/v37/samo15.html`.

Francesca Sargolini, Marianne Fyhn, Torkel Hafting, Bruce L. McNaughton, Menno P. Witter, May-Britt Moser, and Edvard I. Moser. Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, 312(5774):758–762, 2006. doi: 10.1126/science.1125572.

Arno Solin. *Stochastic Differential Equation Methods for Spatio-Temporal Gaussian Process Regression*. Aalto University, 2016. ISBN 978-952-60-6711-7.

Dietrich Stoyan and Antti Penttinen. Recent applications of point process methods in forestry statistics. *Statistical Science*, 15(1):61–78, 2000. ISSN 08834237.

Yee W. Teh and Vinayak Rao. Gaussian process modulated renewal processes. In *Advances in Neural Information Processing Systems 24*, pages 2474–2482, 2011. URL `http://papers.nips.cc/paper/4358-gaussian-process-modulated-renewal-processes.pdf`.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 567–574, 2009. URL `http://proceedings.mlr.press/v5/titsias09a.html`.

Christian J. Walder and Adrian N. Bishop. Fast Bayesian intensity estimation for the permanental process. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3579–3588, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/walder17a.html`.

Florian Wenzel, Théo Galy-Fajou, Christian Donner, Marius Kloft, and Manfred Opper. Scalable logit gaussian process classification. In *Advances in Approximate Bayesian Inference, NIPS Workshop*, 2017. URL `http://approximateinference.org/2017/accepted/WenzelEtAl2017.pdf`.