Journal of Machine Learning Research 16 (2015) 2909-2919 Submitted 9/14; Revised 3/15; Published 12/15

Optimality of Poisson Processes Intensity Learning with Gaussian Processes

Alisa Kirichenko A.KIRICHENKO@UVA.NL
Harry van Zanten HVZANTEN@UVA.NL
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
P.O. Box 94248, 1090 GE Amsterdam, The Netherlands

Editor: Manfred Opper

Abstract

In this paper we provide theoretical support for the so-called "Sigmoidal Gaussian Cox Process" approach to learning the intensity of an inhomogeneous Poisson process on a d-dimensional domain. This method was proposed by Adams, Murray and MacKay (ICML, 2009), who developed a tractable computational approach and showed in simulation and real data experiments that it can work quite satisfactorily. The results presented in the present paper provide theoretical underpinning of the method. In particular, we show how to tune the priors on the hyper parameters of the model in order for the procedure to automatically adapt to the degree of smoothness of the unknown intensity, and to achieve optimal convergence rates.

Keywords: inhomogeneous Poisson process, Bayesian intensity learning, Gaussian process prior, optimal rates, adaptation to smoothness

1. Introduction

Inhomogeneous Poisson processes are widely used models for count and point data in a variety of applied areas. A typical task in applications is to learn the underlying intensity of a Poisson process from a realised point pattern. In this paper we consider nonparametric Bayesian approaches to this problem. These do not assume a specific parametric form of the intensity function and produce posterior distributions which do not only give an estimate of the intensity, for example through the posterior mean or mode, but also give a measure of the remaining uncertainty through the spread of the posterior.

Several papers have explored nonparametric Bayesian approaches in this setting. An early reference is Møller et al. (1998), who study log-Gaussian priors. Gugushvili and Spreij (2013) recently considered Gaussian processes combined with different, non-smooth link functions. Kernel mixtures priors are considered in Kottas and Sansó (2007). Spline-based priors are used in DiMatteo et al. (2001) and Belitser et al. (2013).

The present study is motivated by a method that is not covered by earlier theoretical papers, namely the method of Adams et al. (2009). These authors presented the first approach that is also computationally fully nonparametric in the sense that it does not involve potentially inaccurate finite-dimensional approximations. The method involves a prior on the intensity that is a random multiple of a transformed Gaussian process (GP).

Both the hyper parameters of the GP and the multiplicative constant are endowed with priors as well, resulting in a hierarchical Bayes procedure (details in Section 2.3). Simulation experiments and real data examples in Adams et al. (2009) show that the method can give very satisfactory results.

The aim of this paper is to advance the theoretical understanding of the method of Adams et al. (2009), which they termed "Sigmoidal Gaussian Cox Process" (SGCP). It is by now well known both from theory and practice that nonparametric Bayesian methods need to be tuned very carefully to produce good results. An unfortunate choice of the prior or incorrectly tuned hyper parameters can easily result in procedures that give misleading results or that make sub-optimal use of the information in the training data. See for instance the by now classical reference Diaconis and Freedman (1986), or the more recent paper van der Vaart and van Zanten (2011) and the references therein.

A challenge in this problem (and in nonparametric function learning in general) is to devise a procedure that avoids overfitting and underfitting. The difficulty is that the appropriate degree of "smoothing" depends on the (unknown) regularity of the intensity function that produces the data. Indeed, intuitively it is clear that if the function is very smooth then to learn the intensity at a certain location we can borrow more information from neighboring points than if it is very rough. Ideally we want to have a procedure that automatically uses the appropriate degree of smoothing, that is, that *adapts* to regularity.

To address this issue theoretically it is common to take an asymptotic point of view. Specifically, we assume that we have n independent sets of training data, produced by Poisson processes on the d-dimensional domain $S = [0,1]^d$ (say), with the same intensity function $\lambda_0 : S \to [0,\infty)$. We aim to construct the learning procedure such that we achieve an optimal learning rate, irrespective of the regularity level of the intensity. In the problem at hand it is known that if λ_0 has regularity $\beta > 0$, then the best rate that any procedure can achieve is of the order $n^{-\beta/(d+2\beta)}$. This can be made precise in the minimax framework, for instance. For a fixed estimation or learning procedure, one can determine the largest expected loss that is incurred when the true function generating the data is varied over a ball of functions with fixed regularity β, say. This will depend on n and quantifies the worst-case rate of convergence for that fixed estimator for β-regular truths. The minimax rate is obtained by minimising this over all possible estimators. So it is the best convergence rate that any procedure can achieve, uniformly over a ball of functions with fixed regularity β. See, for example, Tsybakov (2009) for a general introduction to the minimax approach and Kutoyants (1998) or Reynaud-Bouret (2003) for minimax results in the context of the Poisson process model that we consider in this paper.

Note that the smoothness degree is unknown to us, so we can not use it in the construction of the procedure, but still we want that the posterior contracts around λ_0 at the rate $n^{-\beta/(d+2\beta)}$, as $n \to \infty$, if λ_0 is β-smooth. In this paper we prove that with appropriate priors on the hyper parameters, the SGCP approach of Adams et al. (2009) attains this optimal rate (up to a logarithmic factor). It does so for every regularity level $\beta > 0$, so it is fully *rate-adaptive*.

Technically the paper uses the mathematical framework for studying contraction rates for Gaussian and conditionally Gaussian priors as developed in van der Vaart and van Zanten (2008a) and van der Vaart and van Zanten (2009). We also use an extended version of a general result for Bayesian inference for 1-dimensional Poisson processes from Belitser et al.

(2013). On a general level the line of reasoning is similar to that of van der Vaart and van Zanten (2009). However, due to the presence of a link function and a random multiplicative constant in the SGCP model (see Section 2 ahead) the results of the latter paper do not apply in the present setting and additional mathematical arguments are required to prove the desired results.

The paper is organised as follows. In Section 2 we describe the Poisson process observation model and the SGCP prior model, which together determine a full hierarchical Bayesian model. The main result about the performance of the SGCP approach is presented and discussed in Section 3. Mathematical proofs are given in Section 4. In Section 5 we make some concluding remarks.

2. The SGCP Model

In this section we describe the observation model and the SGCP prior model for the intensity.

2.1 Observation Model

We assume we observe n independent copies of an inhomogeneous Poisson process on the d-dimensional unit cube $S = [0,1]^d$ (adaptation to other domains is straightforward). We denote these observed data by $N^1, \ldots, N^n$. Formally every N^i is a counting measure on subsets of S. The object of interest is the underlying *intensity function*. This is a (integrable) function $\lambda : [0,1]^d \to [0,\infty)$ with the property that given λ, every N^j is a random counting measure on $[0,1]^d$ such that $N^j(A)$ and $N^j(B)$ are independent if the sets $A, B \subset [0,1]^d$ are disjoint and the number of points $N^j(B)$ falling in the set B has a Poisson distribution with mean $\int_B \lambda(s)\, ds$. If we want to stress that the probabilities and expectations involving the observations N^j depend on λ, we use the notations P_λ and E_λ, respectively. We note that instead of considering observations from n independent Poisson processes with intensity λ, one could equivalently consider observations from a single Poisson process with intensity $n\lambda$.

2.2 Prior Model

The SGCP model introduced in Adams et al. (2009) postulates a-priori that the intensity function λ is of the form

$$\lambda(s) = \lambda^* \sigma(g(s)), \qquad s \in S, \tag{2.1}$$

where $\lambda^* > 0$ is an upper bound on λ, g is a GP indexed by S and σ is the sigmoid, or logistic function on the real line, defined by $\sigma(x) = (1 + e^{-x})^{-1}$. In the computational section of Adams et al. (2009) g is modeled as a GP with squared exponential covariance kernel and zero mean, with a prior on the length scale parameter. The hyper parameter λ^* is endowed with an independent gamma prior.

In the mathematical results presented in this paper we allow a bit more flexibility in the choice of the covariance kernel of the GP, the link function σ and the priors on the hyper parameters. We assume that g is a zero-mean, homogenous GP with covariance kernel given in spectral form by

$$\mathrm{E} g(s)g(t) = \int e^{-i\langle \xi, \ell(t-s)\rangle} \mu(\xi)\, d\xi, \qquad s,t \in S, \tag{2.2}$$

where $\ell > 0$ is an (inverse) length scale parameter and μ is a spectral density on $\mathbb{R}^d$ such that the map $a \mapsto \mu(a\xi)$ on $(0, \infty)$ is decreasing for every $\xi \in \mathbb{R}^d$ and that satisfies

$$\int e^{\delta||\xi||}\mu(d\xi) < \infty \tag{2.3}$$

for some $\delta > 0$ (the Euclidean inner product and norm are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively). Note that, in particular, the centered Gaussian spectral density satisfies this condition and corresponds to the squared exponential kernel

$$\mathrm{E}g(s)g(t) = e^{-\ell^2\|t-s\|^2}.$$

We endow the length scale parameter ℓ with a prior with density p_ℓ on $[0, \infty)$, for which we assume the bounds, for positive constants C_1, D_1, C_2, D_2, nonnegative constants p, q, and every sufficiently large $x > 0$,

$$C_1 x^p \exp(-D_1 x^d \log^q x) \leqslant p_\ell(x) \leqslant C_2 x^p \exp(-D_2 x^d \log^q x). \tag{2.4}$$

This condition is, for instance, satisfied if ℓ^d has a gamma distribution, which is a common choice in practice. Note however that the technical condition (2.4) is only a condition on the tail of the prior on ℓ. On the upper bound λ^* we put a prior satisfying an exponential tail bound. Specifically, we use a positive, continuous prior density p_{λ^*} on $[0, \infty)$ such that for some $c_0, C_0, \kappa > 0$,

$$\int_{\lambda_0}^{\infty} p_{\lambda^*}(x)\, dx \leqslant C_0 e^{-c_0 \lambda_0^\kappa} \tag{2.5}$$

for all $\lambda_0 > 0$. Note that this condition is fulfilled if we place a gamma prior on λ^*. Finally, we use a strictly increasing, infinitely smooth link function $\sigma : \mathbb{R} \to (0, 1)$ in (2.1) that satisfies

$$|\sqrt{\sigma(x)} - \sqrt{\sigma(y)}| \leqslant c|x - y| \tag{2.6}$$

for all $x, y \in \mathbb{R}$. This condition is in particular fulfilled for the sigmoid function employed by Adams et al. (2009). It holds for other link functions as well, for instance for the cdf of the standard normal distribution.

2.3 Full Hierarchical Model

With the assumptions made in the preceding sections in place, the full hierarchical specification of the prior and observation model can then be summarised as follows:

$$\begin{aligned}
\ell &\sim p_\ell \quad \text{(satisfying (2.4))} \\
\lambda^* &\sim p_{\lambda^*} \quad \text{(satisfying (2.5))} \\
g \,|\, \ell, \lambda^* &\sim \text{GP with kernel given by (2.2)–(2.3)} \\
\lambda \,|\, g, \ell, \lambda^* &\sim \text{defined by (2.1), with smooth } \sigma \text{ satisfying (2.6)} \\
N^1, \ldots, N^n \,|\, \lambda, g, \ell, \lambda^* &\sim \text{independent Poisson processes with intensity } \lambda.
\end{aligned}$$

Note that under the prior, several quantities are, by construction, independent. Specifically, ℓ and λ_* are independent, and g and λ^* are independent.

The main results of the paper concern the posterior distribution of the intensity function λ, that is, the conditional $\lambda \,|\, N^1, \ldots, N^n$. Throughout we will denote the prior on λ by Π and the posterior by $\Pi(\cdot \,|\, N^1, \ldots, N^n)$. In this setting Bayes' formula asserts that

$$\Pi(\lambda \in B \,|\, N^1, \ldots, N^n) = \frac{\int_B p(N^1, \ldots, N^n \,|\, \lambda)\, \Pi(d\lambda)}{\int p(N^1, \ldots, N^n \,|\, \lambda)\, \Pi(d\lambda)}, \tag{2.7}$$

where the likelihood is given by

$$p(N^1, \ldots, N^n \,|\, \lambda) = \prod_{i=1}^{n} e^{\int_S \lambda(x) N^i(dx) - \int_S (\lambda(x)-1)\, dx}$$

(see, for instance, Kutoyants, 1998).

3. Main Result

Consider the prior and observations model described in the preceding section and let $\Pi(\cdot \,|\, N^1, \ldots, N^n)$ be the corresponding posterior distribution of the intensity function λ.

The following theorem describes how quickly the posterior distribution contracts around the true intensity λ_0 that generates the data. The rate of contraction depends on the smoothness level of λ_0. This is quantified by assuming that λ_0 belongs to the Hölder space $C^\beta[0,1]^d$ for $\beta > 0$. By definition a function on $[0,1]^d$ belongs to this space if it has partial derivatives up to the order $\lfloor \beta \rfloor$ and if the $\lfloor \beta \rfloor$th order partial derivatives are all Hölder continuous of the order $\beta - \lfloor \beta \rfloor$. Here $\lfloor \beta \rfloor$ denotes the greatest integer strictly smaller than β. The rate of contraction is measured in the L^2-distance between the square root of intensities. This is the natural statistical metric in this problem, as it can be shown that in this setting the Hellinger distance between the models with intensity functions λ_1 and λ_2 is equivalent to $\min\{\|\sqrt{\lambda_1} - \sqrt{\lambda_2}\|_2, 1\}$ (see Belitser et al., 2013). Here $\|f\|_2$ denotes the L^2-norm of a function on $S = [0,1]^d$, that is, $\|f\|_2^2 = \int_S f^2(s)\, ds$.

Theorem 1 *Suppose that $\lambda_0 \in C^\beta[0,1]^d$ for some $\beta > 0$ and that λ_0 is strictly positive. Then for all sufficiently large $M > 0$,*

$$\mathrm{E}_{\lambda_0} \Pi(\lambda : \|\sqrt{\lambda} - \sqrt{\lambda_0}\|_2 \geqslant M n^{-\beta/(d+2\beta)} \log^\rho n | N^1, \ldots, N^n) \to 0 \tag{3.1}$$

as $n \to \infty$, for some $\rho > 0$.

The theorem asserts that if the intensity λ_0 that generates the data is β-smooth, then, asymptotically, all the posterior mass is concentrated in (Hellinger) balls around λ_0 with a radius that is up to a logarithmic factor of the optimal order $n^{-\beta/(d+2\beta)}$. Since the procedure does not use the knowledge of the smoothness level β, this indeed shows that the method is rate-adaptive, that is, the rate of convergence adapts automatically to the degree of smoothness of the true intensity. Let us mention once again that the conditions of the theorem are in particular fulfilled if in (2.1), λ^* is taken gamma, σ is the sigmoid (logistic) function, and g is a squared exponential GP with length scale ℓ, with ℓ^d a gamma variable.

4. Proof of Theorem 1

To prove the theorem we employ an extended version of a result from Belitser et al. (2013) that gives sufficient conditions for having (3.1) in the case $d = 1$, cf. their Theorem 1. Adaptation to the case of a general $d \in \mathbb{N}$ is straightforward. To state the result we need some (standard) notation and terminology. For a set of positive functions $\mathcal{F}$ we write $\sqrt{\mathcal{F}} = \{\sqrt{f}, f \in \mathcal{F}\}$. For $\varepsilon > 0$ and a norm $\|\cdot\|$ on $\mathcal{F}$, let $N(\varepsilon, \mathcal{F}, ||\cdot||)$ be the minimal number of balls of radius ε with respect to norm $\|\cdot\|$ needed to cover $\mathcal{F}$. The uniform norm $\|f\|_\infty$ of a function f on S is defined, as usual, as $\|f\|_\infty = \sup_{s\in S} |f(s)|$. The space of continuous function on S is denoted by $C(S)$. As usual, $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

Let Π now be a general prior on the intensity function λ and let $\Pi(\cdot \,|\, N^1, \ldots, N^n)$ be the corresponding posterior (2.7).

Theorem 2 *Assume that λ_0 is bounded away from 0. Suppose that for positive sequences $\overline{\delta}_n, \delta_n \to 0$ such that $n(\overline{\delta}_n \wedge \delta_n)^2 \to \infty$ as $n \to \infty$ and constants $c_1, c_2 > 0$, it holds that for all $L > 1$, there exist subsets $\mathcal{F}_n \subset C(S)$ and a constant c_3 such that*

$$1 - \Pi(\mathcal{F}_n) \leqslant e^{-Ln\delta_n^2}, \tag{4.1}$$

$$\Pi(\lambda : ||\lambda - \lambda_0||_\infty \leqslant \delta_n) \geqslant c_1 e^{-nc_2\delta_n^2}, \tag{4.2}$$

$$\log N(\overline{\delta}_n, \sqrt{\mathcal{F}_n}, \|\cdot\|_2) \leqslant c_3 n\overline{\delta}_n^2. \tag{4.3}$$

Then for $\varepsilon_n = \overline{\delta}_n \vee \delta_n$ and all sufficiently large $M > 0$,

$$\mathrm{E}_{\lambda_0}\Pi(\lambda : \|\sqrt{\lambda} - \sqrt{\lambda_0}\|_2 \geqslant M\varepsilon_n | N^1, \ldots N^n) \to 0 \tag{4.4}$$

as $n \to \infty$.

We note that this theorem has a form that is commonly encountered in the literature on contraction rates for nonparametric Bayes procedures. The so-called "prior mass condition" (4.2) requires that the prior puts sufficient mass near the true intensity function λ_0 generating the data. The "remaining mass condition" (4.1) and the "entropy condition" (4.3) together require that "most" of the prior mass should be concentrated on so-called "sieves" $\mathcal{F}_n$ that are not too large in terms of their metric entropy. The sieves grow as $n \to \infty$ and in the limit they capture all the posterior mass.

In the subsequent subsections we will show that the prior defined in Section 2.3 fulfills the conditions of this theorem, for $\delta_n = n^{-\beta/(2\beta+d)}(\log n)^{k_1}$ and $\overline{\delta}_n = L_1 n^{-\beta/(2\beta+d)}(\log n)^{(d+1)/2+2k_1}$, with $L_1 > 0$ and $k_1 = ((1 + d) \vee q)/(2 + d/\beta)$. The proofs build on earlier work, especially from van der Vaart and van Zanten (2009), in which results like (4.1)–(4.3) have been derived for GP's like g. Here we extend and adapt these results to deal with the additional link function σ and the prior on the maximum intensity λ^*.

4.1 Prior Mass Condition

In this section we show that with λ^*, σ and g as specified in Section 2.3 and $\lambda_0 \in C^\beta(S)$, we have

$$\mathrm{P}(\|\lambda^*\sigma(g) - \lambda_0\|_\infty \leqslant \delta_n) \geqslant c_1 e^{-nc_2\delta_n^2} \tag{4.5}$$

for constants $c_1, c_2 > 0$ and δ_n as defined above.

The link function σ is strictly increasing and smooth, hence it has a smooth inverse $\sigma^{-1} : (0,1) \to \mathbb{R}$. Define the function w_0 on S by

$$w_0(s) = \sigma^{-1}\Big(\frac{\lambda_0(s)}{2\|\lambda_0\|_\infty}\Big), \qquad s \in S,$$

so that $\lambda_0 = 2\|\lambda_0\|_\infty \sigma(w_0)$. Since the function λ_0 is positive and continuous on the compact set S, it is bounded away from 0 on S, say $\lambda_0 \geqslant a > 0$. It follows that $\lambda_0(s)/2\|\lambda_0\|_\infty$ varies in the compact interval $[a/2||\lambda_0||_\infty, 1/2]$ as s varies in S, hence w_0 inherits the smoothness of λ_0, that is, $w_0 \in C^\beta(S)$.

Now observe that for $\varepsilon > 0$,

$$\begin{aligned}
&\mathrm{P}(\|\lambda^*\sigma(g) - \lambda_0\|_\infty \leqslant 2\varepsilon) \\
&\qquad = \mathrm{P}(\|(\lambda^* - 2\|\lambda_0\|_\infty)\sigma(g) + 2\|\lambda_0\|_\infty(\sigma(g) - \sigma(w_0))\|_\infty \leqslant 2\varepsilon) \\
&\qquad \geqslant \mathrm{P}(|\lambda^* - 2\|\lambda_0\|_\infty| \leqslant \varepsilon)\mathrm{P}(\|\sigma(g) - \sigma(w_0)\|_\infty \leqslant \varepsilon/2\|\lambda_0\|_\infty).
\end{aligned}$$

Since λ^* has a positive, continuous density the first factor on the right is bounded from below by a constant times ε. Since the function $\sqrt{\sigma}$ is Lipschitz by assumption, the second factor is bounded from below by $\mathrm{P}(\|g - w_0\|_\infty \leqslant c\varepsilon)$ for a constant $c > 0$. By Theorem 3.1 in van der Vaart and van Zanten (2009) we have the lower bound

$$\mathrm{P}(\|g - w_0\|_\infty \leqslant \delta_n) \geqslant e^{-n\delta_n^2},$$

with δ_n as specified above. The proof of (4.5) is now easily completed.

4.2 Construction of Sieves

Let $\mathbb{H}^\ell$ be the RKHS of the GP g with covariance (2.2) and let $\mathbb{H}_1^\ell$ be its unit ball (see van der Vaart and van Zanten, 2008b for background on these notions). Let $\mathbb{B}_1$ be the unit ball in $C[0,1]^d$ relative to the uniform norm. Define

$$\mathcal{F}_n = \bigcup_{\lambda \leqslant \lambda_n} \lambda\sigma(\mathcal{G}_n),$$

where

$$\mathcal{G}_n = \left[M_n\sqrt{\frac{r_n}{\gamma_n}}\mathbb{H}_1^{r_n} + \varepsilon_n\mathbb{B}_1\right] \cup \left[\bigcup_{a \leqslant \gamma_n} (M_n\mathbb{H}_1^a) + \varepsilon_n\mathbb{B}_1\right],$$

and λ_n, M_n, γ_n, r_n and ε_n are sequences to be determined later. In the next two subsections we study the metric entropy of the sieves $\mathcal{F}_n$ and the prior mass of their complements.

4.3 Entropy

Since $\sqrt{\sigma}$ is bounded and Lipschitz we have, for $a, b \in [0, \lambda_n]$, some $c > 0$ and $f, g \in \mathcal{G}_n$,

$$\|\sqrt{a\sigma(f)} - \sqrt{b\sigma(g)}\|_\infty \leqslant |\sqrt{a} - \sqrt{b}| + c\sqrt{\lambda_n}\|f - g\|_\infty.$$

Since $|\sqrt{a}-\sqrt{b}| \leq \sqrt{|a-b|}$ for $a, b > 0$, it follows that for $\varepsilon > 0$,

$$N(2\varepsilon\sqrt{\lambda_n}, \sqrt{\mathcal{F}_n}, \|\cdot\|_2) \leqslant N(\varepsilon\sqrt{\lambda_n}, [0,\lambda_n], \sqrt{|\cdot|}) N(\varepsilon/c, \mathcal{G}_n, \|\cdot\|_\infty),$$

and hence

$$\log N(2\varepsilon\sqrt{\lambda_n}, \sqrt{\mathcal{F}_n}, \|\cdot\|_2) \lesssim \log\left(\frac{1}{\varepsilon}\right) + \log N(\varepsilon/c, \mathcal{G}_n, \|\cdot\|_\infty).$$

By formula (5.4) from van der Vaart and van Zanten (2009),

$$\log N(3\varepsilon_n, \mathcal{G}_n, \|\ \cdot\ \|_\infty) \quad \leqslant \quad K r_n^d \left(\log \frac{d^{1/4} M_n^{3/2} \sqrt{2\tau r_n}}{\varepsilon_n^{3/2}}\right)^{1+d} \quad + \quad 2\log \frac{2M_n\sqrt{||\mu||}}{\varepsilon_n},$$

for $\|\mu\|$ the total mass of the spectral measure μ, τ^2 the second moment of μ, a constant $K > 0$, $\gamma_n = \varepsilon_n/(2\tau\sqrt{d}M_n)$, $r_n > A$ for some constant $A > 0$, and given that the following relations hold:

$$d^{1/4} M_n^{3/2} \sqrt{2\tau r_n} > 2\varepsilon_n^{3/2}, \qquad M_n\sqrt{||\mu||} > \varepsilon_n. \tag{4.6}$$

By substituting $\bar{\eta}_n = \varepsilon_n\sqrt{\lambda_n}$ we get that for some constants K_1 and K_2,

$$\log N(2\bar{\eta}_n, \sqrt{\mathcal{F}_n}, \|\cdot\|_2) \lesssim K_1 r_n^d \left(\log \frac{\lambda_n^{3/4} M_n^{3/2} d^{1/4}\sqrt{2\tau r_n}}{\bar{\eta}_n^{3/2}}\right)^{1+d} + K_2 \log \frac{\lambda_n^{1/2} M_n}{\bar{\eta}_n},$$

when $M_n > 1$. In terms of $\bar{\eta}$ the conditions (4.6) can be rewritten as

$$d^{1/4} M_n^{3/2} \lambda_n^{3/4} \sqrt{2\tau r_n} > 2\bar{\eta}_n^{3/2}, \qquad M_n \lambda_n^{1/2}\sqrt{||\mu||} > \bar{\eta}_n. \tag{4.7}$$

So we conclude that we have the entropy bound

$$\log N(\bar{\eta}_n, \sqrt{\mathcal{F}_n}, \|\cdot\|_2) \lesssim n\bar{\eta}_n^2$$

for sequences λ_n, M_n, r_n and $\bar{\eta}_n$ satisfying (4.7) and

$$K_1 r_n^d \left(\log \frac{\lambda_n^{3/4} M_n^{3/2} d^{1/4}\sqrt{2\tau r_n}}{\bar{\eta}_n^{3/2}}\right)^{1+d} < n\bar{\eta}_n^2, \quad K_2 \log \frac{\lambda_n^{1/2} M_n}{\bar{\eta}_n} < n\bar{\eta}_n^2. \tag{4.8}$$

4.4 Remaining Mass

By conditioning we have

$$\begin{aligned} \mathrm{P}(\lambda^*\sigma(g) \notin \mathcal{F}_n) &= \int_0^\infty \mathrm{P}(\lambda\sigma(g) \notin \mathcal{F}_n) p_{\lambda^*}(\lambda)\, d\lambda \\ &\leqslant \int_0^{\lambda_n} \mathrm{P}(\lambda\sigma(g) \notin \mathcal{F}_n) p_{\lambda^*}(\lambda)\, d\lambda + \int_{\lambda_n}^\infty p_{\lambda^*}(\lambda)\, d\lambda. \end{aligned}$$

By (2.5) the second term is bounded by a constant times $\exp(-c_0\lambda_n^\kappa)$. For the first term, note that for $\lambda \leqslant \lambda_n$ we have

$$\lambda^{-1} \bigcup_{\lambda' \leqslant \lambda_n} \lambda'\sigma(\mathcal{G}_n) \supset \sigma(\mathcal{G}_n),$$

hence $\mathrm{P}(\lambda\sigma(g) \notin \mathcal{F}_n) \leqslant \mathrm{P}(g \notin \mathcal{G}_n)$. From (5.3) in van der Vaart and van Zanten (2009) we obtain the bound

$$\mathrm{P}(g \notin \mathcal{G}_n) \leqslant \frac{K_3 r_n^{p-d+1} e^{-D_2 r_n^d \log^q r_n}}{\log^q r_n} + e^{-M_n^2/8},$$

for some $K_3 > 0$, $\varepsilon_n < \varepsilon_0$ for a small constant $\varepsilon_0 > 0$, and M_n, r_n and ε_n satisfying

$$M_n^2 > 16 K_4 r_n^d (\log(r_n/\varepsilon_n))^{1+d}, \qquad r_n > 1, \tag{4.9}$$

where K_4 is some large constant. It follows that $\mathrm{P}(g \notin \mathcal{G}_n)$ is bounded above by a multiple of $\exp(-Ln\widetilde{\eta}_n^2)$ for a given constant L and $\widetilde{\eta}_n = \lambda_n \varepsilon_n$, provided M_n, r_n, γ_n and ε_n satisfy (4.9) and

$$D_2 r_n^d \log^q r_n \geqslant 2Ln\widetilde{\eta}_n^2, \quad r_n^{p-d+1} \leqslant e^{Ln\widetilde{\eta}_n^2}, \quad M_n^2 \geqslant 8Ln\widetilde{\eta}_n^2. \tag{4.10}$$

Note that in terms of $\tilde{\eta}_n$, (4.9) can be rewritten as

$$M_n^2 > 16 K_4 r_n^d (\log(r_n \lambda_n/\widetilde{\eta}_n))^{1+d}, \qquad r_n > 1. \tag{4.11}$$

We conclude that if (4.11),(4.10) holds and

$$c_0 \lambda_n^\kappa > Ln\widetilde{\eta}_n^2, \tag{4.12}$$

then

$$\mathrm{P}(\lambda^* \sigma(g \notin \mathcal{F}_n)) \lesssim e^{-Ln\widetilde{\eta}_n^2}.$$

4.5 Completion of the Proof

In the view of the preceding it only remains to show that $\widetilde{\eta}_n$, $\bar{\eta}_n$, r_n, $M_n > 1$ and λ_n can be chosen such that relations (4.7), (4.8), (4.10), (4.11) and (4.12) hold.

One can see that it is true for $\widetilde{\eta}_n = \delta_n$ and $\bar{\eta}_n = \overline{\delta}_n$ described in the theorem, with r_n, M_n, λ_n as follows:

$$\begin{aligned} r_n &= L_2\, n^{\frac{1}{2\beta+d}} (\log n)^{\frac{2k_1}{d}}, \\ M_n &= L_3\, n^{\frac{d}{2(2\beta+d)}} (\log n)^{\frac{d+1}{2}+2k_1}, \\ \lambda_n &= L_4\, n^{\frac{d}{\kappa(2\beta+d)}} (\log n)^{\frac{4k_1}{\kappa}} \end{aligned}$$

for some large constants $L_2, L_3, L_4 > 0$.

5. Concluding Remarks

We have shown that the SGCP approach to learning intensity functions proposed by Adams et al. (2009) enjoys very favorable theoretical properties, provided the priors on the hyper parameters are chosen appropriately. The result shows there is some flexibility in the construction of the prior. The squared exponential GP may be replaced by other smooth stationary processes, other link functions may be chosen, and there is also a little room in the choice of the priors on the length scale and the multiplicative parameter. This flexibility is limited, however, and although our result only gives upper bounds on the

contraction rate, results like those of Castillo (2008) and van der Vaart and van Zanten (2011) lead us to believe that one might get sub-optimal performance when deviating too much from the conditions that we have imposed. Strictly speaking the matter is open however and additional research is necessary to make this belief precise and to describe the exact boundaries between good and sub-optimal behaviours.

We expect that a number of generalizations of our results are possible. For instance, it should be possible to obtain generalizations to anisotropic smoothness classes and priors as considered in Bhattacharya et al. (2014), and classes of analytic functions as studied in van der Vaart and van Zanten (2009). These generalizations take considerable additional technical work however and are therefore not worked out in this paper. We believe they would not change the general message of the paper.

Acknowledgments

Research supported by the Netherlands Organisation for Scientific Research, NWO.

References

Adams, R. P., Murray, I. and MacKay, D. J. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 9–16. ACM.

Belitser, E., Serra, P. and van Zanten, J. H. (2013). Rate-optimal Bayesian intensity smoothing for inhomogeneous Poisson processes. *To appear in J. Statist. Plann. Inference, arXiv:1304.6017* .

Bhattacharya, A., Pati, D. and Dunson, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *Ann. Statist.* **42**(1), 352–381.

Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2**, 1281–1299.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**(1), 1–67.

DiMatteo, I., Genovese, C. R. and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**(4), 1055–1071.

Gugushvili, S. and Spreij, P. (2013). A note on non-parametric Bayesian estimation for Poisson point processes. *ArXiv E-prints*.

Kottas, A. and Sansó, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference* **137**(10), 3151–3163.

Kutoyants, Y. A. (1998). *Statistical inference for spatial Poisson processes*. Springer.

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

Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields* **126**(1), 103–153.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer, New York.

van der Vaart, A. W. and van Zanten, J. H. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36**(3), 1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections* **3**, 200–222.

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37**(5B), 2655–2675.

van der Vaart, A. W. and van Zanten, J. H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.* **12**, 2095–2119.