Asymptotic results in nonparametric Bayesian function estimation

Kirichenko, A.

Chapter 4

Poisson intensity estimation

4.1 Introduction

Inhomogeneous Poisson processes are widely used models for counting the occurrence of certain events in a variety of applied areas. A typical task in applications is to learn the underlying intensity function of a Poisson process from a realised point pattern. In this chapter 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, e.g. 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 (GP) 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. [2015].

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 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 GP. Both the hyperparameters of the GP and the multiplicative constant are endowed with priors as well, resulting in a hierarchical Bayes procedure (details in Section 4.2.3). Simulation experiments and real data examples in Adams et al. [2009] show that the method can give satisfactory results.

The aim of this chapter is to advance the theoretical understanding of the method of Adams et al. [2009], which they named "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 hyperparameters can easily result in procedures that give misleading results or that make suboptimal use of the information in the training data. See for instance Diaconis and Freedman [1986b], 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 estimation 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 according to which the data is generated. 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 neighbouring points than if it is very rough. Ideally we want to have a procedure that automatically uses the appropriate degree of smoothing, i.e. that adapts to regularity.

To address this issue theoretically we again take an asymptotic approach. We assume that we have n independent sets of training data, produced by Poisson processes, say, on the d-dimensional domain $S = [0,1]^d$, with the same intensity function $\lambda_0 : S \to [0,\infty)$. We aim to construct a learning procedure that achieves an optimal convergence 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 $n^{-\beta/(d+2\beta)}$. See Kutoyants [1998] or Reynaud-Bouret [2003] for minimax results in the context of the Poisson process model considered in this chapter.

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. We prove that with appropriate priors on the hyperparameters, 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.

In order to study contraction rates for Gaussian and conditionally Gaussian priors we use the mathematical framework developed in van der Vaart and van Zanten [2008a] and van der Vaart and van Zanten [2009]. We also use an extended version of the general result for Bayesian inference for 1-dimensional Poisson processes from Belitser et al. [2015]. The reasoning is mainly 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 4.2 ahead) their results are not applicable in the present setting and additional mathematical arguments are required to derive the contraction rates.

The remainder of this chapter is organised as follows. In the next section 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 4.3. Mathematical proofs are given in Section 4.4.

4.2 The SGCP model

4.2.1 Observation model

We assume we observe n independent copies $N^1, \dots, N^n$ of an inhomogeneous Poisson process with the intensity function λ_0 on the d-dimensional unit cube $S = [0,1]^d$ (adaptation to other domains is straightforward). Formally every N^i is a counting measure on subsets of S. The intensity function is a (Lebesgue integrable) function $\lambda_0 : [0,1]^d \to [0,\infty)$ with the property that given λ_0, 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 $\int_B \lambda_0(s)\, ds$. If we want to stress that the probabilities and expectations involving the observations N^j depend on λ_0, we use the notations $\mathbb{P}_{\lambda_0}$ and $\mathbb{E}_{\lambda_0}$, respectively. We note that instead of considering observations from n independent Poisson processes with intensity λ_0, one could equivalently consider observations from a single Poisson process with intensity $n\lambda_0$.

4.2.2 Prior model

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

$$\lambda(s) = \lambda^* \sigma(g(s)), \qquad s \in S, \tag{4.1}$$

where $\lambda^* > 0$ is an upper bound on λ, g is a GP indexed by S and σ is the sigmoid, or the 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 modelled as a GP with squared exponential covariance kernel and zero mean, with a prior on the length scale parameter. The hyperparameter λ^* is endowed with an independent gamma prior.

In the mathematical results presented here 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 hyperparameters. We assume that g is a zero-mean, homogenous GP with covariance kernel given in spectral form by

$$\mathbb{E}g(s)g(t) = \int e^{-i\langle \xi, \ell(t-s)\rangle} \mu(\xi)\, d\xi, \qquad s, t \in S, \tag{4.2}$$

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

$$\int e^{\delta \|\xi\|} \mu(d\xi) < \infty \tag{4.3}$$

for some $\delta > 0$ (the Euclidean inner product and norm in $\mathbb{R}^d$ are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively). Note that, in particular, the centred Gaussian spectral

density satisfies this condition and corresponds to the squared exponential kernel

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

We endow the length scale parameter ℓ with a prior with density p_ℓ on $[0,\infty)$ that satisfies the following bounds. We assume there exist positive constants C_1, D_1, C_2, D_2, nonnegative constants p, q, and every sufficiently large $x > 0$ such that

$$C_1 x^p \exp(-D_1 x^d \log^q x) \leq p_\ell(x) \leq C_2 x^p \exp(-D_2 x^d \log^q x). \tag{4.4}$$

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 (4.4) is only a condition on the tail of the prior on ℓ. Furthermore, on the upper bound parameter 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 $c_0, C_0, \kappa > 0$,

$$\int_{\lambda_0}^{\infty} p_{\lambda^*}(x)\,dx \leq C_0 e^{-c_0\lambda_0^\kappa} \tag{4.5}$$

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 (4.1) that satisfies

$$|\sqrt{\sigma(x)} - \sqrt{\sigma(y)}| \leq c|x-y| \tag{4.6}$$

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 cumulative distribution function of the standard normal distribution.

4.2.3 Full hierarchical model

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

$$\begin{aligned}
\ell &\sim p_\ell \quad \text{(satisfying (4.4))} \\
\lambda^* &\sim p_{\lambda^*} \quad \text{(satisfying (4.5))} \\
g \,|\, \ell, \lambda^* &\sim \text{GP with kernel given by (4.2)–(4.3)} \\
\lambda \,|\, g, \ell, \lambda^* &\sim \text{defined by (4.1), with smooth } \sigma \text{ satisfying (4.6)} \\
N^1, \ldots, N^n \,|\, \lambda, g, \ell, \lambda^* &\sim \text{independent Poisson processes with intensity } \lambda.
\end{aligned}$$

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 chapter concern the posterior distribution of the intensity function λ, i.e. the conditional distribution $\lambda \,|\, N^1, \ldots, N^n$. We denote the prior on λ by Π and the posterior by $\Pi(\cdot \,|\, N^1, \ldots, N^n)$. In this setting the Bayes' formula

asserts that

$$\Pi(\lambda \in B \,|\, N^1, \dots, N^n) = \frac{\int_B p(N^1, \dots, N^n \,|\, \lambda)\, \Pi(d\lambda)}{\int p(N^1, \dots, N^n \,|\, \lambda)\, \Pi(d\lambda)}, \tag{4.7}$$

where the likelihood is given by

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

(see, e.g., Kutoyants [1998]).

4.3 Main result

Consider the prior and observations model described in the preceding section and let $\Pi(\cdot \,|\, N^1, \dots, N^n)$ 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 according to which the data is generated. 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 roots of intensities. This is a 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. [2015]). Here $\|f\|_2$ denotes the L^2-norm of a function on $S = [0,1]^d$, i.e. $\|f\|_2^2 = \int_S f^2(s)\, ds$.

Theorem 4.3.1. *Suppose that $\lambda_0 \in C^\beta\left([0,1]^d\right)$ for some $\beta > 0$ and that λ_0 is strictly positive. Then for all sufficiently large $M > 0$,*

$$\mathbb{E}_{\lambda_0} \Pi(\lambda : \|\sqrt{\lambda} - \sqrt{\lambda_0}\|_2 \geq M n^{-\beta/(d+2\beta)} \log^\rho n | N^1, \dots, N^n) \to 0 \tag{4.8}$$

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, i.e. 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 (4.1) parameter λ^* is gamma–distributed, σ is the sigmoid (logistic) function, and g is a squared exponential GP with length scale ℓ, where ℓ^d is a gamma variable.

4.4 Proofs

In this section we present the proof of Theorem 4.3.1. To prove the theorem we employ an extended version of a result from Belitser et al. [2015] that gives sufficient conditions for having (4.8) 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 $\mathcal{F}^c$ for its complement and $\sqrt{\mathcal{F}} = \{\sqrt{f}, f \in \mathcal{F}\}$. 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 $C(S)$.

Let Π now be a general prior on the intensity function λ and let $\Pi(\cdot \mid N^1, \dots, N^n)$ be the corresponding posterior (4.7). As usual, we denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

Theorem 4.4.1. *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*

$$1 - \Pi(\mathcal{F}_n) \le e^{-Ln\delta_n^2}, \tag{4.9}$$

$$\Pi(\lambda : \|\lambda - \lambda_0\|_\infty \le \delta_n) \ge c_1 e^{-nc_2\delta_n^2}, \tag{4.10}$$

$$\log N(\overline{\delta}_n, \sqrt{\mathcal{F}_n}, \|\cdot\|_2) \le c_3 n\overline{\delta}_n^2. \tag{4.11}$$

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

$$\mathbb{E}_{\lambda_0}\Pi(\lambda : \|\sqrt{\lambda} - \sqrt{\lambda_0}\|_2 \ge M\varepsilon_n | N^1, \dots N^n) \to 0 \tag{4.12}$$

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" (4.10) requires that the prior puts sufficient mass near the true intensity function λ_0 according to which the data is generated. The "remaining mass condition" (4.9) and the "entropy condition" (4.11) 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 show that the prior defined in Section 4.2.3 fulfils 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 are based on earlier work, especially from van der Vaart and van Zanten [2009], in which results like (4.9)–(4.11) have been derived for the 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^*.

4.4.1 Prior mass condition

In this section we show that with λ^*, σ, and g specified in Section 4.2.3 and $\lambda_0 \in C^\beta(S)$, we have

$$\mathbb{P}(\|\lambda^*\sigma(g) - \lambda_0\|_\infty \le \delta_n) \ge c_1 e^{-nc_2\delta_n^2} \tag{4.13}$$

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 $\lambda_0 \ge a > 0$. 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, i.e. $w_0 \in C^\beta(S)$.

Now observe that for $\varepsilon > 0$,

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

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 $\mathbb{P}(\|g - w_0\|_\infty \le c\varepsilon)$ for a constant $c > 0$. By Theorem 3.1 in van der Vaart and van Zanten [2008b] we have the lower bound

$$\mathbb{P}(\|g - w_0\|_\infty \le \delta_n) \ge e^{-n\delta_n^2},$$

with δ_n as specified above. The proof of (4.13) is now easily completed.

4.4.2 Construction of sieves

Let $\mathbb{H}^\ell$ be the RKHS of the GP g with covariance (4.2) 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

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

where

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

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.

4.4.3 Entropy

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

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

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

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

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],

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

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

$$d^{1/4} M_n^{3/2} \sqrt{2\tau r_n} > 2\varepsilon_n^{3/2}, \qquad M_n \sqrt{\|\mu\|} > \varepsilon_n. \tag{4.14}$$

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

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

when $M_n > 1$. In terms of $\overline{\eta}$ the conditions (4.14) can be rewritten as

$$d^{1/4} M_n^{3/2} \lambda_n^{3/4} \sqrt{2\tau r_n} > 2\overline{\eta}_n^{3/2}, \qquad M_n \lambda_n^{1/2} \sqrt{\|\mu\|} > \overline{\eta}_n. \tag{4.15}$$

So we conclude that we have the entropy bound

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

for sequences λ_n, M_n, r_n and $\overline{\eta}_n$ satisfying (4.15) and

$$K_1 r_n^d \left(\log \frac{\lambda_n^{3/4} M_n^{3/2} d^{1/4} \sqrt{2\tau r_n}}{\overline{\eta}_n^{3/2}} \right)^{1+d} < n\overline{\eta}_n^2, \quad K_2 \log \frac{\lambda_n^{1/2} M_n}{\overline{\eta}_n} < n\overline{\eta}_n^2. \tag{4.16}$$

4.4.4 Remaining mass

By conditioning we have

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

By (4.5) the second term is bounded by a constant times $\exp(-c_0\lambda_n^\kappa)$. For the first term, note that for $\lambda \leq \lambda_n$ we have

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

Hence, $\mathbb{P}(\lambda\sigma(g) \notin \mathcal{F}_n) \leq \mathbb{P}(g \notin \mathcal{G}_n)$. From (5.3) in van der Vaart and van Zanten [2009] we obtain the bound

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

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

$$M_n^2 > 16 K_4 r_n^d (\log(r_n/\varepsilon_n))^{1+d}, \qquad r_n > 1, \tag{4.17}$$

where K_4 is some large constant. It follows that $\mathbb{P}(g \notin \mathcal{G}_n)$ 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 (4.17) and

$$D_2 r_n^d \log^q r_n \geq 2Ln\widetilde{\eta}_n^2, \quad r_n^{p-d+1} \leq e^{Ln\widetilde{\eta}_n^2}, \quad M_n^2 \geq 8Ln\widetilde{\eta}_n^2. \tag{4.18}$$

Note that in terms of $\tilde{\eta}_n$, (4.17) can be rewritten as

$$M_n^2 > 16 K_4 r_n^d (\log(r_n\lambda_n/\widetilde{\eta}_n))^{1+d}, \qquad r_n > 1. \tag{4.19}$$

We conclude that if (4.19),(4.18) holds and

$$c_0 \lambda_n^\kappa > Ln\widetilde{\eta}_n^2, \tag{4.20}$$

then

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

4.4.5 Completion of the proof

In the view of the preceding it only remains to show that $\widetilde{\eta}_n$, $\overline{\eta}_n$, r_n, $M_n > 1$ and λ_n can be chosen such that relations (4.15), (4.16), (4.18), (4.19) and (4.20) hold.

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

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

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