Nonparametric estimation of regression functions in point process models

Sebastian Döhler and Ludger Rüschendorf
University of Freiburg

Abstract

We prove that the empirical L^2-risk minimizing estimator over some general type of sieve classes is universally, strongly consistent for the regression function in a class of point process models of Poissonian type (random sampling processes). The universal consistency result needs weak assumptions on the underlying distributions and regression functions. It applies in particular to neural net classes and to radial basis function nets. For the estimation of the intensity functions of a Poisson process a similar technique yields consistency of the sieved maximum likelihood estimator for some general sieve classes.

Keywords: nonparametric estimation, regression functions, point process models, neural nets, sieve classes

1 Introduction

In this paper we consider estimation of the regression function and the intensity function for Poissonian type point process models. The main result of this paper is a uniform strong consistency result for empirical L^2-risk minimizing estimators for the regression function over some general type of sieve classes. The method of empirical risk minimization was proposed in the work of Vapnik and Chervonenkis (1974) and then developed into a powerful method by various authors. For general review see the books of Devroye, Györfi, and Lugosi (1996) and Györfi, Kohler, Krzyzak, and Walk (2002) while for the estimation of the regression function for fixed sample size see Devroye et al. (1994), Lugosi and Zeger (1995), and Krzyzak et al. (1996).

The main steps of this approach are an error decomposition into a bias term and a stochastic error term. For handling the bias term some approximation result is needed while the treatment of the stochastic error term needs to establish some uniform law of large numbers for the empirical risk measure term. The proof of this uniform law is based on empirical process theory.

For a concise presentation of this method see Lugosi and Zeger (1995) and Krzyzak, Linder, and Lugosi (1996).

In this paper we give an extension of this approach from the fixed sample case to the framework of random sampling processes of Poissonian type. For the estimation of the regression function we develop in detail the extension of the method of empirical L^2-risk minimizing estimation. For the estimation of the intensity function it is more natural to switch to sieved maximum likelihood estimation. Since the technical details are similar we give only a sketch of the corresponding proof of the consistency result. The estimation of the intensity function of Poisson processes is a well motivated and studied subject in the statistical literature. The results in this paper supplement the classes of estimation procedures which were investigated for Poisson processes by many authors (for general review on statistical problems for point processes see the books of Karr (1991) and Kutoyants (1998)) by nonparametric estimators of general sieve type which turned out to be very useful and promising for various kinds of applications in the fixed sample case.

Let $\Phi = \sum_{i=1}^{N} \varepsilon_{(X_i,Y_i)}$ be a point process with $(X_i, Y_i) \sim (X, Y)$ an iid sequence with values in $\mathbb{R}^k \times \mathbb{R}$ and a random number N of points independent of the sequence $((X_i, Y_i))$. Our aim is to estimate the regression function $m(x) = E(Y \mid X = x)$ based on an iid sample $\Phi_1, \ldots, \Phi_n$ where $\Phi_i = \sum_{j=1}^{N_i} \varepsilon_{(X_{ij},Y_{ij})} \sim \Phi$.

A finite sample estimation theory of this type of point process models has been developed in Rüschendorf (1989) (where they are called random sampling processes). A main example are Poisson processes with a finite intensity in the observation window. Kernel estimators for the estimation of the intensity in the Poisson case were studied in Rudemo (1982) and Kutoyants (1998). The arithmetic mean estimator for the intensity was investigated in Liese (1990) and Liese and Ziegler (1999). Kim and Koo (2000) studied maximum likelihood type sieve estimators of the intensity for wavelet sieves while Reynaud-Bouret (2001) established and analysed adaptive properties of penalized projection estimators in various smoothness classes for the intensity function.

Our main aim in this paper is to prove a universal consistency result for estimators of the regression function minimizing the empirical L^2 risk over some sieve class $(\mathcal{F}_n)$, i.e. we consider

$$\widehat{m}_n = \widehat{m}_n(\mathcal{F}_n) = \underset{f \in \mathcal{F}_n}{\arg\min} \sum_{i=1}^{n} \sum_{j=1}^{N_i} (Y_{ij} - f(X_{ij}))^2 . \tag{1.1}$$

Here and throughout the paper we assume for the sake of simplicity the existence of minimizing functions in $\mathcal{F}_n$. If the minima do not exist the same analysis can be carried out with functions approximating the infimum. For a similar remark cf. also the paper of Lugosi and Zeger (1995).

Concerning the sieve class $(\mathcal{F}_n)$ in (1.1) we assume that $\mathcal{F}_n$ approximates a dense subclass $\mathcal{F}$ in L^2 which is generated by a VC-class as a vector space.

Condition A)

1) Assume that $\mathcal{F} \subset L^2(P)$ is dense for all probability measures P on $\mathbb{R}^k$.

2) $\mathcal{F} = \{\sum_i a_i f_i; a_i \in \mathbb{R}, f_i \in \mathcal{F}_0\}$ is a vector space generated by a VC-class $\mathcal{F}_0$ of functions $f : \mathbb{R}^k \to [0,1]$.

The sieve $\mathcal{F}_n \subset \mathcal{F}$ is constructed as a sequence of subclasses of $\mathcal{F}$, such that $\mathcal{F}_n \uparrow \mathcal{F}$.

This type of condition was used in Döhler (1999) in the context of hazard regression estimation in censoring models. Conditions 1), 2) hold true in particular for some general classes of neural nets and radial basis functions.

Let $\sigma : \mathbb{R} \to [0,1]$ be monotonically nondecreasing and $\lim_{x\to\infty} \sigma(x) = 1$, $\lim_{x\to-\infty} \sigma(x) = 0$ and define the neural net class

$$\mathcal{F} = \mathcal{F}(\sigma) \quad := \quad \Big\{ f : \mathbb{R}^k \to \mathbb{R}; x \to \sum_{i=1}^{K} c_i \sigma(a_i x + b_i) + c_0; \qquad K \in \mathbb{N}, a_i \in \mathbb{R}^k, c_i \in \mathbb{R} \Big\}. \tag{1.2}$$

Furthermore, for $h : \mathbb{R}_+ \to [0,1]$, with $\int h(\|x\|) d\lambda^k(x) < \infty$ let

$$\mathcal{F} = \mathcal{R}(h) \quad := \quad \Big\{ f : \mathbb{R}^k \to \mathbb{R}_+; x \to \sum_{i=1}^{K} c_i h(\|A_i x + b_i\|) + c_0; \qquad K \in \mathbb{N}, A_i \in \mathbb{R}^{k\times k}, b_i \in \mathbb{R}^k, c_i \in \mathbb{R} \Big\} \tag{1.3}$$

denote the radial basis function net generated by h.

Then $\mathcal{F}(\sigma)$ and $\mathcal{R}(h)$ satisfy conditions 1), 2) in condition A). The denseness assertion follows from Hornik (1991) resp. Krzyzak, Linder, and Lugosi (1996) while the finiteness of the VC-dimension follows from the fact that in both cases each member of $\mathcal{F}_0$ is the composition of a fixed increasing or decreasing function with a finite dimensional vector space for functions).

An estimator sequence $\widehat{\mu}_n = \widehat{\mu}_n(\Phi_1, \ldots, \Phi_n) = \widehat{\mu}_n(x)$ is called universally strongly consistent for $m(x) = E(Y \mid X = x)$ in L^2-sense if $E(\widehat{\mu}_n(X) - m(X))^2 = \int (\widehat{\mu}_n(x) - m(x))^2 d\mu(x) \to 0$ a.s. for all (X,Y) such that $EY^2 < \infty$, where $\mu = P^X$. Our aim is to prove universal strong consistency for the empirical risk minimizing estimator $\widehat{m}_n$.

2 Universal consistency of sieved empirical L^2-risk minimizing estimators

In this section we state the main universal consistency result for the empirical L^2-risk minimizing estimator sequence $(\widehat{m}_n)$ and establish an extension of the truncation technique of Lugosi and Zeger (1995) to the case of random sampling processes.

The estimator is based on a sieve $\mathcal{F}_n = \mathcal{F}(\beta_n, K_n) \subset \mathcal{F}$ with $\beta_n \uparrow \infty$, $K_n \uparrow \infty$ with elements $f = \sum_{i=1}^K c_i f_i \in \mathcal{F}_n$ such that $|c_i| \le \beta_n$ and $K \le K_n$ (cp. (3.1)).

Theorem 2.1 (Universal strong consistency of empirical risk estimators) *Let $\Phi = \sum_{i=1}^N \varepsilon_{(X_i,Y_i)}$ be a (random sampling) point process in $\mathbb{R}^k \times \mathbb{R}$ such that $EY^2 < \infty$ and $EN^r < \infty$ for some $r > 6$ and let (Φ_i) be an iid sample of Φ. Then under assumption A) there exists a sieve $(\mathcal{F}_n) = (\mathcal{F}(\beta_n, K_n)) \subset \mathcal{F}$ such that $\mathcal{F}_n \uparrow \mathcal{F}$ and the empirical L^2-risk minimizing estimator $\widehat{m}_n = \widehat{m}_n(\mathcal{F}_n)$ is universally strongly consistent (in L^2-sense) for the regression function $m = E(Y \mid X = \cdot)$, i.e. $\int (\widehat{m}_n(x) - m(x))^2 dP^X(x) \to 0$ a.s. for all random variables (X, Y) with $EY^2 < \infty$.*

For the proof of Theorem 2.1 we need the following extension of Theorem 1 of Lugosi and Zeger (1995) which allows us to restrict to bounded Y.

Theorem 2.2 (Truncation method) *Let $\mathcal{F}$ be dense in $L^2(P)$ for all probability measures P on $\mathbb{R}^k$. Assume there exists a sequence $\mathcal{F}_n \uparrow \mathcal{F}$ such that for all bounded r.v.s Y we have*

$$\sup_{f\in\mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{N_i} (Y_{ij} - f(X_{ij}))^2 - \mathbb{E}\left(\sum_{j=1}^N (Y_j - f(X_j))^2 \right) \right| \to 0 \qquad (2.1)$$

then $\widehat{m}_n = \widehat{m}_n(\mathcal{F}_n)$ is strongly consistent for $m = E(Y \mid X = \cdot)$ for all $Y \in L^2$.

Proof: For the proof we use the truncation argument of Lugosi and Zeger (1995). For $L > 0$ denote the truncated random variables $Y^L := T_L \circ Y$, $Y_{ij}^L := T_L \circ Y_{ij}$, where

$$T_L(x) = \begin{cases} L & \text{if } x \ge L \\ x & \text{if } -L < x < L \\ -L & \text{if } x \le -L. \end{cases}$$

Denote the empirical risk estimator for the truncated process by

$$\widehat{m}_n^L := \arg\min_{f\in\mathcal{F}_n} \sum_{i=1}^n \sum_{j=1}^{N_i} (Y_{ij}^L - f(X_{ij}))^2$$

and the related optimal approximations by

$$f_n^* := \arg\min_{f\in\mathcal{F}_n} \mathbb{E}\left(\sum_{j=1}^N (Y_j - f(X_j))^2 \right), \text{ and}$$

$$f_n^{L,*} := \arg\min_{f\in\mathcal{F}_n} \mathbb{E}\left(\sum_{j=1}^N (Y_j^L - f(X_j))^2 \right).$$

Furthermore, define

$$\begin{aligned}
\delta_n &:= d_{L^2}(Y, \widehat{m}_n(X)) - \inf_{f \in \mathcal{F}_n} d_{L^2}(Y, f(X)),\\
\delta_n^L &:= d_{L^2}(Y^L, \widehat{m}_n(X)) - \inf_{f \in \mathcal{F}_n} d_{L^2}(Y^L, f(X)),
\end{aligned}$$

with L^2-distance $d_{L^2} = d_{L^2(P)}$ and the empirical risk measure

$$\Delta_n^L := \sup_{f \in \mathcal{F}_n} \left| \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N_i} (Y_{ij}^L - f(X_{ij}))^2 \right)^{\frac{1}{2}} - \left(\mathbb{E} \left(\sum_{j=1}^{N} (Y_j^L - f(X_j))^2 \right) \right)^{\frac{1}{2}} \right|.$$

In the first step we establish

$$\delta_n \le \delta_n^L + 2 d_{L^2}(Y^L, Y). \tag{2.2}$$

For the proof of (2.2) note that

$$\begin{aligned}
\inf_{f \in \mathcal{F}_n} d_{L^2}(Y^L, f(X)) &\le d_{L^2}(Y^L, f_n^*(X))\\
&\le d_{L^2}(Y, f_n^*(X)) + d_{L^2}(Y^L, Y)\\
&= \inf_{f \in \mathcal{F}_n} d_{L^2}(Y, f(X)) + d_{L^2}(Y^L, Y),
\end{aligned}$$

using independence of N and (X, Y). From the triangle inequality we obtain

$$\begin{aligned}
\delta_n &\le d_{L^2}(Y^L, \widehat{m}_n(X)) + d_{L^2}(Y^L, Y) + d_{L^2}(Y^L, Y) - \inf_{f \in \mathcal{F}_n} d_{L^2}(Y^L, f(X))\\
&= \delta_n^L + 2 d_{L^2}(Y^L, Y).
\end{aligned}$$

In the next step we prove

$$\sqrt{\mathbb{E} N} \delta_n^L \le 2 \Delta_n^L + 2 \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N_i} (Y_{ij}^L - Y_{ij})^2 \right)^{\frac{1}{2}}. \tag{2.3}$$

For the proof of (2.3) note that by the triangle inequality for $i = 1, \ldots, n$, $j = 1, \ldots, N_i$ and $a_{ij}, b_{ij}, c_{ij} \in \mathbb{R}$ holds

$$\left(\sum_{i=1}^{n} \sum_{j=1}^{N_i} (a_{ij} - b_{ij})^2 \right)^{\frac{1}{2}} \le \left(\sum_{i=1}^{n} \sum_{j=1}^{N_i} (a_{ij} - c_{ij})^2 \right)^{\frac{1}{2}} + \left(\sum_{i=1}^{n} \sum_{j=1}^{N_i} (c_{ij} - b_{ij})^2 \right)^{\frac{1}{2}}. \tag{2.4}$$

Using (2.4) and the definition of $d_n^{L,*}$ we obtain

$$
\begin{aligned}
&\sqrt{\mathbb{E}N}\,\delta_n^L \\
&= \sqrt{\mathbb{E}N}\,d_{L^2}(Y^L,\widehat{m}_n(X)) - \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}(Y_{ij}^L-\widehat{m}_n(X_{ij}))^2\right)^{\frac{1}{2}} \\
&\quad + \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}(Y_{ij}^L-\widehat{m}_n(X_{ij}))^2\right)^{\frac{1}{2}} - \sqrt{\mathbb{E}N}\inf_{f\in\mathcal{F}_n} d_{L^2}(Y^L,f(X)) \\
&\le \Delta_n^L + \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}(Y_{ij}^L-Y_{ij})^2\right)^{\frac{1}{2}} + \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}(Y_{ij}-\widehat{m}_n(X_{ij}))^2\right)^{\frac{1}{2}} \\
&\quad - \sqrt{\mathbb{E}N}\inf_{f\in\mathcal{F}_n} d_{L^2}(Y^L,f(X)) \\
&\le \Delta_n^L + \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}(Y_{ij}^L-Y_{ij})^2\right)^{\frac{1}{2}} + \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}(Y_{ij}-f_n^{L,*}(X_{ij}))^2\right)^{\frac{1}{2}} \\
&\quad - \sqrt{\mathbb{E}N}\inf_{f\in\mathcal{F}_n} d_{L^2}(Y^L,f(X)) \\
&\le \Delta_n^L + 2\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}(Y_{ij}^L-Y_{ij})^2\right)^{\frac{1}{2}} + \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}(Y_{ij}^L-f_n^{L,*}(X_{ij}))^2\right)^{\frac{1}{2}} \\
&\quad - \sqrt{\mathbb{E}N}\inf_{f\in\mathcal{F}_n} d_{L^2}(Y^L,f(X)) \\
&\le 2\Delta_n^L + 2\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}(Y_{ij}^L-Y_{ij})^2\right)^{\frac{1}{2}}.
\end{aligned}
$$

(2.2) and (2.3) imply

$$
\delta_n \le \frac{2}{\sqrt{\mathbb{E}N}}\Delta_n^L + \frac{2}{\sqrt{\mathbb{E}N}}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}(Y_{ij}^L-Y_{ij})^2\right)^{\frac{1}{2}} + 2d_{L^2}(Y^L,Y). \quad (2.5)
$$

Y^L is bounded. Therefore, using assumption (2.1) we obtain

$$
\limsup_{n\to\infty}\Delta_n^L = 0 \qquad \text{a.s..}
$$

Relation (2.5) implies using the strong law of large numbers $\limsup_{n\to\infty}\delta_n \le 4d_{L^2}(Y^L,Y)$. The denseness assumption on $\mathcal{F}$ implies $\lim_{n\to\infty}\inf_{f\in\mathcal{F}_n} d_{L^2}(Y, f(X)) = d_{L^2}(Y,\mathbb{E}(Y|X))$. Therefore, from $\lim_{n\to\infty}\delta_n = 0$ a.s. we obtain $\lim_{n\to\infty} d_{L^2}(Y,\widehat{m}_n(X)) = d_{L^2}(Y,\mathbb{E}(Y|X))$ a.s., which yields the strong consistency property stated in Theorem 2.2. For details of this last step see also the proof of Theorem 2.1 in section 3. □

3 A uniform law of large numbers for the empirical risk measure

In order to establish a universal consistency result for the empirical risk estimator it is by Theorem 2.2 sufficient to consider the case where Y is bounded. So let us assume in this section

Assumption B) Y is bounded, i.e. $|Y| \leq L < \infty$ for some constant L.

For $\beta > 0, K \in \mathbb{N}$ we introduce

$$\mathcal{F}(\beta, K) := \left\{ f : \mathbb{R}^k \longrightarrow [-K\beta, K\beta];\ f = \sum_{i=1}^{K} c_i f_i; \right. \qquad \left. \text{with } f_i \in \mathcal{F}_0, |c_i| \leq \beta \right\} \tag{3.1}$$

and the induced class

$$\mathcal{G}(\beta, K, L) := \left\{ g_f : \mathbb{R}^k \times [-L, L] \to [0, (L + K\beta)^2], \right. \qquad \left. g_f(x, y) \to (y - f(x))^2; f \in \mathcal{F}(\beta, K) \right\}. \tag{3.2}$$

The maximal deviation functional as in (2.1) for the class $\mathcal{F}(\beta, K)$ then can be written in the form

$$\begin{aligned} &\Delta_n(\beta, K, L)(\omega) \\ &:= \sup_{f \in \mathcal{F}(\beta, K)} \left| \sum_{i=1}^{n} \sum_{j=1}^{N_i} (Y_{ij} - f(X_{ij}))^2 - \mathbb{E}\left(\sum_{i=1}^{n} \sum_{j=1}^{N} (Y_{ij} - f(X_{ij}))^2 \right) \right| \\ &= \sup_{g \in \mathcal{G}(\beta, K, L)} \left| \sum_{i=1}^{n} \langle g, \Phi_i(\omega) \rangle - \mathbb{E}\left(\sum_{i=1}^{n} \langle g, \Phi_i(\omega) \rangle \right) \right|, \end{aligned} \tag{3.3}$$

with $\langle g, \mu \rangle := \int g \, d\mu$. To estimate $\Delta_n(\beta, K, L)$ we shall make use of a variant of Pollard's inequality due to Ziegler (1994). For the relevant notions from empirical process theory we refer to van der Vaart and Wellner (1996). As usual $\dim_{VC}(\mathcal{F})$ denotes the Vapnik-Cervonenkis dimension of $\mathcal{F}$, $N(\varepsilon, \mathcal{F}, d_{L^2(\mu)})$ denotes the ε-covering number of $\mathcal{F}$ w.r.t. $L^2(\mu)$, $N_2(\varepsilon, \mathcal{Z})$, $D_2(\varepsilon, \mathcal{Z})$ denote the ε-covering resp. ε-packing numbers of $\mathcal{Z} \subset \mathbb{R}^n$ with respect to the Euclidean metric.

Lemma 3.1 *Let $\mathcal{F}$ be a VC-class with majorant F and let $d = \dim_{VC}(\mathcal{F})$. Then there exist some constant $C(d) \geq 0$ such that*

$$N(\varepsilon \|F\|_{L^2(\mu)}, \mathcal{F}, d_{L^2(\mu)}) \leq C(d) \varepsilon^{-4(d-1)} \tag{3.4}$$

for $0 < \varepsilon \leq 1$ and all measures μ with $\|F\|_{L^2(\mu)} < \infty$.

We also make use of the following stability properties of the covering numbers which are variants of those in Pollard (1990, section 5), see also Döhler (1999, 2000).

Lemma 3.2 (Stability properties of covering numbers) *Let $\mathcal{F}$ and $\mathcal{G}$ be two function classes on $\mathbb{R}^k$ and μ be a measure, then:*

a) $N(\varepsilon+\delta, \mathcal{F}\oplus\mathcal{G}, d_{L^2(\mu)}) \;\leq\; N(\varepsilon, \mathcal{F}, d_{L^2(\mu)})N(\delta, \mathcal{G}, d_{L^2(\mu)})$
where $\mathcal{F}\oplus\mathcal{G}$ denotes the sumclass.

b) If $|f|\leq K$ for all $f\in\mathcal{F}$, then

$$N(\varepsilon, \mathcal{F}^2, d_{L^2(\mu)}) \;\leq\; N\left(\frac{\varepsilon}{2K}, \mathcal{F}, d_{L^2(\mu)}\right)$$

where $\mathcal{F}^2=\{f^2; f\in\mathcal{F}\}$.

c) If μ is a finite measure and $|f|\leq K$ for all $f\in\mathcal{F}$, then for any $a>0$

$$N(\varepsilon, [-a,a]\cdot\mathcal{F}, d_{L^2(\mu)}) \;\leq\; N\left(\frac{\varepsilon}{2a}, \mathcal{F}, d_{L^2(\mu)}\right)\frac{4aK\|1\|_{L^2(\mu)}}{\varepsilon}.$$

The empirical risk measure $\Delta_n(\beta,K,L)$ involves the following random set $\mathcal{Z}_n$ of
points in $\mathbb{R}^n$:

$$\mathcal{Z}_n(\beta,K,L)(\omega) := \Big\{(\langle g,\Phi_1(\omega)\rangle,\ldots,\langle g,\Phi_n(\omega)\rangle)\,\Big|\, g\in\mathcal{G}(\beta,K,L)\Big\}\subset\mathbb{R}^n. \quad (3.5)$$

Let us denote the corresponding random entropy integral by

$$J_n(\omega) \;:=\; 9\int_0^{\delta_n(\omega)}\sqrt{\log D_2(\varepsilon,\mathcal{Z}_n(\beta,K,L)(\omega))}\,d\varepsilon, \quad (3.6)$$

where $\delta_n(\omega):=\sup\{\|y\|_2 \mid y\in\mathcal{Z}_n(\beta,K,L)(\omega)\}$ and $D_2(\varepsilon,\mathcal{Z}_n)$ denotes the ε-packing number of $\mathcal{Z}_n$ in Euclidean metric. In the following we use the notation $J_n=J_n(\beta,K)=J_n(\beta,K,L)$ depending on the parameters of interest.

Theorem 3.3 (Uniform law for the empirical risk measure, bounded case) *Let $1\leq p,q<\infty$, $q+1<p$ let $L>0$ be fixed and assume that for $K\in\mathbb{N}$, $\beta>0$ and any bounded random variable $|Y|\leq L$ holds*

$$EJ_n^p(\beta,K)=O(n^q). \quad (3.7)$$

Then there exist sequences $K_n\uparrow\infty, \beta_n\uparrow\infty$ such that for any bounded r.v. $|Y|\leq L$ holds

$$\lim_{n\to\infty}\frac{1}{n}\Delta_n(\beta_n,K_n,L)=0 \qquad \text{a.s.}$$

Proof: In the first step of the proof we establish: There exists $\delta > 0$ and $\varepsilon_n \downarrow 0$, such that for any $|Y| \le L$ and $K \in \mathbb{N}$, $\beta > 0$ holds

$$P\left(\frac{1}{n}\Delta_n(\beta, K, L) > \varepsilon_n\right) \le \frac{1}{n^{1+\delta}}. \tag{3.8}$$

for all $n \ge N_0 = N_0(\beta, K, L) \in \mathbb{N}$. Define $\delta' := \frac{1}{2}(p - (q+1))$ and $\varepsilon_n := \left(\frac{1}{n}\right)^{\frac{1}{p}(p-(q+1+\delta'))}$ and note that by assumption $\delta' > 0$ and $q + 1 + \delta' < p$. We apply Pollard's maximal inequality for empirical processes indexed by function classes (cp. Pollard (1990, (7.5)), to obtain the estimate

$$P\left(\frac{1}{n}\Delta_n(\beta, K, L) > \varepsilon_n\right) \le \frac{(2C_p)^p\,\mathbb{E}J_n^p(\beta, K, L)}{(n\varepsilon_n)^p}$$

with some constant $C_p > 0$. Therefore, by the growth assumption on the random entropy integral there exist $N_0 = N_o(\beta, K, L) \in \mathbb{N}$ and $C = C(\beta, K, L) > 0$ such that for $n \ge N_0$ and with $\delta = \frac{\delta'}{2}$

$$\begin{aligned} P\left(\frac{1}{n}\Delta_n(\beta, K, L) > \varepsilon_n\right) &\le \frac{(2C_p)^p C(\beta, K, L) n^q}{n^p n^{q+1+\delta'-p}} \\ &= \frac{(2C_p)^p C(\beta, K, L)}{n^{1+\delta'}} \\ &\le \frac{1}{n^{1+\delta}}. \end{aligned} \tag{3.9}$$

Let $N_{ij} = N_0(i, i, j)$ and assume w.l.g. $N_{i1} < N_{i2} < \dots$ and define $N_i := N_{ii}$. Assume w.l.g. $N_i \uparrow \infty$ and define for $n \in \{N_p, \dots, N_{p+1} - 1\}$, $K_n := \beta_n := p$. Then we obtain for $n \ge N_L$

$$P\left(\frac{1}{n}\Delta_n(\beta_n, K_n, L) > \varepsilon_n\right) \le \frac{1}{n^{1+\delta}}, \tag{3.10}$$

since $n \ge N_L$ implies $n \in \{N_{p_0}, \dots, N_{p_0+1} - 1\}$ for some $p_0 \ge L$, i.e. $K_n = \beta_n = p_0$, and $n \ge N_{p_0} \ge N_{p_0 L} = N_0(p_0, p_0, L)$ and so (3.10) follows from (3.9).

Define $A_n = \{\frac{1}{n}\Delta_n(\beta_n, K_n, L) > \varepsilon_n\}$, then $P(A_n) \le \frac{1}{n^{1+\delta}}$ for n sufficiently large. This implies the statement of the theorem by application of the Borel-Cantelli Lemma. □

For the application of Theorem 3.3 we need to establish an upper bound for the random entropy integral in order to verify condition (3.7).

Proposition 3.4 *Let $1 \le p < \infty$ and $EN^{\frac{3}{2}p} < \infty$. Then*

$$EJ_n^p(\beta, K) = O(n^{\frac{3}{4}p}) \tag{3.11}$$

for any bounded r.v. Y, $|Y| \le L$ and any $K \in \mathbb{N}$ and $\beta > 0$.

Proof: Define

$$F_n(\beta,K,L)(\omega) := \left(\langle (L+K\beta)^2, \Phi_1(\omega)\rangle, \ldots, \langle (L+K\beta)^2, \Phi_n(\omega)\rangle\right) \in \mathbb{R}^n.$$

Then

$$\begin{aligned} & J_n(\beta,K,L)(\omega) && (3.12)\\ \leq\ & C|F_n(\beta,K,L)(\omega)| \int_0^1 \sqrt{\log N_2(\varepsilon |F_n(\beta,K,L)(\omega)|, \mathcal{Z}_n(\beta,K,L)(\omega))}\, d\varepsilon. \end{aligned}$$

For the proof of (3.12) note that for all $z \in \mathcal{Z}_n(\beta,K,L)(\omega)$ and all $1 \leq i \leq n$ $|z_i| \leq (F_n(\beta,K,L)(\omega))_i$, i.e. F_n majorizes $\mathcal{Z}_n$. As $F_n(\beta,K,L)(\omega) = (L+K\beta)^2(N_1(\omega),\ldots,N_n(\omega))$ we obtain $|F_n(\beta,K,L)(\omega)| = (L+K\beta)^2\sqrt{\sum\limits_{i=1}^n N_i^2(\omega)}$ and $\|z\|_2 \leq |F_n(\beta,K,L)(\omega)|$. By a substitution as in Pollard (1990, (7.7)) we obtain

$$\begin{aligned} & J_n(\beta,K,L)(\omega)\\ \leq\ & 9\,|F_n(\beta,K,L)(\omega)| \int_0^1 \sqrt{\log D_2(\varepsilon |F_n(\beta,K,L)(\omega)|, \mathcal{Z}_n(\beta,K,L)(\omega))}\, d\varepsilon. \end{aligned}$$

Using that the packing number D_2 is related to the covering number N_2 by $D_2(\varepsilon,\mathcal{F}) \leq N_2\left(\frac{\varepsilon}{2},\mathcal{F}\right)$ we obtain (3.12).

In the next step we prove

$$\begin{aligned} & N_2\left(\varepsilon\,|F_n(\beta,K,L)(\omega)|, \mathcal{Z}_n(\beta,K,L)(\omega)\right) && (3.13)\\ \leq\ & C(\beta,K)\left(\frac{1}{\varepsilon}\right)^{K(V_0+1)}\left(\sqrt{\textstyle\sum_{i=1}^n N_i(\omega)}\right)^{K(V_0+1)} \end{aligned}$$

with $V_0 = 4(\dim_{VC}(\mathcal{F}_0)-1)$, $C(\beta,K) := C(d)^k 2^{KV_0}(\beta K)^{K(K+V_0)}$, and $C(d) > 0$ a universal constant.

For the proof define $F^2 := (L+K\beta)^2$ and $M(\omega) := (\langle F^2, \sum_{i=1}^n \Phi_i(\omega)\rangle)^{\frac{1}{2}}$. We first establish

$$\begin{aligned} & N_2(\varepsilon|F_n(\beta,K,L)(\omega)|, \mathcal{Z}_n(\beta,K,L)(\omega)) && (3.14)\\ \leq\ & N\left(\frac{\varepsilon M(\omega)}{\sqrt{\sum_{i=1}^n N_i(\omega)}}, \mathcal{G}(\beta,K,L), d_{L^2(\sum_{i=1}^n \Phi_i(\omega))}\right) \end{aligned}$$

To this purpose define $r := N(\varepsilon M(\omega), \mathcal{G}(\beta,K,L), d_{L^2(\sum_{i=1}^n \Phi_i(\omega))})$, and let $g_1,\ldots,g_r \in \mathcal{G}(\beta,K,L)$ be an $\varepsilon M(\omega)$ approximation of $\mathcal{G}(\beta,K,L)$ w.r.t. the $d_{L^2(\sum_{i=1}^n \Phi_i(\omega))}$-metric. Define $y_i := (\langle g_i, \Phi_1(\omega)\rangle, \ldots, \langle g_i, \Phi_n(\omega)\rangle)$. Then for all $z \in \mathcal{Z}_n(\beta,K,L)(\omega)$ corresponding to some $g \in \mathcal{G}(\beta,K,L)$ there exist some y_i with

$$d_2(y,y_i) \ \leq\ \varepsilon|F_n(\beta,K,L)(\omega)|\sqrt{\textstyle\sum_{i=1}^n N_i(\omega)}.$$

This inequality results from an application of Jensens's inequality and the approximation of g by g_i. Together we obtain (3.14).

Define next $\Phi_i'(\omega) := \sum_{j=1}^{N_i(\omega)} \varepsilon_{X_{ij}}$. From Pollard's inequality (3.4) we obtain

$$N\left(\varepsilon\sqrt{\sum_{i=1}^n N_i(\omega)}, \mathcal{F}_0, d_{L^2(\sum_{i=1}^n \Phi_i'(\omega))}\right) \;\leq\; A\left(\frac{1}{\varepsilon}\right)^{4(d_0-1)}, \tag{3.15}$$

with $d_0 = \dim_{VC}(\mathcal{F}_0)$. Denote

$$\begin{aligned}\mathcal{G}'(\beta,K,L) \;:=\; \big\{g_f' : \mathbb{R}^k \times [-L,L] &\longrightarrow [-(L+K\beta),(L+K\beta)],\\ (x,y) &\longmapsto y - f(x) \,|\, f \in \mathcal{F}(\beta,K)\big\}.\end{aligned}$$

Then $\mathcal{G}(\beta,K,L) \subset (\mathcal{G}'(\beta,K,L))^2$ and $\mathcal{F}(\beta,K) \subset \bigoplus_{i=1}^K[-\beta,\beta]\cdot\mathcal{F}_0$. Using the stability properties of the covering numbers in Lemma 3.2 these properties imply omitting the arguments in the following formulas:

$$\begin{aligned}A &= N\left(\frac{\varepsilon M}{\sqrt{\sum N_i}}, \mathcal{G}, d_{L^2(\sum\Phi_i)}\right)\\ &= N\left(\varepsilon F, \mathcal{G}, d_{L^2(\sum\Phi_i)}\right) && \text{as } M = F\sqrt{\sum N_i}\\ &\leq N\left(\frac{\varepsilon F}{2(L+K\beta))}, \mathcal{G}', d_{L^2(\sum\Phi_i)}\right) && \text{by Lemma 3.2, b) using}\\ & && |g_f'| \leq L+K\beta\\ &= N\left(\frac{\varepsilon}{2}, \mathcal{G}', d_{L^2(\sum\Phi_i)}\right) \;\leq\; N\left(\frac{\varepsilon}{2}, \mathcal{F}, d_{L^2(\sum\Phi_i)}\right).\end{aligned}$$

Since $\mathcal{F} \subset \oplus_{i=1}^K[-\beta,\beta]\cdot\mathcal{F}_0$ we obtain from Lemma 3.2, a) and c)

$$\begin{aligned}A &\leq \left[N\left(\frac{\varepsilon}{2K}, [-\beta,\beta]\cdot\mathcal{F}_0, d_{L^2(\sum\Phi_i')}\right)\right]^K\\ &\leq \left[N\left(\frac{\varepsilon}{2K\beta}, \mathcal{F}_0, d_{L^2(\sum\Phi_i')}\right)\frac{8\beta K\sqrt{\sum N_i}}{\varepsilon}\right]^K\\ &\leq \left[N\left(\frac{\varepsilon}{2K\beta\sqrt{\sum N_i}}\sqrt{\sum N_i}, \mathcal{F}_0, d_{L^2(\sum\Phi_i')}\right)\frac{8\beta K\sqrt{\sum N_i}}{\varepsilon}\right]^K.\end{aligned}$$

This implies by Pollard's inequality in (3.15)

$$\begin{aligned}A &\leq \left[C(d)\left(\frac{\varepsilon}{2K\beta\sqrt{\sum N_i}}\right)^{-V_0}\frac{8\beta K\sqrt{\sum N_i}}{\varepsilon}\right]^K\\ &= C(d)^K 2^{KV_0}(\beta K)^{K(K+V_0)}\left(\frac{1}{\varepsilon}\right)^{K(V_0+1)}\left(\sum N_i\right)^{\frac{K(V_0+1)}{2}}\end{aligned}$$

and thus we obtain (3.13).

Define

$$R_1(\beta, K) := \int_0^1 \sqrt{\log\left[C(\beta,K)\left(\frac{1}{\varepsilon}\right)^{K(V_0+1)}\right]}\, d\varepsilon \tag{3.16}$$

and $R_2(K) := \sqrt{K(V_0+1)}$ then from (3.13) and (3.12) and using some simple calculations (see Döhler and Rüschendorf (2000, proof of Proposition 15) we obtain the estimate

$$\begin{aligned} & J_n(\beta, K, L)(\omega) \\ \leq\ & C|F_n(\beta,K)(\omega)|\left[R_1(\beta,K,L) + R_2(K)\sqrt[4]{\sum_{i=1}^n N_i(\omega)}\,\right]. \end{aligned} \tag{3.17}$$

With $Z_n(\omega) := \sqrt{\sum_{i=1}^n N_i^2(\omega)}$ this yields

$$J_n(\beta,K,L)(\omega) \leq C_1(\beta,K,L) Z_n(\omega)\left(1 + C_2(\beta,K,L) Z_n(\omega)^{\frac{1}{2}}\right). \tag{3.18}$$

Define $p' := [p] \in \mathbb{N}$ and $\delta := p - p'$, then

$$\begin{aligned} J_n^p(\beta,K,L)(\omega) \ &\leq\ C_1(\beta,K,L) Z_n^p(\omega)\left[1 + C_2(\beta,K,L)Z_n(\omega)^{\frac{1}{2}}\right]^{p'} \\ &\qquad \left[1 + C_2(\beta,K,L)Z_n(\omega)^{\frac{1}{2}}\right]^{\delta} \\ &\leq\ C_1(\beta,K,L) Z_n^p(\omega)\left[1 + C_2(\beta,K,L)Z_n(\omega)^{\frac{1}{2}}\right]^{p'} \\ &\qquad \left[1 + C_2^{\delta}(\beta,K,L)Z_n(\omega)^{\frac{\delta}{2}}\right], \end{aligned}$$

using $(1+x)^\delta \leq 1 + x^\delta$ for $x \geq 0$. Therefore, by the binomial formula we obtain

$$\begin{aligned} \mathbb{E} J_n^p(\beta,K,L) \ &\leq\ C_1(\beta,K,L)\sum_{k=0}^{p'}\binom{p'}{k} C_2^k(\beta,K,L) \\ &\qquad \mathbb{E}(Z_n^p Z_n^{\frac{k}{2}}(1 + C_3(\beta,K,L)Z_n(\omega)^{\frac{\delta}{2}}) \\ &\leq\ C_4(\beta,K,L)\mathbb{E} Z_n^{p+\frac{p'}{2}+\frac{\delta}{2}} \\ &\leq\ C_4(\beta,K,L)\mathbb{E} Z_n^{\frac{3p}{2}}. \end{aligned}$$

As $\mathbb{E} N^{\frac{3p}{2}} < \infty$ this implies by a wellknown moment estimate $\mathbb{E} Z_n^{\frac{3p}{2}} = \mathbb{E}\left(\sum_{i=1}^n N_i^2(\omega)\right)^{\frac{3p}{4}} = O(n^{\frac{3p}{4}})$ which yields the statement of the theorem. □

Note that the various constants $C_i(\beta, K.L)$ in the proof of Proposition 3.4 can be made explicit.

For the application of Proposition 3.4 to Theorem 3.3 we obtain with $q = \frac{3}{4}p$ the condition $q + 1 = \frac{3}{4}p + 1 < p$, i.e. $p > 4$ and, therefore, we need the assumption $EN^r < \infty$ where $r = \frac{3}{2}p > 6$. This leads to the moment assumption on N in Theorem 2.1.

For the proof of Theorem 2.1 we shall make use of the usual decomposition of the error into approximation and estimation error (for related decomposition results see e.g. Krzyzak, Linder, and Lugosi (1996)).

Lemma 3.5 *Define $J_2(f) = E(Y - f(X))^2 EN$ and the empirical counterpart $J_{2,n}(f) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^{N_i}(Y_{ij} - f(X_{ij}))^2$. Then*

a) $J_2(f) - J_2(m) = (EN)\int (f - m)^2 dP_X$

b) $J_2(\widehat{m}_n) - J_2(m) = (J_2(\widehat{m}_n) - \inf_{f\in\mathcal{F}_n} J_2(f)) + (\inf_{f\in\mathcal{F}_n} J_2(f) - J_2(m))$

c) $|J_2(\widehat{m}_n) - \inf_{f\in\mathcal{F}_n} J_2(f)| \leq 2\sup_{f\in\mathcal{F}_n} |J_{2,n}(f) - J_2(f)|$.

Proof: a) and b) are obvious. For the proof of c) note that

$$\begin{aligned} J_2(\widehat{m}_n) - \inf_{f\in\mathcal{F}_n} J_2(f)) &= J_2(\widehat{m}_n) - J_{2,n}(\widehat{m}_n) + J_{2,n}(\widehat{m}_n) - \inf_{f\in\mathcal{F}_n} J_2(f) \\ &= J_2(\widehat{m}_n) - J_{2,n}(\widehat{m}_n) + \sup_{f\in\mathcal{F}_n}(J_{2,n}(\widehat{m}_n) - J_2(f)) \end{aligned}$$

Therefore,

$$\begin{aligned} |J_2(\widehat{m}_n) - \inf_{f\in\mathcal{F}_n} J_2(f)| &\leq |J_2(\widehat{m}_n) - J_{2,n}(\widehat{m}_n)| + \sup_{f\in\mathcal{F}_n} |J_{2,n}(f) - J_2(f)| \\ &\leq 2\sup_{f\in\mathcal{F}_n} |J_{2,n}(f) - J_2(f)|. \end{aligned}$$ □

Proof of Theorem 2.1: The bounds on the random entropy numbers in Proposition 3.4 together with Theorem 3.3 imply convergence of the empirical risk measure $\sup_{f\in\mathcal{F}_n} |J_{2,n}(f) - J_2(f)| \to 0$ for bounded Y. Theorem 2.2 implies that this convergence extends to all $Y \in L^2(P)$. By the denseness assumption on the sieve class $\mathcal{F}_n$ also the approximation term in Lemma 3.5 b) converges to zero, $\inf_{f\in\mathcal{F}_n} J_2(f) - J_2(m) \to 0$. So from Lemma 3.5 b) and c) we obtain $J_2(\widehat{m}_n) - J_2(m)) = EN\int(\widehat{m}_n(x) - m(x))^2 dP_X(x) \to 0$ a.s. for any $Y \in L^2(P)$. □

Remark 3.6 *The conditions for the universal consistency result are stated in quite weak form. Stone (1977) first showed the existence of weakly universal consistent estimators. Since then many results on weakly and strongly universal consistent estimators have been obtained by various methods. For a list of references see Devroye, Györfi, Krzyzak, and Lugosi (1994). It is known that no nontrivial upper bounds for the rate of convergence can be obtained under this general type of assumptions (see Lugosi and Zeger (1995)).*

Under additional assumptions on the moments of Y and on the approximation properties of the class $\mathcal{F}$ the estimates in this paper can be made explicit to obtain convergence rates and the order of K_n and β_n. In the related paper of Döhler (1999) this is made explicit in the problem of estimation of the hazard function in censoring models. An essential step in this approach is an exponential bound in the maximal inequality (instead of the polynomial bound used here). This exponential bound holds true if the random entropy integral is bounded by a constant (see Pollard (1990, (7.3)).

4 Estimation of the intensity function of a Poisson process

For the related problem of estimating the intensity function of a Poisson process it is more natural to switch from empirical L^2-risk minimization to sieved maximum likelihood estimation. Various parametric and nonparametric estimators of the intensity have been already studied in the literature (see the references in our introduction). The empirical risk minimization method applies in much the same way also to the maximum likelihood risk function. For the fixed sample case and the estimation of the hazard function in censoring models this has been worked out in Döhler (1999). Since many technical details are similar to those in the first part of this paper we give only a sketch of the proof of the strong consistency of the sieved maximum likelihood estimator for general sieve classes satisfying assumption A.

Let Φ be a Poisson process on $[0,1]^k$ with finite intensity measure μ and intensity function $\widetilde{\alpha}_0$ then $\mu(A) = E\Phi(A) = \int_A \widetilde{\alpha}_0 d\lambda$. Let $\Phi_1, \Phi_2, \ldots, \Phi_n$ be an iid sample, $\Phi_i \sim \Phi$ then the log-likelihood of $P_{\widetilde{\alpha}}$ w.r.t. $P_{\widetilde{\alpha}_0}$ is given by

$$\begin{aligned} L_n(\widetilde{\alpha}) &= n\int(1-\widetilde{\alpha})d\lambda + \int(\log\widetilde{\alpha})d\sum_{i=1}^n \Phi_i. \text{ Define} \\ \Lambda(\widetilde{\alpha}) &= EL_1(\widetilde{\alpha}) = \int(1-\widetilde{\alpha}+\widetilde{\alpha}_0\log\widetilde{\alpha})d\lambda. \end{aligned} \tag{4.1}$$

In order to avoid technical problems with the logarithm let $\widetilde{\alpha} = \exp(\alpha)$, $\widetilde{\alpha}_0 = \exp(\alpha_0)$, and so we obtain (using the notation $L_n(\widetilde{\alpha}) =: L_n(\alpha)$)

$$\frac{1}{n}L_n(\alpha) - \Lambda(\alpha) = \frac{1}{n}\int \alpha d\left(\sum_{i=1}^n \Phi_i\right) - \int \alpha\exp\alpha_0 d\lambda. \tag{4.2}$$

With respect to the function-class $\mathcal{F}(\beta, K)$ as defined in (3.1) the maximal deviation is given by

$$\Delta_n(\beta, K) := \sup_{g\in\mathcal{G}(\beta,K)}\left|\sum_{i=1}^n\langle g,\Phi_i\rangle - E\sum_{i=1}^n\langle g,\Phi_i\rangle\right| \tag{4.3}$$

where $\langle g, \mu \rangle = \int g d\mu$ and $\mathcal{G}(\beta, K) = \exp(\mathcal{F}(\beta, K))$. The random point set $\mathcal{Z}_n(\beta, K)$ and the random entropy integral $J_n(\beta, K)$ then are of a similar structure as in Section 3 (see (3.5), (3.6)). To estimate $J_n(\beta, K)$ we need estimates for the covering number resp. packing number $D_2(\varepsilon, \mathcal{Z}_n(\beta, K))$. These estimates are obtained similar as in the proof of Proposition 3.4 under the corresponding assumptions on the Poisson process Φ. As consequence we obtain a maximal-inequality as in the proof of Theorem 3.3. Applying the Borel-Cantelli lemma this yields the corresponding uniform law. Thus we can find suitable sequences $\beta_n \uparrow \infty$, $K_n \uparrow \infty$ such that with the corresponding sieve $\mathcal{F}_n = \mathcal{F}(\beta_n, K_n)$, $\mathcal{F}_n \uparrow \mathcal{F}$, and

$$\sup_{\alpha \in \mathcal{F}_n} \left| \frac{1}{n} L_n(\alpha) - \Lambda(\alpha) \right| \to 0 \quad \text{a.s.} \tag{4.4}$$

In consequence we obtain as in Döhler (1999) that the sieved maximum-likelihood estimator $\widehat{\alpha}_n$ converges to α_0 with respect to the statistical distance $\Lambda(\alpha) - \Lambda(\alpha_0)$ (the Kullback-Leibler distance), i.e.

$$\Lambda(\widehat{\alpha}_n) - \Lambda(\alpha_0) \to 0 \quad \text{a.s.} \tag{4.5}$$

Using (4.1) the Λ-distance has the representation

$$|\Lambda(\alpha_0) - \Lambda(\alpha)| = \Lambda(\alpha_0) - \Lambda(\alpha) = \int G(\alpha - \alpha_0) \exp(\alpha_0) d\lambda\!\!\lambda \tag{4.6}$$

with $G(y) := \exp(y) - (1+y)$ (see Döhler (1999)). This representation implies L^1-convergence

$$\int |\widehat{\alpha}_n - \alpha_0| \exp(\alpha_0) d\lambda\!\!\lambda \to 0 \quad \text{a.s.} \tag{4.7}$$

by some simple estimates of G.

As consequence therefore we obtain strong consistency in L^1-sense of the sieved ML-estimator for the intensity function using the parametrization of the intensity as in (4.2).

Theorem 4.1 (Strong consistency of the ML-estimator) *Let (Φ_i) be an iid sample of Poisson processes on $[0,1]^k$ with finite intensity parametrized as in (4.2). Let $\mathcal{F}$ satisfy assumption A. Then there exist sequences $\beta_n \uparrow \infty$, $K_n \uparrow \infty$ such that the ML-estimator $(\widehat{\alpha}_n)$ based on the sieve $\mathcal{F}_n = \mathcal{F}(\beta_n, K_n)$ is strongly consistent for the intensity function α_0 in L^1-sense i.e.* $\int |\widehat{\alpha}_n - \alpha_0| \exp \alpha_0 d\lambda\!\!\lambda \to 0$ *a.s.*

Acknowledgement The authors thank two reviewers for several useful and interesting comments on the subject and presentation of the paper.

References

Devroye, L., L. Györfi, A. Krzyzak, and G. Lugosi (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Stat. 22*, 1371–1385.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A probabilistic theory of pattern recognition.*, Volume 31 of *Applications of Mathematics.* Springer.

Döhler, S. (1999). Consistent hazard regression estimation by sieved maximum likelihood estimators. In *Proceedings of* Conference on limit theorems *in Balatonlelle.* To appear.

Döhler, S. (2000). *Empirische Risiko-Minimierung bei zensierten Daten.* Dissertation, University of Freiburg.

Döhler, S. and L. Rüschendorf (2000). Adaptive estimation of hazard functions. In *Probability and Mathematical Statistics.* To appear.

Györfi, L., M. Kohler, A. Krzyzak, and H. Walk (2002). *A Distribution Free Theory of Nonparametric Regression.* Springer Series in Statistics.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks 4*, 251–257.

Karr, A. (1991). *Point Processes and their Statistical Inference.* Marcel Dekker.

Kim, W.-C. and J.-Y. Koo (2000). Inhomogeneous Poisson intensity via information projections onto wavelets subspaces. Preprint.

Kohler, M. (1997). On the universal consistency of a least squares spline regression estimator. *Mathem. Methods Statistics 6*, 349–364.

Kohler, M. (2000). Universal consistency of local polynomial kernel regression estimates. Preprint.

Krzyzak, A. and T. Linder (1998). Radial basis function networks and complexity regularization in function learning. *Neural Networks 9*, 247–256.

Krzyzak, A., T. Linder, and G. Lugosi (1996). Nonparametric estimation and classification using radial basis function nets and empirical risk minimization. *IEEE Transactions on Neural Networks 7*(2), 475–487.

Kutoyants, Y. A. (1998). *Statistical Inference for Spatial Poisson Processes*, Volume 134 of *Lecture Notes in Statistics.* Springer.

Liese, F. (1990). Estimation of intensity measures of Poisson point processes. In *Proceedings of the 11th Prague Conference*, pp. 121–139.

Liese, F. and K. Ziegler (1999). A note on empirical process methods in the theory of point processes. *Scand. Journal Statist. 26*, 533–537.

Lugosi, G. and K. Zeger (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory 41*(3), 677–687.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* Institute of Mathematical Statistics.

Reynaud-Bouret, P. (2001). Concentration inequalities for inhomogeneous Poisson processes and adaptive estimation of the intensity. Preprint.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Stat. Theory Appl. 9*, 65–78.

Rüschendorf, L. (1989). Inference for random sampling processes. *Stoch. Proc. Appl. 32*, 129–140.

Stone, C. J. (1977). Consistent nonparametric regression. Discussion. *Ann. Stat. 5*, 595–645.

van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes.* Springer.

Vapnik, V. and A. Chervonenkis (1974). *The Theory of Pattern Recognition.* Nauka, Moscow.

Ziegler, K. (1994). *On functional central limit theorems and uniform laws of large numbers for sums of independent processes.* Dissertation, University of Munich.

University of Freiburg
Institute for Mathematical Stochastics
Eckerstr. 1
79104 Freiburg
Germany

ruschen@stochastik.uni-freiburg.de