Available online at www.sciencedirect.com

ScienceDirect

Journal of Multivariate Analysis 98 (2007) 1072 – 1094

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Asymptotic confidence intervals for Poisson regression

Michael Kohler[a], Adam Krzyżak[b,*]

[a] *Fachrichtung 6.1-Mathematik, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany*

[b] *Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Que., Canada H3G 1M8*

Received 16 August 2006

Available online 20 September 2006

Abstract

Let (X, Y) be a $\mathbb{R}^d \times \mathbb{N}_0$-valued random vector where the conditional distribution of Y given $X = x$ is a Poisson distribution with mean $m(x)$. We estimate m by a local polynomial kernel estimate defined by maximizing a localized log-likelihood function. We use this estimate of $m(x)$ to estimate the conditional distribution of Y given $X = x$ by a corresponding Poisson distribution and to construct confidence intervals of level α of Y given $X = x$. Under mild regularity conditions on $m(x)$ and on the distribution of X we show strong convergence of the integrated L_1 distance between Poisson distribution and its estimate. We also demonstrate that the corresponding confidence interval has asymptotically (i.e., for sample size tending to infinity) level α, and that the probability that the length of this confidence interval deviates from the optimal length by more than one converges to zero with the number of samples tending to infinity.

AMS 1991 subject classification: 62G08; 62G15; 62H12

Keywords: Poisson regression; Local polynomial kernel estimate; Confidence interval

1. Introduction

Let (X, Y) be a $\mathbb{R}^d \times \mathbb{R}$-valued random variable. In regression analysis the dependency of the value of Y on the value of X is studied, e.g., by considering the so-called regression function $m(x) = \mathbf{E}\{Y|X = x\}$. Usually in applications there is little or no a priori knowledge on the structure of m and, therefore, nonparametric methods for analyzing m are of interest. For a general introduction to nonparametric regression see, e.g., [12] and the literature cited therein. In this paper we are interested in the special case that Y takes on with probability one only values

* Corresponding author. Fax: +1 514 848 2830.

E-mail addresses: kohler@math.uni-sb.de (M. Kohler), krzyzak@cs.concordia.ca (A. Krzyżak).

0047-259X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.

doi:10.1016/j.jmva.2006.07.006

in the set of nonnegative integers $\mathbb{N}_0$, and we assume that the conditional distribution of Y given $X = x$ is a Poisson distribution, i.e., we assume

$$\mathbf{P}\{Y = y|X = x\} = \frac{m(x)^y}{y!} \cdot e^{-m(x)} \quad (y \in \mathbb{N}_0, x \in \mathbb{R}^d).$$

In case of a linear function m this is the well-known generalized linear model (cf. [26]) with Poisson likelihood. In the sequel we do not want to make any parametric assumption on m. In this situation we want to use the observed value of X to make some inference about the value of Y, in particular we are interested in constructing confidence intervals for Y given $X = x$.

To do this we assume that a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of the distribution of (X, Y) is given, where $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots$ are independent and identically distributed. In a first step we use the given data

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

to construct an estimate $m_n(x) = m_n(x, \mathcal{D}_n)$ of $m(x)$ and estimate the above conditional probabilities of $Y = y$ given $X = x$ by

$$\widehat{\mathbf{P}}_n\{Y = y|X = x\} = \frac{m_n(x)^y}{y!} \cdot e^{-m_n(x)}. \tag{1}$$

Of course, any of the standard nonparametric regression estimates (like local polynomial kernel estimates, least-squares estimates, or smoothing spline estimates) could be used to estimate the regression function m at this point. However, we are not so much interested in good estimates of m but instead in good estimates of $\mathbf{P}\{Y = y|X = x\}$. Our main aim is to construct estimates such that the integrated L_1 distance between $\mathbf{P}\{Y = y|X = x\}$ and $\widehat{\mathbf{P}}_n\{Y = y|X = x\}$ converges to zero. Since convergence of the L_1 distance between densities to zero is equivalent to convergence to zero of the total variation distance between the corresponding distributions (cf., e.g., [7]), this automatically implies that the level of confidence regions of Y given $X = x$ based on $\widehat{\mathbf{P}}_n\{Y = y|X = x\}$ converges in the average and for sample sizes tending to infinity to the nominal value (cf. Corollary 1 below).

We define regression estimates with this property similarly to [10] by maximizing a localized log-likelihood function with respect to polynomials. This kind of estimate can be considered as an adaptation of the famous local polynomial kernel regression estimate (cf., e.g., [11]) to Poisson regression. The main result of this paper is that we show (under some mild conditions on the underlying distribution) almost sure convergence to zero of the integrated L_1 distance between $\mathbf{P}\{Y = y|X = x\}$ and its estimate (1).

A number of papers have been devoted to Poisson regression and its applications. Automatic methods for the choice of the bandwidth of the Nadaraya–Watson kernel estimate (cf. [27,36]) in Poisson regression have been investigated in [5,15], while in the first paper, in addition, the estimation of a direction vector in a single index model is considered. The Nadaraya–Watson kernel estimate can be also defined as localized log-likelihood estimate provided polynomials of degree zero are used. Related penalized log-likelihood estimates have been investigated (in particular in view of automatic choice of the parameters) in [29,38]. For the related local maximum likelihood estimates the choice of the bandwidth was investigated in [10] in particular in the context of nonparametric logistic regression. Cross-validation of deviance in generalized linear models was discussed in [16]. The amount of smoothing for Poisson intensity reconstructions in medical computed tomography was studied in [30]. Poisson regression estimation has been applied by

Kolaczyk [20] to estimation of X-ray and γ-ray burst intensity maps in astrophysics using Haar wavelets. Regression estimation minimizing roughness-penalized Poisson likelihood was applied to tomographic image reconstruction by La Riviere and Pan [22]. Nowak and Kolaczyk [28] applied Bayesian multiscale methods to Poisson inverse problem.

In the proofs of the main results we use ideas developed in empirical process theory for the analysis of local-likelihood density estimates as described in Chapter 4 of [33] (see also [23,24,3,4]) and apply them to Poisson regression. The proofs are tailor-made for the proposed estimate.

The definition of the estimate is given in Section 2, the main results are described in Section 3, an outline of the proof of the main theorem is given in Section 4, and Section 5 contains the proofs.

2. Definition of the estimate

We define the estimate by maximizing a localized version of the log-likelihood-function

$$L(\theta) = \sum_{i=1}^{n} \log\left(\frac{\theta^{Y_i}}{Y_i!} \cdot e^{-\theta}\right)$$

of a Poisson distribution and throughout the paper log is a natural logarithm. To define such a localized log-likelihood function, let $K : \mathbb{R}^d \to \mathbb{R}$ be a so-called kernel function, e.g., $K(u) = 1_{\{\|u\| \leqslant 1\}}$ (where 1_A denotes the indicator function of a set A and $\|u\|$ is the Euclidean norm of $u \in \mathbb{R}^d$), and let $h_n > 0$ be the so-called bandwidth, which we will choose later such that

$$h_n \to 0 \quad (n \to \infty).$$

The localized log-likelihood of a function $g : \mathbb{R}^d \to \mathbb{R}_+$ at point $x \in \mathbb{R}^d$ is defined by

$$L_{\text{loc}}(g|x) = \sum_{i=1}^{n} \log\left(\frac{g(X_i)^{Y_i}}{Y_i!} \cdot e^{-g(X_i)}\right) \cdot K\left(\frac{x - X_i}{h_n}\right).$$

We estimate $m(x)$ by maximizing $L_{\text{loc}}(g|x)$ with respect to functions of the form

$$g(x^{(1)}, \ldots, x^{(d)}) = \exp\left(\sum_{j_1,\ldots,j_d=0,\ldots,M} a_{j_1,\ldots,j_d} \cdot (x^{(1)})^{j_1} \cdot \cdots \cdot (x^{(d)})^{j_d}\right). \tag{2}$$

More precisely, let $M \in \mathbb{N}_0$, $\beta_n > 1$ and set

$$\mathcal{F}_{M,\beta_n} = \Bigg\{ f : \mathbb{R}^d \to \mathbb{R}: f(x^{(1)}, \ldots, x^{(d)}) = \sum_{j_1,\ldots,j_d=0,\ldots,M} a_{j_1,\ldots,j_d} (x^{(1)})^{j_1} \cdot \ldots \cdot (x^{(d)})^{j_d}$$
$$(x^{(1)}, \ldots, x^{(d)} \in \mathbb{R}) \text{ for some } a_{j_1,\ldots,j_d} \in \mathbb{R} \text{ with } |a_{j_1,\ldots,j_d}| \leqslant \frac{\log(\beta_n)}{(M+1)^d} \Bigg\}$$

and

$$\mathcal{G}_{M,\beta_n} = \left\{ g : \mathbb{R}^d \to \mathbb{R}_+ : g(x) = \exp(f(x)) \ (x \in \mathbb{R}^d) \text{ for some } f \in \mathcal{F}_{M,\beta_n} \right\}.$$

Note that M controls the degrees of polynomials in (2) and β_n controls the size of polynomial coefficients. The bound on the coefficients in the definition of $\mathcal{F}_{M,\beta_n}$ implies

$$\frac{1}{\beta_n} \leqslant g(x) \leqslant \beta_n \quad \text{for all } x \in [0,1]^d$$

for all $g \in \mathcal{G}_{M,\beta_n}$. Later we will choose β_n such that

$$\beta_n \to \infty \quad (n \to \infty).$$

We note that β_n is needed in the proof of consistency of the estimate. For an example of a nonconsistent regression estimate without β_n in context of L_2 regression estimation using local polynomial kernel estimates with degree greater zero refer to Problem 10.3 in [12]. With this notation we define our estimate by

$$m_n(x) = \hat{g}_x(x), \tag{3}$$

where $\hat{g}_x \in \mathcal{G}_{M,\beta_n}$ satisfies

$$\hat{g}_x = \arg \max_{g \in \mathcal{G}_{M,\beta_n}} \sum_{i=1}^n \log\left(\frac{g(X_i)^{Y_i}}{Y_i!} \cdot e^{-g(X_i)}\right) \cdot K\left(\frac{x - X_i}{h_n}\right).$$

(Here $z_0 = \arg \max_{z \in D} f(z)$ is the value at which the function $f : D \to \mathbb{R}$ takes on its maximum, i.e., $z_0 \in D$ satisfies $f(z_0) = \max_{z \in D} f(z)$.) For notational simplicity we assume here and in the sequel that the maximum above does indeed exist. In case that it does not exist, it is easy to see that the results below do also hold if we define the value of the estimate at point x as the value of a function $\hat{g} \in \mathcal{G}_{M,\beta_n}$ which satisfies

$$\sum_{i=1}^n \log\left(\frac{\hat{g}(X_i)^{Y_i}}{Y_i!} \cdot e^{-\hat{g}(X_i)}\right) \cdot K\left(\frac{x - X_i}{h_n}\right)$$
$$\geqslant \sup_{g \in \mathcal{G}_{M,\beta_n}} \sum_{i=1}^n \log\left(\frac{g(X_i)^{Y_i}}{Y_i!} \cdot e^{-g(X_i)}\right) \cdot K\left(\frac{x - X_i}{h_n}\right) - \varepsilon_n,$$

provided $\varepsilon_n > 0$ is chosen such that

$$\varepsilon_n \to 0 \quad (n \to \infty).$$

Results concerning practical implementation of the estimate with degree zero are described in Remark 5 in the next section.

3. Main results

In the next theorem, we formulate our main result which concerns convergence to zero of the integrated L_1 distance between the conditional Poisson distribution and its estimate.

Theorem 1. *Let* $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots$ *be independent and identically distributed* $\mathbb{R}^d \times \mathbb{N}_0$*-valued random vectors which satisfy*

$$\mathbf{P}\{Y = y | X = x\} = \frac{m(x)^y}{y!} \cdot e^{-m(x)} \quad (y \in \mathbb{N}_0, x \in \mathbb{R}^d)$$

for some function $m : \mathbb{R}^d \to (0, \infty)$. *Assume*

$$X \in [0, 1]^d \quad a.s. \tag{4}$$

and

$$|m(x) - m(z)| \leqslant C_{\text{lip}}(m) \cdot \|x - z\| \quad (x, z \in \mathbb{R}^d) \tag{5}$$

for some constant $C_{\text{lip}}(m) \in \mathbb{R}$, *i.e., assume that* $\|X\|$ *is bounded a.s. and m is Lipschitz continuous with Lipschitz constant* $C_{\text{lip}}(m)$.

Consider a kernel $K : \mathbb{R}^d \to \mathbb{R}_+$ *satisfying*

$$K(u) = \tilde{K}(\|u\|) \quad (u \in \mathbb{R}^d)$$

for some $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ *which has total variation* $V < \infty$ *and satisfies for some* $r, R, b, B > 0$

$$b \cdot 1_{[0,r]}(v) \leqslant \tilde{K}(v) \leqslant B \cdot 1_{[0,R]}(v) \quad (v \in \mathbb{R}_+). \tag{6}$$

Choose $\beta_n, h_n > 0$ *such that*

$$\beta_n \to \infty \quad (n \to \infty), \tag{7}$$

$$h_n \beta_n^5 \exp(c \cdot \beta_n) \to 0 \quad (n \to \infty) \tag{8}$$

for any constant $c > 0$, *and*

$$\frac{n \cdot h_n^{2d}}{\log(n)^6} \to \infty \quad (n \to \infty). \tag{9}$$

Define the estimate $\widehat{\mathbf{P}}_n\{Y = y|X = x\}$ *as in* (1) *and* (3). *Then*

$$\int \sum_{y=0}^{\infty} \left|\widehat{\mathbf{P}}_n\{Y = y|X = x\} - \mathbf{P}\{Y = y|X = x\}\right| \mathbf{P}_X(dx) \to 0 \quad a.s.$$

The finite total variation condition and (6) is satisfied by many standard kernels with compact support, e.g., by naive, triangular and parabolic kernel.

By a discrete version of Scheffe's theorem (which follows, e.g., from the proof of Theorem 1.1 in [6]) we have for $x \in \mathbb{R}^d$,

$$\sum_{y=0}^{\infty} \left|\widehat{\mathbf{P}}_n\{Y = y|X = x\} - \mathbf{P}\{Y = y|X = x\}\right|$$
$$= 2 \sup_{A \subseteq \mathbb{N}_0} \left| \sum_{y \in A} \widehat{\mathbf{P}}_n\{Y = y|X = x\} - \sum_{y \in A} \mathbf{P}\{Y = y|X = x\} \right|, \tag{10}$$

therefore, under the assumptions of Theorem 1 the integrated total variation distance between $\mathbf{P}\{Y = \cdot|X = x\}$ and $\widehat{\mathbf{P}}_n\{Y = \cdot|X = x\}$ converges to zero almost surely. This can be used to construct asymptotic confidence intervals for Y given $X = x$. Let $\alpha \in (0, 1)$. Assume that given X we want to find an interval $I(X)$ of the form $I(X) = [0, u(X)]$, which is as small as possible and satisfies

$$\mathbf{P}\{Y \in I(X)\} \approx 1 - \alpha.$$

To construct such a confidence interval we choose the smallest value $u_n(x) \in \mathbb{R}$ such that

$$\sum_{y \in \mathbb{N}_0, y \leqslant u_n(x)} \widehat{\mathbf{P}}_n\{Y = y | X = x\} \geqslant 1 - \alpha, \tag{11}$$

and set $I_n(x) = [0, u_n(x)]$. From Theorem 1, we have the following result

Corollary 1. *Under the assumptions of Theorem* 1 *we have*

$$\liminf_{n \to \infty} \mathbf{P}\{Y \in I_n(X) | \mathcal{D}_n\} \geqslant 1 - \alpha \quad a.s.$$

Proof. By (11) we have

$$\begin{aligned}
&\mathbf{P}\{Y \in I_n(X) | \mathcal{D}_n\} \\
&\quad = \int \sum_{y \in I_n(x) \cap \mathbb{N}_0} \mathbf{P}\{Y = y | X = x\} \, \mathbf{P}_X(dx) \\
&\quad \geqslant 1 - \alpha - \left| \int \sum_{y \in I_n(x) \cap \mathbb{N}_0} \widehat{\mathbf{P}}_n\{Y = y | X = x\} \, \mathbf{P}_X(dx) \right. \\
&\qquad \left. - \int \sum_{y \in I_n(x) \cap \mathbb{N}_0} \mathbf{P}\{Y = y | X = x\} \, \mathbf{P}_X(dx) \right|.
\end{aligned}$$

Because of

$$\begin{aligned}
&\left| \int \sum_{y \in I_n(x) \cap \mathbb{N}_0} \widehat{\mathbf{P}}_n\{Y = y | X = x\} \, \mathbf{P}_X(dx) - \int \sum_{y \in I_n(x) \cap \mathbb{N}_0} \mathbf{P}\{Y = y | X = x\} \, \mathbf{P}_X(dx) \right| \\
&\quad \leqslant \int \sup_{A \subseteq \mathbb{N}_0} \left| \sum_{y \in A} \widehat{\mathbf{P}}_n\{Y = y | X = x\} - \sum_{y \in A} \mathbf{P}\{Y = y | X = x\} \right| \mathbf{P}_X(dx),
\end{aligned}$$

(10) and Theorem 1 yield the assertion. □

Next we investigate whether the length $u_n(X)$ of the confidence interval $I_n(X)$ converges to the optimal length $u(X)$, where for $x \in \mathbb{R}^d$ we define $u(x)$ as the smallest natural number which satisfies

$$\sum_{y \in \mathbb{N}_0, y \leqslant u(x)} \mathbf{P}\{Y = y | X = x\} \geqslant 1 - \alpha.$$

If the case

$$\sum_{y \in \mathbb{N}_0, y \leqslant u(x)} \mathbf{P}\{Y = y | X = x\} = 1 - \alpha$$

occurs, a very small error in the estimate of $m(x)$ may result in $|u_n(x) - u(x)| \geqslant 1$. Therefore, in general we cannot expect that $u_n(X)$ converges to $u(X)$. Instead we show below, that the probability that $u_n(X)$ deviates from $u(X)$ by more than one converges to zero.

Corollary 2. *Under the assumptions of Theorem* 1 *we have*

$$\mathbf{P}\{|u_n(X)-u(X)|>1\}\to 0 \quad (n\to\infty).$$

Proof. Set

$$\widehat{\mathbf{P}}_n\{Y=y|X\}=\frac{m_n(X)^y}{y!}\cdot e^{-m_n(X)} \quad \text{and} \quad \mathbf{P}\{Y=y|X\}=\frac{m(X)^y}{y!}\cdot e^{-m(X)}.$$

Since m is bounded away from zero and infinity on $[0,1]^d$ we can conclude that $u(x)$ is bounded and that

$$\mathbf{P}\{Y=y|X=x\}>c_1 \quad \text{for } y\leqslant u(x)+1$$

for some constant $c_1>0$. Assume that $|u_n(x)-u(x)|>1$. In case $u_n(x)>u(x)+1$ we have

$$\begin{aligned}
&\sum_{y\in\mathbb{N}_0,\, y\leqslant u(x)+1}\widehat{\mathbf{P}}_n\{Y=y|X=x\}-\sum_{y\in\mathbb{N}_0,\, y\leqslant u(x)+1}\mathbf{P}\{Y=y|X=x\}\\
&\quad\leqslant(1-\alpha)-\sum_{y\in\mathbb{N}_0,\, y\leqslant u(x)}\mathbf{P}\{Y=y|X=x\}-\mathbf{P}\{Y=u(x)+1|X=x\}\\
&\quad\leqslant(1-\alpha)-(1-\alpha)-c_1=-c_1.
\end{aligned}$$

In case $u(x)>u_n(x)+1$ we have $u(x)-2\geqslant u_n(x)$ which implies

$$\begin{aligned}
&\sum_{y\in\mathbb{N}_0,\, y\leqslant u(x)-2}\widehat{\mathbf{P}}_n\{Y=y|X=x\}-\sum_{y\in\mathbb{N}_0,\, y\leqslant u(x)-2}\mathbf{P}\{Y=y|X=x\}\\
&\quad\geqslant(1-\alpha)-\sum_{y\in\mathbb{N}_0,\, y\leqslant u(x)-1}\mathbf{P}\{Y=y|X=x\}+\mathbf{P}\{Y=u(x)-1|X=x\}\\
&\quad\geqslant(1-\alpha)-(1-\alpha)+c_1=c_1.
\end{aligned}$$

From this we conclude that

$$|u_n(X)-u(X)|>1$$

implies

$$\max_{k\in\{u(X)-2,u(X)+1\}}\left|\sum_{y\in\mathbb{N}_0,\, y\leqslant k}\widehat{\mathbf{P}}_n\{Y=y|X\}-\sum_{y\in\mathbb{N}_0,\, y\leqslant k}\mathbf{P}\{Y=y|X\}\right|>c_1.$$

From this we get by Markov inequality

$$\begin{aligned}
\mathbf{P}\{|u_n(X)-u(X)|>1\}&\leqslant\mathbf{P}\left\{\sup_{A\subseteq\mathbb{N}_0}\left|\sum_{y\in A}\widehat{\mathbf{P}}_n\{Y=y|X\}-\sum_{y\in A}\mathbf{P}\{Y=y|X\}\right|>c_1\right\}\\
&\leqslant 2\cdot\mathbf{E}\sup_{A\subseteq\mathbb{N}_0}\left|\sum_{y\in A}\widehat{\mathbf{P}}_n\{Y=y|X\}-\sum_{y\in A}\mathbf{P}\{Y=y|X\}\right|\Big/ 2c_1.
\end{aligned}$$

By (10), dominated convergence theorem and Theorem 1 we have

$$2\cdot\mathbf{E}\sup_{A\subseteq\mathbb{N}_0}\left|\sum_{y\in A}\widehat{\mathbf{P}}_n\{Y=y|X\}-\sum_{y\in A}\mathbf{P}\{Y=y|X\}\right|$$

$$
\begin{aligned}
&= \mathbf{E} \sum_{y=0}^{\infty} \left| \widehat{\mathbf{P}}_n\{Y = y|X\} - \mathbf{P}\{Y = y|X\} \right| \\
&= \mathbf{E} \sum_{y=0}^{\infty} \int \left| \widehat{\mathbf{P}}_n\{Y = y|X = x\} - \mathbf{P}\{Y = y|X = x\} \right| \mathbf{P}_X(dx) \\
&\to 0 \quad (n \to \infty),
\end{aligned}
$$

which implies the assertion. □

Remark 1. We would like to stress that in the above results there is no assumption on the distribution of X besides $X \in [0, 1]^d$ a.s. In particular X is not required to have a density with respect to the Lebesgue–Borel measure. We can handle the case of unbounded X using truncation.

Remark 2. It was suggested by the referee that is possible to weaken with considerable effort condition (9) to $nh_n^d/(\log n)^M \to \infty$ for some $M > 0$ by making use of localization property of the function $K(\frac{\cdot - x}{h_n})$, $x \in [0, 1]^d$, by a properly chosen discretization of $[0, 1]^d$. Similar approach has been applied by Einmahl and Mason [9] to study uniform consistency of kernel estimators.

Remark 3. If we assume that the regression function is bounded by some constant L and that we know this bound (this assumption is not required in the results above), we can construct a strong pointwise consistent estimate $m_n(x)$ of m, i.e., an estimate which satisfies for $\mathbf{P}_X$-almost all x,

$$
m_n(x) \to m(x) \quad \text{a.s.,}
$$

which is bounded by L, too (the last property can be ensured by truncation of the estimate). Since the function $f(z) = z^y \cdot e^{-z}$ is Lipschitz continuous on $[0, L]$ with Lipschitz constant $(y+1) \cdot L^y$, this pointwise consistency implies

$$
\int \sum_{y=0}^{\infty} \left| \frac{m_n(x)^y}{y!} \cdot e^{-m_n(x)} - \frac{m(x)^y}{y!} \cdot e^{-m(x)} \right| \mathbf{P}_X(dx) \to 0 \quad \text{a.s.}
$$

Therefore, for truncated versions of estimates which are strong universal pointwise consistent, the result of Theorem 1 does hold, too, provided a bound on the supremum norm of the regression function is known a priori. Various strong universal pointwise consistent estimates have been constructed in [1,2,21,35]. For related universal consistency result see, e.g., [32,31,8,13,14,25,19].

In view of this, the main new results in Theorem 1 are, that firstly the bound on m does not have to be known in advance, and secondly the consistency result in Theorem 1 holds also for the localized maximum likelihood estimate which has not been considered in the papers above, but which seems to be especially suitable in the context of this paper where the main aim is not the estimation of the regression function but estimation of $\mathbf{P}\{Y = y|X = x\}$.

Remark 4. In the paper we only consider the prediction intervals of the shape $I(X) = [0, u(X)]$. It is possible to extend the paper to intervals $I(X) = [u_1(X), u_2(X)]$ of the smallest length.

Remark 5. In order to investigate the behavior of the estimate in practice we provide the results of Monte-Carlo experiments. Take degree $M = 0$. This implies that our estimate is the Nadaraya–Watson estimate with the naive kernel. Let $m(x) = \sin(2\pi x) + 2$ and assume that X is uniform on $[0, 1]$. In the simulations we used sample size of 100, 200, 400 and the bandwidth $h = 0.1, 0.07,$

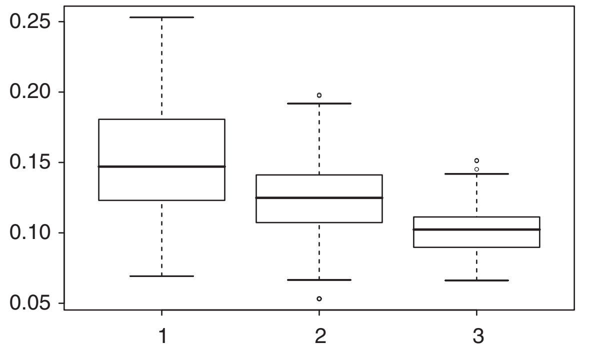

Fig. 1. Boxplots of the integrated L_1 errors of the estimates of m for sample sizes 100, 200, 400.

0.05 (approximately const $* \, 1/\sqrt(n)$). Each simulation was repeated 100 times, we calculated for these 100 times the corresponding integrated L_1 distance, and we produced three box plots of the estimate corresponding to the three sample sizes. The results are shown in Fig. 1.

4. Outline of the proof of Theorem 1

In the proof of Theorem 1 we approximate the integrated distance between the estimated and the true conditional Poisson distribution by the integrated Hellinger distance

$$\int \sum_{y=0}^{\infty} \left(\sqrt{\widehat{\mathbf{P}}_n\{Y = y|X = x\}} - \sqrt{\mathbf{P}\{Y = y|X = x\}} \right)^2 \mathbf{P}_X(dx) \tag{12}$$

between the two conditional distributions and we show that it converges to zero almost surely. Then we bound this integrated Hellinger distance from above by some constant times

$$-\mathbf{E}\left\{ \log \frac{\widehat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X\}}{2\mathbf{P}\{Y|X\}} \middle| \mathcal{D}_n \right\},$$

where

$$\widehat{\mathbf{P}}_n\{Y|X\} = \frac{m_n(X)^Y}{Y!} \cdot e^{-m_n(X)} \quad \text{and} \quad \mathbf{P}\{Y|X\} = \frac{m(X)^Y}{Y!} \cdot e^{-m(X)}.$$

Next, we take advantage of Lipschitz continuity of m to approximate (12) by an expected value of the smoothed conditional expectation

$$-\int \frac{\mathbf{E}\left\{ \log \frac{\widehat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X = x\}}{2\mathbf{P}\{Y|X = x\}} \cdot K\left(\frac{x - X}{h_n}\right) \middle| \mathcal{D}_n \right\}}{\mathbf{E}K\left(\frac{x - X}{h_n}\right)} \mathbf{P}_X(dx),$$

where

$$\widehat{\mathbf{P}}_n\{Y|X\} = \frac{\hat{g_x}(X)^Y}{Y!} \cdot e^{-\hat{g_x}(X)} \quad \text{and} \quad \mathbf{P}\{Y|X = x\} = \frac{m(x)^Y}{Y!} \cdot e^{-m(x)}.$$

We then approximate the nominator in the integral above by its empirical version and relate it to the estimate. We note that by definition of the estimate and concavity of the log-function, the empirical version of the nominator above

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\frac{\hat{g_x}(X_i)^{Y_i}}{Y_i!}\cdot e^{-\hat{g_x}(X_i)}+\frac{m(x)^{Y_i}}{Y_i!}\cdot e^{-m(x)}}{2\frac{m(x)^{Y_i}}{Y_i!}\cdot e^{-m(x)}}\right)\cdot K\left(\frac{x-X_i}{h_n}\right)$$

is always nonnegative. Therefore, it suffices to show that the difference between the nominator above and its empirical version is asymptotically small, which we prove by using results of empirical process theory.

5. Proofs

Proof of Theorem 1. In the *first step of the proof* we observe that

$$\int\sum_{y=0}^{\infty}\left|\widehat{\mathbf{P}}_n\{Y=y|X=x\}-\mathbf{P}\{Y=y|X=x\}\right|\mathbf{P}_X(dx)\to 0 \quad \text{a.s.} \tag{13}$$

follows from

$$\int\sum_{y=0}^{\infty}\left(\sqrt{\widehat{\mathbf{P}}_n\{Y=y|X=x\}}-\sqrt{\mathbf{P}\{Y=y|X=x\}}\right)^2\mathbf{P}_X(dx)\to 0 \quad \text{a.s.} \tag{14}$$

For the sake of completeness we repeat the proof of this well-known fact (cf., e.g., [7]). Observe that for $a, b > 0$

$$|a-b|=\left|\sqrt{a}-\sqrt{b}\right|\cdot\left|\sqrt{a}+\sqrt{b}\right|\leqslant\left(\sqrt{a}-\sqrt{b}\right)^2+2\sqrt{b}\cdot\left|\sqrt{a}-\sqrt{b}\right|$$

and conclude from this and the Cauchy–Schwarz inequality

$$\begin{aligned}
&\int\sum_{y=0}^{\infty}\left|\widehat{\mathbf{P}}_n\{Y=y|X=x\}-\mathbf{P}\{Y=y|X=x\}\right|\mathbf{P}_X(dx)\\
&\quad\leqslant\int\sum_{y=0}^{\infty}\left(\sqrt{\widehat{\mathbf{P}}_n\{Y=y|X=x\}}-\sqrt{\mathbf{P}\{Y=y|X=x\}}\right)^2\mathbf{P}_X(dx)\\
&\qquad+2\cdot\int\sum_{y=0}^{\infty}\sqrt{\mathbf{P}\{Y=y|X=x\}}\cdot\left|\sqrt{\widehat{\mathbf{P}}_n\{Y=y|X=x\}}-\sqrt{\mathbf{P}\{Y=y|X=x\}}\right|\mathbf{P}_X(dx)\\
&\quad\leqslant\int\sum_{y=0}^{\infty}\left(\sqrt{\widehat{\mathbf{P}}_n\{Y=y|X=x\}}-\sqrt{\mathbf{P}\{Y=y|X=x\}}\right)^2\mathbf{P}_X(dx)\\
&\qquad+2\cdot\int\sqrt{\sum_{y=0}^{\infty}\mathbf{P}\{Y=y|X=x\}}\\
&\qquad\times\sqrt{\sum_{y=0}^{\infty}\left(\sqrt{\widehat{\mathbf{P}}_n\{Y=y|X=x\}}-\sqrt{\mathbf{P}\{Y=y|X=x\}}\right)^2}\mathbf{P}_X(dx).
\end{aligned}$$

With

$$\sqrt{\sum_{y=0}^{\infty} \mathbf{P}\{Y = y|X = x\}} = \sqrt{1} = 1$$

and

$$\int \sqrt{\sum_{y=0}^{\infty} \left(\sqrt{\widehat{\mathbf{P}}_n\{Y = y|X = x\}} - \sqrt{\mathbf{P}\{Y = y|X = x\}} \right)^2} \, \mathbf{P}_X(dx)$$
$$\leqslant 1 \cdot \sqrt{\int \sum_{y=0}^{\infty} \left(\sqrt{\widehat{\mathbf{P}}_n\{Y = y|X = x\}} - \sqrt{\mathbf{P}\{Y = y|X = x\}} \right)^2 \mathbf{P}_X(dx)}$$

(which follows from another application of the Cauchy–Schwarz inequality) the assertion of the first step follows.

In the *second step of the proof* we show

$$\int \sum_{y=0}^{\infty} \left(\sqrt{\widehat{\mathbf{P}}_n\{Y = y|X = x\}} - \sqrt{\mathbf{P}\{Y = y|X = x\}} \right)^2 \mathbf{P}_X(dx)$$
$$\leqslant -16 \cdot \mathbf{E}\left\{ \log \left(\frac{\widehat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X\}}{2\mathbf{P}\{Y|X\}} \right) \middle| \mathcal{D}_n \right\}, \tag{15}$$

where

$$\widehat{\mathbf{P}}_n\{Y|X\} = \frac{m_n(X)^Y}{Y!} \cdot e^{-m_n(X)} \quad \text{and} \quad \mathbf{P}\{Y|X\} = \frac{m(X)^Y}{Y!} \cdot e^{-m(X)}.$$

Lemma 4.2 in [33] yields

$$\sum_{y=0}^{\infty} \left(\sqrt{\widehat{\mathbf{P}}_n\{Y = y|X = x\}} - \sqrt{\mathbf{P}\{Y = y|X = x\}} \right)^2$$
$$\leqslant 16 \cdot \sum_{y=0}^{\infty} \left(\sqrt{\frac{\widehat{\mathbf{P}}_n\{Y = y|X = x\} + \mathbf{P}\{Y = y|X = x\}}{2}} - \sqrt{\mathbf{P}\{Y = y|X = x\}} \right)^2. \tag{16}$$

The rest of the proof mimics the proof of Lemma 1.3 in [33]. Define

$$a(x, y) = \frac{\widehat{\mathbf{P}}_n\{Y = y|X = x\} + \mathbf{P}\{Y = y|X = x\}}{2}$$

and

$$b(x, y) = \mathbf{P}\{Y = y|X = x\}.$$

Inequality $\frac{1}{2}\log u \leqslant \sqrt{u} - 1$ implies

$$\frac{1}{2} \log \frac{a(x, y)}{b(x, y)} \leqslant \sqrt{\frac{a(x, y)}{b(x, y)}} - 1$$

or equivalently

$$\frac{1}{2}\log\frac{b(x,y)}{a(x,y)} \geqslant 1-\sqrt{\frac{a(x,y)}{b(x,y)}}.$$

Hence

$$\begin{aligned}\frac{1}{2}\sum_{y=0}^{\infty} b(x,y)\log\frac{b(x,y)}{a(x,y)} &\geqslant 1-\sum_{y=0}^{\infty}\sqrt{\frac{a(x,y)}{b(x,y)}}b(x,y)\\ &= 1-\sum_{y=0}^{\infty}\sqrt{a(x,y)b(x,y)}\\ &= \frac{1}{2}\sum_{y=0}^{\infty}a(x,y)+\frac{1}{2}\sum_{y=0}^{\infty}b(x,y)-\sum_{y=0}^{\infty}\sqrt{a(x,y)b(x,y)}\\ &= \frac{1}{2}\sum_{y=0}^{\infty}\left(\sqrt{a(x,y)}-\sqrt{b(x,y)}\right)^2.\end{aligned}$$

Thus we get

$$\sum_{y=0}^{\infty}\left(\sqrt{a(x,y)}-\sqrt{a(x,y)}\right)^2 \leqslant -\sum_{y=0}^{\infty} b(x,y)\,\log\frac{a(x,y)}{b(x,y)}.$$

Using the definitions of $a(x,y)$ and $b(x,y)$ and inequality (16) we get

$$\begin{aligned}&\sum_{y=0}^{\infty}\left(\sqrt{\widehat{\mathbf{P}}_n\{Y=y|X=x\}}-\sqrt{\mathbf{P}\{Y=y|X=x\}}\right)^2\\ &\quad\leqslant -16\cdot\mathbf{E}_{\mathcal{D}_n}\left\{\log\left(\frac{\widehat{\mathbf{P}}_n\{Y|X\}+\mathbf{P}\{Y|X\}}{2\cdot\mathbf{P}\{Y|X\}}\right)\middle| X=x\right\},\end{aligned}$$

where in $\mathbf{E}_{\mathcal{D}_n}\{\cdot|X=x\}$ we take the expectation only with respect to Y for fixed $X=x$ and fixed $\mathcal{D}_n$. By integrating this inequality with respect to $\mathbf{P}_X$ we get (15).

In the *third step of the proof* we show

$$\begin{aligned}&\mathbf{E}\left\{\log\left(\frac{\widehat{\mathbf{P}}_n\{Y|X\}+\mathbf{P}\{Y|X\}}{2\cdot\mathbf{P}\{Y|X\}}\right)\middle|\mathcal{D}_n\right\}\\ &\quad-\int\frac{\mathbf{E}\left\{\log\left(\frac{\widehat{\mathbf{P}}_n\{Y|X\}+\mathbf{P}\{Y|X=x\}}{2\cdot\mathbf{P}\{Y|X=x\}}\right)\cdot K\left(\frac{x-X}{h_n}\right)\middle|\mathcal{D}_n\right\}}{\mathbf{E}K\left(\frac{x-X}{h_n}\right)}\mathbf{P}_X(dx)\\ &\quad\to 0\quad\text{a.s.,}\end{aligned}\tag{17}$$

where

$$\widehat{\mathbf{P}}_n\{Y|X\} = \frac{\hat{g}(X)^Y}{Y!} \cdot e^{-\hat{g}(X)} \quad \text{and} \quad \mathbf{P}\{Y|X=x\} = \frac{m(x)^Y}{Y!} \cdot e^{-m(x)}.$$

The first expectation on the left-hand side of (17) can be written as

$$\begin{aligned}
&\int \mathbf{E}_{\mathcal{D}_n} \left\{ \log \left(\frac{\widehat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X\}}{2 \cdot \mathbf{P}\{Y|X\}} \right) \Bigg| X = x \right\} \mathbf{P}_X(dx) \\
&= \int \sum_{y=0}^{\infty} \log \left(\frac{\widehat{\mathbf{P}}_n\{Y=y|X=x\} + \mathbf{P}\{Y=y|X=x\}}{2\mathbf{P}\{Y=y|X=x\}} \right) \mathbf{P}\{Y=y|X=x\} \mathbf{P}_X(dx) \\
&= \int \phi_n(x) \mathbf{P}_X(dx).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\frac{\mathbf{E} \left\{ \log \left(\frac{\widehat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X=x\}}{2 \cdot \mathbf{P}\{Y|X=x\}} \right) \cdot K\left(\frac{x-X}{h_n} \right) \Bigg| \mathcal{D}_n \right\}}{\mathbf{E} K\left(\frac{x-X}{h_n} \right)} \\
&= \frac{\int \phi_{n,x}(u) \cdot K\left(\frac{x-u}{h_n} \right) \mathbf{P}_X(du)}{\int K\left(\frac{x-u}{h_n} \right) \mathbf{P}_X(du)},
\end{aligned}$$

where

$$\begin{aligned}
\phi_{n,x}(u) &= \mathbf{E}_{\mathcal{D}_n} \left\{ \log \left(\frac{\widehat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X=x\}}{2 \cdot \mathbf{P}\{Y|X=x\}} \right) \Bigg| X = u \right\} \\
&= \sum_{y=0}^{\infty} \log \left(\frac{\frac{\hat{g}(u)^y}{y!} \cdot e^{-\hat{g}(u)} + \frac{m(x)^y}{y!} \cdot e^{-m(x)}}{2 \frac{m(x)^y}{y!} \cdot e^{-m(x)}} \right) \cdot \frac{m(u)^y}{y!} \cdot e^{-m(u)}.
\end{aligned}$$

Because of $m_n(x) = \hat{g}(x)$ we have

$$\phi_{n,x}(x) = \phi_n(x).$$

We will show in Lemma 1 below that there exists $c_n > 0$ with

$$c_n h_n \to 0 \quad (n \to \infty)$$

such that for all $x, u, v \in [0,1]^d$

$$|\phi_{n,x}(u) - \phi_{n,x}(v)| \leqslant c_n \cdot \|u - v\|,$$

(i.e., such that $\phi_{n,x}$ is Lipschitz continuous with Lipschitz constant c_n independent of x). We use Lipschitz continuity here to avoid making assumptions on the existence of a density of X. Using this, we can bound the absolute value of the left-hand side of (17) by

$$
\begin{aligned}
&\left| \int \phi_{n,x}(x)\mathbf{P}_X(dx) - \int \frac{\int \phi_{n,x}(u)\cdot K\left(\frac{x-u}{h_n}\right)\mathbf{P}_X(du)}{\int K\left(\frac{x-u}{h_n}\right)\mathbf{P}_X(du)}\mathbf{P}_X(dx)\right| \\
&\leqslant \int \frac{\int |\phi_{n,x}(x)-\phi_{n,x}(u)|\cdot K\left(\frac{x-u}{h_n}\right)\mathbf{P}_X(du)}{\int K\left(\frac{x-u}{h_n}\right)\mathbf{P}_X(du)}\mathbf{P}_X(dx) \\
&\leqslant c_n \cdot R \cdot h_n \to 0 \quad (n\to\infty),
\end{aligned}
$$

where we have used in the first inequality that the set of all x with

$$
\int K\left(\frac{x-u}{h_n}\right)\mathbf{P}_X(du) = 0
$$

has $\mathbf{P}_X$-measure zero (for a related argument see, e.g., the last step in the proof of Lemma 24.5 in [12]), and where the second inequality follows from $K((x-u)/h_n) = 0$ for $\|x-u\| > R\cdot h_n$.

In the *fourth step of the proof* we show

$$
\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\frac{\hat{g}(X_i)^{Y_i}}{Y_i!}\cdot e^{-\hat{g}(X_i)} + \frac{m(x)^{Y_i}}{Y_i!}\cdot e^{-m(x)}}{2\frac{m(x)^{Y_i}}{Y_i!}\cdot e^{-m(x)}}\right)\cdot K\left(\frac{x-X_i}{h_n}\right) \geqslant 0 \tag{18}
$$

for n sufficiently large (i.e., whenever $\log(\beta_n)/(M+1)^d \geqslant \log(\|m\|_\infty)$, where $\|m\|_\infty$ is the supremum norm of m) and all $x \in [0,1]^d$.

Let n be such that $\log(\beta_n)/(M+1)^d \geqslant \log(\|m\|_\infty)$. By concavity of the log-function we have

$$
\log\frac{a+b}{2b} = \log\left(\frac{1}{2}\cdot\frac{a}{b}+\frac{1}{2}\cdot 1\right) \geqslant \frac{1}{2}\cdot\log\frac{a}{b} + \frac{1}{2}\cdot\log 1 = \frac{1}{2}\cdot\log\frac{a}{b}
$$

for all $a, b > 0$, which implies

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\frac{\hat{g}(X_i)^{Y_i}}{Y_i!}\cdot e^{-\hat{g}(X_i)} + \frac{m(x)^{Y_i}}{Y_i!}\cdot e^{-m(x)}}{2\frac{m(x)^{Y_i}}{Y_i!}\cdot e^{-m(x)}}\right)\cdot K\left(\frac{x-X_i}{h_n}\right) \\
&\geqslant \frac{1}{2}\cdot\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\frac{\hat{g}(X_i)^{Y_i}}{Y_i!}\cdot e^{-\hat{g}(X_i)}}{\frac{m(x)^{Y_i}}{Y_i!}\cdot e^{-m(x)}}\right)\cdot K\left(\frac{x-X_i}{h_n}\right) \\
&= \frac{1}{2}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\hat{g}(X_i)^{Y_i}}{Y_i!}\cdot e^{-\hat{g}(X_i)}\right)\cdot K\left(\frac{x-X_i}{h_n}\right)\right. \\
&\qquad \left. -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{m(x)^{Y_i}}{Y_i!}\cdot e^{-m(x)}\right)\cdot K\left(\frac{x-X_i}{h_n}\right)\right) \\
&\geqslant 0
\end{aligned}
$$

by definition of $\hat{g}$. This proves (18).

In the *fifth step of the proof* we set

$$\widehat{\mathbf{P}}_n\{Y_i|X_i\} = \frac{\hat{g}(X_i)^{Y_i}}{Y_i!} \cdot e^{-\hat{g}(X_i)} \quad \text{and} \quad \mathbf{P}\{Y_i|X_i = x\} = \frac{m(x)^{Y_i}}{Y_i!} \cdot e^{-m(x)},$$

and show that

$$A_n := \frac{1}{h_n^d} \cdot \sup_{x\in[0,1]^d} \Bigg| \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{\widehat{\mathbf{P}}_n\{Y_i|X_i\} + \mathbf{P}\{Y_i|X_i = x\}}{2 \cdot \mathbf{P}\{Y_i|X_i = x\}}\right) \cdot K\left(\frac{x - X_i}{h_n}\right) - \mathbf{E}\left\{\log\left(\frac{\widehat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X = x\}}{2 \cdot \mathbf{P}\{Y|X = x\}}\right) \cdot K\left(\frac{x - X}{h_n}\right) \Bigg| \mathcal{D}_n\right\}\Bigg| \to 0 \quad \text{a.s.} \tag{19}$$

implies the assertion.

From step 2 we conclude

$$\begin{aligned} 0 &\leqslant \int \sum_{y=0}^{\infty} \left(\sqrt{\widehat{\mathbf{P}}_n\{Y = y|X = x\}} - \sqrt{\mathbf{P}\{Y = y|X = x\}}\right)^2 \mathbf{P}_X(dx) \\ &\leqslant -16 \cdot (B_n - C_n) - 16 \cdot C_n, \end{aligned}$$

where

$$B_n = \mathbf{E}\left\{\log\left(\frac{\widehat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X\}}{2\mathbf{P}\{Y|X\}}\right)\Bigg| \mathcal{D}_n\right\}$$

and

$$C_n = \int \frac{\mathbf{E}\left\{\log\left(\frac{\widehat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X = x\}}{2 \cdot \mathbf{P}\{Y|X = x\}}\right) \cdot K\left(\frac{x - X}{h_n}\right)\Bigg| \mathcal{D}_n\right\}}{\mathbf{E}K\left(\frac{x - X}{h_n}\right)} \mathbf{P}_X(dx).$$

By step 3 we have

$$B_n - C_n \to 0 \quad \text{a.s.}$$

so by step 1 the assertion of Theorem 1 follows from

$$\limsup_{n\to\infty}(-C_n) \leqslant 0 \quad \text{a.s.} \tag{20}$$

Set

$$D_n = \int \frac{\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\widehat{\mathbf{P}}_n\{Y_i|X_i\} + \mathbf{P}\{Y_i|X_i = x\}}{2 \cdot \mathbf{P}\{Y_i|X_i = x\}}\right) \cdot K\left(\frac{x - X_i}{h_n}\right)}{\mathbf{E}K\left(\frac{x - X}{h_n}\right)} \mathbf{P}_X(dx).$$

In step 4 we have shown

$$D_n \geqslant 0,$$

so

$$-C_n = (D_n - C_n) - D_n \leqslant (D_n - C_n)$$

and (20) follows from

$$D_n - C_n \to 0 \quad \text{a.s.}$$

But this in turn is implied by (19), since

$$|D_n - C_n| \leqslant A_n \cdot \int \frac{1}{\mathbf{E}\left\{\frac{1}{h_n^d} \cdot K\left(\frac{x-X}{h_n}\right)\right\}} \mathbf{P}_X(dx)$$

and

$$\int \frac{1}{\mathbf{E}\left\{\frac{1}{h_n^d} \cdot K\left(\frac{x-X}{h_n}\right)\right\}} \mathbf{P}_X(dx) < \infty$$

by Lemma 3.1(b) in [18].

In the *sixth* (*and final*) *step of the proof* we apply the results of empirical process theory to show (19). Let $\mathcal{H}_n$ be the set of all functions

$$h : \mathbb{R}^d \times \mathbb{N}_0 \to \mathbb{R}$$

which satisfy

$$h(x, y) = \log\left(\frac{\frac{g(x)^y}{y!} \cdot e^{-g(x)} + \frac{\alpha^y}{y!} \cdot e^{-\alpha}}{2 \cdot \frac{\alpha^y}{y!} \cdot e^{-\alpha}}\right) \cdot K\left(\frac{u-x}{h_n}\right)$$

for some $g \in \mathcal{G}_{M,\beta_n}$, $u \in \mathbb{R}^d$ and $\alpha \in [c_2, c_3]$, where $c_2 = \min_{x\in[0,1]^d} m(x) > 0$ and $c_3 = \max_{x\in[0,1]^d} m(x) < \infty$. Let $k_n = \lceil \log n \rceil$ be the smallest integer greater than or equal to $\log n$. Then

$$A_n \leqslant \frac{1}{h_n^d} \cdot \sup_{h\in\mathcal{H}_n} \left| \frac{1}{n}\sum_{i=1}^n h(X_i, Y_i) - \mathbf{E}h(X, Y)\right| \leqslant \sum_{i=1}^3 T_{i,n},$$

where

$$T_{1,n} = \frac{1}{h_n^d} \cdot \sup_{h\in\mathcal{H}_n} \left| \frac{1}{n}\sum_{i=1}^n h(X_i, Y_i) \cdot 1_{\{Y_i \leqslant k_n\}} - \mathbf{E}\left\{h(X, Y) 1_{\{Y \leqslant k_n\}}\right\}\right|,$$

$$T_{2,n} = \frac{1}{h_n^d} \cdot \frac{1}{n}\sum_{i=1}^n \sup_{h\in\mathcal{H}_n} |h(X_i, Y_i)| \cdot 1_{\{Y_i > k_n\}}$$

and

$$T_{3,n} = \frac{1}{h_n^d} \cdot \mathbf{E}\left\{\sup_{h\in\mathcal{H}_n} |h(X, Y)| 1_{\{Y > k_n\}}\right\}.$$

Notice that

$$
\begin{aligned}
|h(x,y)| &\leqslant B \cdot \log\left(2 \cdot \max\left\{\left(\frac{1}{2}\right) \cdot \left(\frac{g(x)}{\alpha}\right)^y e^{-g(x)+\alpha}, \frac{1}{2}\right\}\right) \\
&\leqslant B \cdot |y \cdot \log(g(x)/\alpha) - g(x) + \alpha| \\
&\leqslant B \cdot \left(y \cdot \log(\beta_n/c_2) + c_3 + \beta_n\right) \leqslant c_4 \cdot y \cdot \log n
\end{aligned} \tag{21}
$$

for $x \in [0,1]^d$, $y \in \mathbb{N}$ and $h \in \mathcal{H}_n$, cf. (7)–(9). By Markov's inequality we then get, for an arbitrary $\varepsilon > 0$, for all large n

$$
\begin{aligned}
&\mathbf{P}\left\{T_{2,n} > \varepsilon\right\} \\
&= \mathbf{P}\left\{\sum_{k=k_n+1}^{\infty} \sum_{i=1}^{n} \sup_{h \in \mathcal{H}_n} |h(X_i, Y_i)| \cdot 1_{\{Y_i = k\}} > n \cdot h_n^d \cdot \varepsilon\right\} \\
&\leqslant \frac{\mathbf{E}\left\{\sum_{k=k_n+1}^{\infty} \sum_{i=1}^{n} \sup_{h \in \mathcal{H}_n} |h(X_i, Y_i)| \cdot 1_{\{Y_i = k\}}\right\}}{n \cdot h_n^d \cdot \varepsilon} \\
&\leqslant \frac{n \cdot \sum_{k=k_n+1}^{\infty} c_4 \cdot k \cdot \log n \cdot \sup_{x \in [0,1]^d} \dfrac{m(x)^k}{k!} \cdot e^{-m(x)}}{n \cdot h_n^d \cdot \varepsilon} \\
&\leqslant \frac{c_4 \log n}{h_n^d \cdot \varepsilon} \cdot c_3 \cdot e^{-c_2} \cdot \sum_{k=k_n+1}^{\infty} \frac{c_3^{k_n}}{k_n!} \cdot \frac{c_3^{k-1-k_n}}{(k-1-k_n)!} \\
&= \frac{c_5 \log n}{h_n^d \cdot \varepsilon} \cdot \frac{c_3^{k_n}}{k_n!} \\
&\leqslant \frac{c_5 \log n}{h_n^d \cdot \varepsilon} \cdot c_3^{k_n} \cdot \left(\frac{k_n}{2}\right)^{-\frac{k_n}{2}} \\
&\leqslant \frac{c_5}{\varepsilon} \cdot \exp\left(\log \frac{\log n}{h_n^d} + k_n \cdot \log c_3 - \frac{k_n}{2} \cdot \log \frac{k_n}{2}\right).
\end{aligned}
$$

Since

$$
\frac{\log \dfrac{\log n}{h_n^d}}{\log(n) \cdot \log(\log n)} \to 0 \quad (n \to \infty),
$$

the last term is summable for each $\varepsilon > 0$. Application of the Borel–Cantelli lemma yields

$$
T_{2,n} \to 0 \quad \text{a.s.}
$$

Similarly we get

$$
\begin{aligned}
T_{3,n} &= \frac{1}{h_n^d} \sum_{k=k_n+1}^{\infty} \mathbf{E}\left\{\sup_{h \in \mathcal{H}_n} |h(X,Y)| \cdot 1_{\{Y=k\}}\right\} \\
&\leqslant \frac{c_6 \log n}{h_n^d} \cdot \sum_{k=k_n+1}^{\infty} k \cdot \sup_{x \in [0,1]^d} \frac{m(x)^k}{k!} \cdot e^{-m(x)} \\
&\leqslant c_7 \frac{\log n}{h_n^d} \cdot \frac{c_8^{k_n}}{k_n!} \to 0 \quad (n \to \infty).
\end{aligned}
$$

So it remains to show

$$T_{1,n} \to 0 \quad \text{a.s.} \tag{22}$$

To do this, we apply Theorem 9.1 in [12] and Lemma 2 below. From these we get for an arbitrary $\varepsilon > 0$

$$\mathbf{P}\left\{T_{1,n} > \varepsilon\right\} \leqslant 8 \cdot \left(c_9 \frac{\beta_n^{k_n} \cdot k_n}{h_n^d \cdot \varepsilon}\right)^{c_{10}} \cdot \exp\left(-\frac{n \cdot \varepsilon^2 \cdot h_n^{2d}}{c_{11} \cdot k_n^2 \cdot (\log\, n)^2}\right).$$

By the assumptions of Theorem 1 we have

$$n \cdot h_n^d \to \infty \quad (n \to \infty) \quad \text{and} \quad \frac{\beta_n}{n} \to 0 \quad (n \to \infty).$$

Indeed, (7) and (8) imply $nh_n \exp(c\beta_n)\frac{\beta_n^5}{n} \to 0$ and thus $\frac{\beta_n^5}{n} \to 0$ because $nh_n\, \exp(c\beta_n) \to 0$. This yields $\frac{\beta_n}{n} \to 0$. Using this we get

$$\mathbf{P}\left\{T_{1,n} > \varepsilon\right\} \leqslant c_{12} \cdot \exp\left(c_{13} \cdot k_n \cdot \log\, n - c_{14} \frac{n \cdot h_n^{2d} \cdot \varepsilon^2}{\log(n)^4}\right).$$

Because of

$$\frac{n \cdot h_n^{2d}}{\log(n)^6} \to \infty \quad (n \to \infty)$$

the right-hand side above is summable for each $\varepsilon > 0$. Application of the Borel–Cantelli lemma yields (22). The proof of Theorem 1 is complete. □

Lemma 1. *Let $\phi_{n,x}$ be defined as in the third step of the proof of Theorem 1 and assume that the assumptions of Theorem 1 are satisfied. Then there exists $c_n > 0$ with*

$$c_n h_n \to 0 \quad (n \to \infty)$$

such that for all $x, u, v \in [0, 1]^d$,

$$|\phi_{n,x}(u) - \phi_{n,x}(v)| \leqslant c_n \cdot \|u - v\|.$$

Proof. The functions in $\mathcal{G}_{M,\beta_n}$ are bounded in absolute value by β_n and are Lipschitz continuous on $[0, 1]^d$ with Lipschitz constant bounded by

$$c_{15} \cdot \beta_n\, \log\, \beta_n$$

for some constant c_{15} depending on M. In addition, the function $f(z) = z^k \cdot e^{-z}$ satisfies

$$|f'(z)| \leqslant (k+1) \cdot \beta_n^k \quad \text{for } z \in [0, \beta_n],$$

from which we can conclude that the function

$$u \mapsto \frac{\hat{g}(u)^k e^{-\hat{g}(u)} + m(x)^k e^{-m(x)}}{2m(x)^k e^{-m(x)}} = \frac{\hat{g}(u)^k e^{-\hat{g}(u)}}{2m(x)^k e^{-m(x)}} + \frac{1}{2} \tag{23}$$

is Lipschitz continuous on $[0, 1]^d$ with Lipschitz constant bounded by

$$c_{16}(k+1)\beta_n^{k+1}\, \log\, \beta_n \cdot \frac{1}{c_2^k},$$

where $c_2 = \min_{x \in [0,1]^d} m(x)$. Here we have used that m is bounded away from zero and infinity on $[0, 1]^d$ (since it is Lipschitz continuous and always greater than zero).

The function in (23) is always greater than or equal to 0.5. In this range the derivative of the log-function is bounded, and since with f_1 and f_2 also $f_1 \cdot f_2$ is Lipschitz continuous with Lipschitz constant bounded by

$$(\|f_1\|_\infty + \|f_2\|_\infty) \cdot (c_{\mathrm{Lip}}(f_1) + c_{\mathrm{Lip}}(f_2)),$$

we can conclude that

$$u \mapsto \log\left(\frac{\hat{g}(u)^k e^{-\hat{g}(u)} + m(x)^k e^{-m(x)}}{2m(x)^k e^{-m(x)}}\right) \cdot m(u)^k e^{-m(u)}$$

is on $[0, 1]^d$ continuous with Lipschitz constant bounded by

$$c_{17}(k \cdot \log \beta_n + \beta_n + c_{18}^k) \cdot \left((k+1) \cdot \beta_n^{k+2} \cdot \frac{1}{c_2^k} + (k+1) \cdot c_{19}^k\right) \leqslant c_{20}(k+1)^2 \beta_n^{k+3} \cdot \frac{1}{c_2^k}.$$

From this we conclude that $\phi_{n,x}$ is on $[0, 1]^d$ Lipschitz continuous with Lipschitz constant bounded by

$$c_n = \sum_{k=0}^{\infty} \frac{c_{20}(k+1)^2 \beta_n^{k+3}}{c_2^k k!} \leqslant c_{21} \beta_n^5 e^{\beta_n / c_2}.$$

With (8) we get the assertion. $\square$

To formulate our next lemma we need the notion of covering numbers. Let $x_1, \ldots, , x_n \in \mathbb{R}^d$ and set $x_1^n = (x_1, \ldots, x_n)$. Define the distance $d_1(f, g)$ between $f, g : \mathbb{R}^d \to \mathbb{R}$ by

$$d_1(f, g) = \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - g(x_i)|.$$

Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. An ε-cover of $\mathcal{F}$ (w.r.t. the distance d_1) is a set of functions $f_1, \ldots, f_k : \mathbb{R}^d \to \mathbb{R}$ with the property

$$\min_{1 \leqslant j \leqslant k} d_1(f, f_j) < \varepsilon \quad \text{for all } f \in \mathcal{F}.$$

Let $\mathcal{N}(\varepsilon, \mathcal{F}, x_1^n)$ denote the size k of the smallest ε-cover of $\mathcal{F}$ w.r.t. the distance d_1, and set $\mathcal{N}(\varepsilon, \mathcal{F}, x_1^n) = \infty$ if there does not exist any ε-cover of $\mathcal{F}$ of finite size.

Lemma 2. *Assume that the assumptions of Theorem* 1 *are satisfied. Set $k_n = \lceil \log n \rceil$ and let $\mathcal{H}_{n,1}$ be the set of all functions $h : \mathbb{R}^d \times \mathbb{N}_0 \to \mathbb{R}$ which satisfy*

$$h(x, y) = \log\left(\frac{\frac{g(x)^y}{y!} \cdot e^{-g(x)} + \frac{\alpha^y}{y!} \cdot e^{-\alpha}}{2 \cdot \frac{\alpha^y}{y!} \cdot e^{-\alpha}}\right) \cdot K\left(\frac{u - x}{h_n}\right) \cdot 1_{\{y \leqslant k_n\}} \quad (x \in \mathbb{R}^d, y \in \mathbb{N}_0)$$

for some $g \in \mathcal{G}_{M,\beta_n}$, $u \in [0,1]^d$ *and* $\alpha \in [c_2, c_3]$. *Then we have for any* $(x,y)_1^n \in (\mathbb{R}^d \times \mathbb{N}_0)^n$ *and any* $\varepsilon > 0$

$$\mathcal{N}\left(\frac{h_n^d \varepsilon}{8}, \mathcal{H}_{n,1}, (x,y)_1^n\right) \leqslant \left(c_{22} \frac{\beta_n^{k_n} \cdot k_n}{h_n^d \cdot \varepsilon}\right)^{c_{23}}$$

for some constants $c_{22}, c_{23} \in \mathbb{R}$.

Proof. Let $\mathcal{H}_{n,2}$ be the set of all functions $h_{n,2} : \mathbb{R}^d \times \mathbb{N}_0 \to \mathbb{R}$ which satisfy

$$h_{n,2}(x,y) = K\left(\frac{u-x}{h_n}\right) = \tilde{K}\left(\frac{\|u-x\|}{h_n}\right) \quad (x \in \mathbb{R}^d, y \in \mathbb{N}_0)$$

for some $u \in [0,1]^d$, and let $\mathcal{H}_{n,3}$ be the set of all functions $h_{n,3} : \mathbb{R}^d \times \mathbb{N}_0 \to \mathbb{R}$ which satisfy

$$h_{n,3}(x,y) = \log\left(\frac{\frac{g(x)^y}{y!} \cdot e^{-g(x)} + \frac{\alpha^y}{y!} \cdot e^{-\alpha}}{2 \cdot \frac{\alpha^y}{y!} \cdot e^{-\alpha}}\right) \cdot 1_{\{y \leqslant k_n\}} \quad (x \in \mathbb{R}^d, y \in \mathbb{N}_0)$$

for some $g \in \mathcal{G}_{M,\beta_n}$ and $\alpha \in [c_2, c_3]$. The functions in $\mathcal{H}_{n,2}$ and $\mathcal{H}_{n,3}$ are bounded in absolute value by B and $c_4 \cdot k_n \cdot \log n$ (cf. (21)) for n sufficiently large, resp. By Lemma 16.5 in [12] we have

$$\mathcal{N}\left(\frac{h_n^d \varepsilon}{8}, \mathcal{H}_{n,1}, (x,y)_1^n\right) \leqslant \mathcal{N}\left(\frac{h_n^d \varepsilon}{16 \cdot c_4 \cdot k_n \cdot \log n}, \mathcal{H}_{n,2}, (x,y)_1^n\right) \cdot \mathcal{N}\left(\frac{h_n^d \varepsilon}{16B}, \mathcal{H}_{n,3}, (x,y)_1^n\right).$$

Next we generalize the results of the eighth step in the proof of Theorem 2.1 in [18]. Since $\tilde{K}$ is of bounded variation it can be written as the difference of two monotone decreasing functions: $\tilde{K} = \tilde{K}_1 - \tilde{K}_2$ (see Corollary 2.7 of [37]). Let $\mathcal{G}$ be the collection of functions $\frac{\|u-x\|}{h}$ parameterized by $u \in \mathbb{R}^d$ and $h \in \mathbb{R}$. Also, let $\mathcal{H}_{n,2}^{(i)} = \{\tilde{K}_i(g(\cdot)) : g \in \mathcal{G}\}$ $(i = 1, 2)$. Clearly $\mathcal{H}_{n,2} = \{\tilde{K}(g(\cdot)) : g \in \mathcal{G}\}$. By Lemma 16.4 of [12] we have

$$\mathcal{N}(\delta, \mathcal{H}_{n,2}, (x,y)_1^n) \leqslant \mathcal{N}(\delta/2, \mathcal{H}_{n,2}^{(1)}, (x,y)_1^n) \mathcal{N}(\delta/2, \mathcal{H}_{n,2}^{(2)}, (x,y)_1^n), \tag{24}$$

$\delta > 0$ because $\mathcal{H}_{n,2} \subset \{f_1 - f_2 : f_1 \in \mathcal{H}_{n,2}^{(1)}, f_2 \in \mathcal{H}_{n,2}^{(2)}\}$. Since $\mathcal{G}$ spans a $(d+1)$-dimensional vector space, by Theorem 9.5 of [12] (see also [34]) the collection of sets

$$\mathcal{G}^+ = \{\{(x,t) : g(x) - t \geqslant 0\} : \ g \in \mathcal{G}\}$$

has VC dimension $V_{\mathcal{G}^+} \leqslant d+1$. Since $\tilde{K}_i$ is monotone, it follows from Lemma 16.3 of [12] that $V_{\mathcal{H}_{n,2}^{(i)+}} \leqslant d+1$. Let V_i be the total variations of $\tilde{K}_i$ $(i = 1, 2)$. Then $V = V_1 + V_2$ and $0 \leqslant \tilde{K}_i(x) \leqslant V_i$, $x \in \mathbb{R}$ $(i = 1, 2)$. Since $0 \leqslant f(x) \leqslant V$ for all $f \in \mathcal{H}_{n,2}$ and x, Theorem 9.4 of [12] and (24) imply

$$\begin{aligned}\mathcal{N}(\delta, \mathcal{H}_{n,2}, (x,y)_1^n) &\leqslant 3\left(\frac{2eV}{\delta} \log\left(\frac{3eV}{\delta}\right)\right)^{V_{\mathcal{H}_{n,2}^{(1)+}}} \cdot 3\left(\frac{2eV}{\delta} \log\left(\frac{3eV}{\delta}\right)\right)^{V_{\mathcal{H}_{n,2}^{(2)+}}} \\ &\leqslant 9\left(\frac{3eV}{\delta}\right)^{V_{\mathcal{H}_{n,2}^{(1)+}} + V_{\mathcal{H}_{n,2}^{(2)+}}}\end{aligned}$$

or

$$\mathcal{N}\left(\frac{h_n^d\varepsilon}{16c_4\cdot k_n\cdot \log n}, \mathcal{H}_{n,2}, (x,y)_1^n\right) \leqslant \left(\frac{c_{24}k_n\ \log n}{h_n^d\varepsilon}\right)^{2(d+1)}.$$

Let $y\leqslant k_n$ and consider the function

$$\phi(u,v) = \log\left(\frac{\frac{u^y}{y!}e^{-u}+\frac{v^y}{y!}e^{-v}}{2\frac{v^y}{y!}e^{-v}}\right) = \log\left(\frac{1}{2}\cdot u^y\cdot v^{-y}\cdot e^{v-u}+\frac{1}{2}\right) \quad (u\in[1/\beta_n,\beta_n], v\in[c_2,c_3]).$$

The partial derivatives of the function inside the log-function are for $y\leqslant k_n$ bounded in absolute value by $c_{25}\cdot k_n\cdot \beta_n^{2k_n}$. Since the log-function is on $[\frac{1}{2},\infty)$ Lipschitz continuous with Lipschitz constant 2, we can conclude that ϕ is for $y\leqslant k_n$ on $[1/\beta_n,\beta_n]\times[c_2,c_3]$ Lipschitz continuous with Lipschitz constant $c_{26}\cdot k_n\cdot \beta_n^{2k_n}$. By Lemma 16.4 of [12] which provides the bound on the covering number for the sums of families of bounded functions we get

$$\mathcal{N}\left(\frac{h_n^d\varepsilon}{16B}, \mathcal{H}_{n,3}, (x,y)_1^n\right) \leqslant \mathcal{N}\left(\frac{h_n^d\varepsilon}{c_{27}\cdot k_n\cdot \beta_n^{2k_n}}, \mathcal{H}_{n,4}, (x,y)_1^n\right) \cdot \mathcal{N}\left(\frac{h_n^d\varepsilon}{c_{27}\cdot k_n\cdot \beta_n^{2k_n}}, \mathcal{H}_{n,5}, (x,y)_1^n\right),$$

where $\mathcal{H}_{n,4}$ and $\mathcal{H}_{n,5}$ are the sets of all functions

$$h_{n,4}(x,y) = \frac{g(x)^y}{y!}\cdot e^{-g(x)} \quad (x\in\mathbb{R}^d, y\in\mathbb{N}_0)$$

with $g\in\mathcal{G}_{M,\beta_n}$, and

$$h_{n,5}(x,y) = \frac{\alpha^y}{y!}\cdot e^{-\alpha} \quad (x\in\mathbb{R}^d, y\in\mathbb{N}_0)$$

with $\alpha\in[c_2,c_3]$, resp., and we can assume w.l.o.g. $(x,y)_1^n\in(\mathbb{R}^d\times\{0,1,\ldots,k_n\})^n$ in the covering numbers on the right-hand side.

It is easy to see that for $y\leqslant k_n$ the derivative of $\psi(z)=z^ye^{-z}/(y!)$ is on $[0,\beta_n]$ bounded in absolute value by some constant times $k_n\beta_n^{k_n}$, which implies

$$\mathcal{N}\left(\frac{h_n^d\varepsilon}{c_{27}\cdot k_n\cdot \beta_n^{2k_n}}, \mathcal{H}_{n,4}, (x,y)_1^n\right) \leqslant \mathcal{N}\left(\frac{h_n^d\varepsilon}{c_{28}\cdot k_n^2\cdot \beta_n^{3k_n}}, \mathcal{G}_{M,\beta_n}, (x,y)_1^n\right) \leqslant \left(\frac{c_{29}\beta_n}{h_n^d\varepsilon/(k_n^2\cdot\beta_n^{3k_n})}\right)^{2(M+1)^d+2},$$

where the last inequality followed from monotonicity of the exponential function and Lemma 9.2, Theorems 9.4, 9.5 and Lemma 16.3 in [12].

Similarly we get

$$\mathcal{N}\left(\frac{h_n^d \varepsilon}{c_{27} \cdot k_n \cdot \beta_n^{2k_n}}, \mathcal{H}_{n,5}, (x, y)_1^n\right) \leqslant \frac{c_{30}}{h_n^d \varepsilon/(k_n^2 \cdot \beta_n^{3k_n})}$$

Putting together the above results we get the assertion. □

Acknowledgments

The authors wish to thank Jürgen Dippon and Jose Santos for several helpful discussions and the referees for constructive and helpful suggestions. Research of the second author was supported by the Natural and Sciences and Engineering Research Council of Canada and by the Alexander von Humboldt Foundation.

References

[1] P. Algoet, Universal schemes for learning the best nonlinear predictor given the infinite past and side information, IEEE Trans. Inform. Theory 45 (1999) 1165–1185.
[2] P. Algoet, L. Györfi, Strong universal pointwise consistency of some regression function estimation, J. Multivariate Anal. 71 (1999) 125–144.
[3] L. Birgé, Approximation dans les espaces métriques et théorie de l'estimation, Z. Wahrscheinlichkeitstheorie Verwandte Gebiete 65 (1983) 181–237.
[4] L. Birgé, P. Massart, Rates of convergence for minimum contrast estimators, Probab. Theory Related Fields 97 (1993) 113–150.
[5] D. Climov, J. Hart, L. Simar, Automatic smoothing and estimation in single index Poisson regression, J. Nonparametric Statist. 14 (2002) 307–323.
[6] L. Devroye, A Course in Density Estimation, Birkhäuser, Basel, 1987.
[7] L. Devroye, L. Györfi, Nonparametric Density Estimation: The L_1 View, Wiley, New York, 1985.
[8] L. Devroye, L. Györfi, A. Krzyżak, G. Lugosi, On the strong universal consistency of nearest neighbor regression function estimates, Ann. Statist. 22 (1994) 1371–1385.
[9] U. Einmahl, M. Mason, An empirical process approach to the uniform consistency of kernel-type function estimators, J. Theoret. Probab. 13 (2000) 1–37.
[10] J. Fan, M. Farmen, I. Gijbels, Local maximum likelihood estimation and inference, J. Roy. Statist. Soc. Ser. B 60 (1998) 591–608.
[11] J. Fan, I. Gijbels, Local Polynomial Modelling and its Applications, Chapman & Hall, London, 1996.
[12] L. Györfi, M. Kohler, A. Krzyżak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer Series in Statistics, Springer, New York, 2002.
[13] L. Györfi, H. Walk, On the strong universal consistency of a series type regression estimate, Math. Methods Statist. 5 (1996) 332–342.
[14] L. Györfi, H. Walk, On the strong universal consistency of a recursive regression estimate by Pál Révész, Statist. Probab. Lett. 31 (1997) 177–183.
[15] J. Hannig, T.C.M. Lee, On Poisson signal estimation under Kullback–Leibler discrepancy and squared risk, J. Statist. Plann. Inference 136 (2006) 882–908.
[16] T.J. Hastie, R.J. Tibshirani, Generalized Additive Models, Chapman & Hall, London, 1990.
[17] H.M. Hudson, T.C.M. Lee, Maximum likelihood restoration and choice of smoothing parameter in deconvolution of image data subject to Poisson noise, Comput. Statist. Data Anal. 26 (1998) 393–410.
[18] M. Kohler, Universal consistency of local polynomial kernel regression estimates, Ann. Inst. Statist. Math. 54 (2002) 879–899.
[19] M. Kohler, A. Krzyżak, Nonparametric regression estimation using penalized least squares, IEEE Trans. Inform. Theory 47 (2001) 3054–3058.
[20] E.D. Kolaczyk, Nonparametric estimation of gamma-ray burst intensities using Haar wavelets, Astrophys. J. 483 (1997) 340–349.
[21] A.S. Kozek, J.R. Leslie, E.F. Schuster, On a universal strong law of large numbers for conditional expectations, Bernoulli 4 (1998) 143–165.

[22] P.J. La Riviere, X. Pan, Nonparametric regression sinogram smoothing using a roughness-penalized Poisson likelihood of objective function, IEEE Trans. Med. Imaging 19 (2000) 773–786.
[23] L. Le Cam, On the assumptions used to prove asymptotic normality of maximum likelihood estimates, Ann. Math. Statist. 41 (1970) 802–828.
[24] L. Le Cam, Convergence of estimates under dimensionality restrictions, Ann. Statist. 1 (1973) 38–53.
[25] G. Lugosi, K. Zeger, Nonparametric estimation via empirical risk minimization, IEEE Trans. Inform. Theory 41 (1995) 677–687.
[26] P. McCullagh, J.A. Nelder, Generalized Linear Models. Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1983.
[27] E.A. Nadaraya, On estimating regression, Theory Probab. Appl. 9 (1964) 141–142.
[28] R.D. Nowak, E.D. Kolaczyk, A statistical multiscale framework for Poisson inverse problems, IEEE Trans. Inform. Theory 46 (2000) 1811–1825.
[29] F. O'Sullivan, B.S. Yandell, W.J. Raynon Jr., Automatic smoothing of regression functions in generalized linear models, J. Amer. Statist. Assoc. 81 (1986) 96–103.
[30] Y. Pawitan, F. O'Sullivan, Data-dependent bandwidth selection for emission computed tomography reconstruction, IEEE Trans. Med. Imaging 12 (2) (1993) 167–172.
[31] C. Spiegelman, J. Sacks, Consistent window estimation in nonparametric regression, Ann. Statist. 8 (1980) 240–246.
[32] C.J. Stone, Consistent nonparametric regression, Ann. Statist. 5 (1977) 595–645.
[33] S. van de Geer, Empirical Processes in M-estimation, Cambridge University Press, Cambridge, MA, 2000.
[34] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer, New York, 1996.
[35] H. Walk, Strong universal pointwise consistency of recursive regression estimates, Ann. Inst. Statist. Math. 53 (2001) 691–707.
[36] G.S. Watson, Smooth regression analysis, Sankhya Ser. A 26 (1964) 359–372.
[37] R.L. Wheeden, A. Zygmund, Measure and Integral, Marcel Dekker, New York, 1977.
[38] M. Yuan, Automatic smoothing for Poisson regression, Commun. Statist. Theory Methods 34 (2005) 603–617.