Asymptotic confidence intervals for Poisson regression *

Michael Kohler

Fachrichtung 6.1-Mathematik, Universität des Saarlandes, Postfach 151150,

D-66041 Saarbrücken, Germany, email: kohler@math.uni-sb.de

and

Adam Krzyżak

Department of Computer Science and Software Engineering, Concordia University, 1455

De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email:

krzyzak@cs.concordia.ca

August 11, 2005

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. Using this estimate of $m(x)$ we estimate the conditional distribution of Y given $X = x$ by a corresponding Poisson distribution and use this distribution to construct confidence intervals of level α of Y given $X = x$. Under mild regularity assumption on $m(x)$ and on the distribution of X we show that the corresponding confidence interval has asymptotically (i.e., for sample size tending to infinity) level α, and that the probability that the length

*Running title: *Confidence intervals for Poisson regression*

Please send correspondence and proofs to: Adam Krzyżak, Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec Canada H3G 1M8, email: krzyzak@cs.concordia.ca, phone: +1-514-848-2424, ext. 3007, fax: +1-514-848-2830.

of this confidence interval deviates from the optimal length by more than one converges to zero with the number of samples tending to infinity.

Key words and phrases: 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., Györfi et al. (2002) and the literature cited therein. In this paper we are interested in the special case that Y takes on with probability one only values 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. McCullagh and Nelder (1983)) 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), ... 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

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

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 $\hat{\mathbf{P}}_n\{Y = y|X = x\}$ 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., Devroye and Györfi (1985)), this automatically implies that the level of confidence regions of Y given $X = x$ based on $\hat{\mathbf{P}}_n\{Y = y|X = x\}$ 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 Fan, Farmen and Gijbels (1998) 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., Fan and Gijbels (1996)) 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).

Automatic methods for the choice of the bandwidth of the Nadaraya-Watson kernel estimate (cf. Nadaraya (1964), Watson (1964)) in Poisson regression have been investigated

in Climov, Hart and Simar (2002) and Hannig and Lee (2003), when 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 O'Sullivan, Yandell and Raynor (1986) and Yuan (2003). For related local maximum likelihood estimates the choice of the bandwidth was investigated in Fan, Farmen and Gijbels (1998) in particular in the context of nonparametric logistic regression.

In the proof 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 van de Geer (2000) (see also Le Cam (1970, 1973), Birgé (1983) and Birgé and Massart (1993)) and apply them to Poisson regression.

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. To define such a localized log-likelihood function, let $K : \mathbb{R}^d \to \mathbb{R}$ be a so-called kernel function, e.g., $K(u) = 1_{\{\|u\| \leq 1\}}$ (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

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

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

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

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

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

and

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

The bound on the coefficients in the definition of $\mathcal{F}_{M,\beta_n}$ implies

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

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

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

With this notation we define our estimate by

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

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

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

provided $\epsilon_n > 0$ is chosen such that

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

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), ... 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

$$X \in [0, 1]^d \quad a.s. \tag{2}$$

and

$$|m(x) - m(z)| \leq C_{lip}(m) \cdot \|x - z\| \quad (x, z \in \mathbb{R}^d) \tag{3}$$

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

Define the kernel function $K : \mathbb{R}^d \to \mathbb{R}_+$ by

$$K(u) = \tilde{K}(\|u\|^2) \quad (u \in \mathbb{R}^d)$$

for some $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ *which is monotone decreasing, left-continuous and satisfies for some* $r, R, b, B > 0$

$$b \cdot 1_{[0,r^2]}(v) \leq \tilde{K}(v) \leq B \cdot 1_{[0,R^2]}(v) \quad (v \in \mathbb{R}_+).$$

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

$$\beta_n \to \infty \quad (n \to \infty), \tag{4}$$

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

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

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

Define the estimate $\hat{\mathbf{P}}_n\{Y = y|X = x\}$ *as above. Then*

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

By a discrete version of Scheffe's theorem (which follows, e.g., from the proof of Theorem 1.1 in Devroye (1987)) we have for $x \in \mathbb{R}^d$

$$\begin{aligned} &\sum_{y=0}^{\infty} \left| \hat{\mathbf{P}}_n\{Y = y|X = x\} - \mathbf{P}\{Y = y|X = x\} \right| \\ &= 2 \sup_{A \subseteq \mathbb{N}_0} \left| \sum_{y \in A} \hat{\mathbf{P}}_n\{Y = y|X = x\} - \sum_{y \in A} \mathbf{P}\{Y = y|X = x\} \right|, \end{aligned} \tag{7}$$

therefore under the assumptions of Theorem 1 the integrated total variation distance between $\mathbf{P}\{Y = \cdot|X = x\}$ and $\hat{\mathbf{P}}_n\{Y = \cdot|X = x\}$ 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

$$\sum_{y \in \mathbb{N}_0, y \le u_n(x)} \hat{\mathbf{P}}_n\{Y = y | X = x\} \ge 1 - \alpha, \tag{8}$$

and set $I_n(x) = [0, u_n(x)]$. From Theorem 1 we can conclude

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

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

Proof. By (8) we have

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

Because of

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

(7) 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

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

If the case

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

occurs, a very small error in the estimate of $m(x)$ may result in $|u_n(x)-u(x)| \ge 1$. 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*

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

Proof. Set

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

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

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

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

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

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

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

From this we conclude that

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

implies

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

From this we get

$$\mathbf{P}\{|u_n(X) - u(X)| > 1\} \le \mathbf{P}\left\{ \sup_{A \subseteq \mathbb{N}_0} \left| \sum_{y \in A} \hat{\mathbf{P}}_n\{Y = y|X\} - \sum_{y \in A} \mathbf{P}\{Y = y|X\} \right| > c_1 \right\}.$$

By (7) and Theorem 1 we have

$$\begin{aligned}
&2 \cdot \mathbf{E} \sup_{A \subseteq \mathbb{N}_0} \left| \sum_{y \in A} \hat{\mathbf{P}}_n\{Y = y|X\} - \sum_{y \in A} \mathbf{P}\{Y = y|X\} \right| \\
&= \mathbf{E} \sum_{y=0}^{\infty} \left| \hat{\mathbf{P}}_n\{Y = y|X\} - \mathbf{P}\{Y = y|X\} \right| \\
&\le \mathbf{E} \sum_{y=0}^{\infty} \int \left| \hat{\mathbf{P}}_n\{Y = y|X = x\} - \mathbf{P}\{Y = y|X = x\} \right| \mathbf{P}_X(dx) \\
&\to 0 \quad (n \to \infty),
\end{aligned}$$

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 it is not required that X have a density with respect to the Lebesgue-Borel measure.

Remark 2. 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

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

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

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

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 Algoet (1999), Algoet and Györfi (1999), Kozek, Leslie and Schuster (1998) and Walk (2001). For related universal consistency result see, e.g., Stone (1977), Spiegelman and Sachs (1980), Devroye et al. (1994), Györfi and Walk (1996, 1997), Lugosi and Zeger (1995) and Kohler and Krzyżak (2001),

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 maximimum likelihood estimate which has not been considered in the papers above, but which seems to be especially suited in the context of this paper where the main aim is not estimation of the regression function but estimation of $\mathbf{P}\{Y=y|X=x\}$.

4 Outline of the proof of Theorem 1

In the proof of Theorem 1 we observe first that it suffices to show that the integrated Hellinger distance

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

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

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

where

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

Using the Lipschitz continuity of m we approximate this term by

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

where

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

By definition of the estimate and concavity of the log-function, the empirical version

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

of the nominator above is always greater than or equal to zero. 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

$$\int \sum_{y=0}^{\infty} \left|\hat{\mathbf{P}}_n\{Y=y|X=x\} - \mathbf{P}\{Y=y|X=x\}\right| \mathbf{P}_X(dx) \to 0 \quad a.s. \tag{9}$$

follows from

$$\int \sum_{y=0}^{\infty} \left(\sqrt{\hat{\mathbf{P}}_n\{Y=y|X=x\}} - \sqrt{\mathbf{P}\{Y=y|X=x\}}\right)^2 \mathbf{P}_X(dx) \to 0 \quad a.s. \tag{10}$$

For the sake of completeness we repeat a proof of this well-known fact (cf., e.g., Devroye and Györfi (1985)). Observe that for $a, b > 0$

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

and conclude from this and the Cauchy-Schwarz inequality

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

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

With

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

and

$$
\begin{aligned}
&\int \sqrt{\sum_{y=0}^{\infty} \left(\sqrt{\hat{\mathbf{P}}_n\{Y=y|X=x\}} - \sqrt{\mathbf{P}\{Y=y|X=x\}} \right)^2} \mathbf{P}_X(dx) \\
&\leq 1 \cdot \sqrt{\int \sum_{y=0}^{\infty} \left(\sqrt{\hat{\mathbf{P}}_n\{Y=y|X=x\}} - \sqrt{\mathbf{P}\{Y=y|X=x\}} \right)^2 \mathbf{P}_X(dx)}
\end{aligned}
$$

(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

$$
\begin{aligned}
&\int \sum_{y=0}^{\infty} \left(\sqrt{\hat{\mathbf{P}}_n\{Y=y|X=x\}} - \sqrt{\mathbf{P}\{Y=y|X=x\}} \right)^2 \mathbf{P}_X(dx) \\
&\leq -16 \cdot \mathbf{E} \left\{ \log \left(\frac{\hat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X\}}{2\mathbf{P}\{Y|X\}} \right) \middle| \mathcal{D}_n \right\} \qquad (11)
\end{aligned}
$$

where

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

By Lemma 4.2 and Lemma 1.3 in van de Geer (2000) we get

$$
\sum_{y=0}^{\infty} \left(\sqrt{\hat{\mathbf{P}}_n\{Y=y|X=x\}} - \sqrt{\mathbf{P}\{Y=y|X=x\}} \right)^2
$$

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

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 (11).

In the *third step of the proof* we show

$$
\begin{aligned}
&\mathbf{E}\left\{ \log \left(\frac{\hat{\mathbf{P}}_n\{Y|X\} + \mathbf{P}\{Y|X\}}{2 \cdot \mathbf{P}\{Y|X\}} \right) \middle| \mathcal{D}_n \right\} \\
&\qquad - \int \frac{\mathbf{E}\left\{ \log \left(\frac{\hat{\mathbf{P}}_x\{Y|X\} + \mathbf{P}_x\{Y|X\}}{2 \cdot \mathbf{P}_x\{Y|X\}} \right) \cdot K\left(\frac{x-X}{h_n} \right) \middle| \mathcal{D}_n \right\}}{\mathbf{E}K\left(\frac{x-X}{h_n} \right)} \mathbf{P}_X(dx) \to 0 \quad a.s. \quad (12)
\end{aligned}
$$

where

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

The first expectation on the left-hand side of (12) can be written as

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

Furthermore

$$
\frac{\mathbf{E}\left\{ \log \left(\frac{\hat{\mathbf{P}}_x\{Y|X\} + \mathbf{P}_x\{Y|X\}}{2 \cdot \mathbf{P}_x\{Y|X\}} \right) \cdot K\left(\frac{x-X}{h_n} \right) \middle| \mathcal{D}_n \right\}}{\mathbf{E}K\left(\frac{x-X}{h_n} \right)} = \frac{\int \phi_{n,x}(u) \cdot K\left(\frac{x-u}{h_n} \right) \mathbf{P}_X(du)}{\int K\left(\frac{x-u}{h_n} \right) \mathbf{P}_X(du)},
$$

where

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

Because of $m_n(x)=\hat{g}_x(x)$ 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 $x,u,v\in[0,1]^d$

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

(i.e., such that $\phi_{n,x}$ is Lipschitz continuous with Lipschitz constant c_n independent of x). Using this, we can bound the absolute value of the left-hand side of (12) by

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

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 Györfi et al. (2002)), 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

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\frac{\hat{g}_x(X_i)^{Y_i}}{Y_i!}\cdot e^{-\hat{g}_x(X_i)}+\frac{m(x)^{Y_i}}{Y_i!}\cdot e^{-m(x)}}{2\frac{m(x)^{Y_i}}{Y_i!}\cdot e^{-m(x)}}\right)\cdot K\left(\frac{x-X_i}{h_n}\right)\geq 0 \tag{13}$$

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

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

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

for all $a,b>0$ which implies

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

by definition of $\hat{g}_x$. This proves (13).

In the *fifth step of the proof* we set

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

and show that

$$\begin{aligned}
A_n:=\frac{1}{h_n^d}\cdot\sup_{x\in[0,1]^d}\left|\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\hat{\mathbf{P}}_x\{Y_i|X_i\}+\mathbf{P}_x\{Y_i|X_i\}}{2\cdot\mathbf{P}_x\{Y_i|X_i\}}\right)\cdot K\left(\frac{x-X_i}{h_n}\right)\right.\\
\left.-\mathbf{E}\left\{\log\left(\frac{\hat{\mathbf{P}}_x\{Y|X\}+\mathbf{P}_x\{Y|X\}}{2\cdot\mathbf{P}_x\{Y|X\}}\right)\cdot K\left(\frac{x-X}{h_n}\right)\middle|\mathcal{D}_n\right\}\right|\to 0\quad a.s.
\end{aligned}\tag{14}$$

implies the assertion.

From step 2 we conclude

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

where

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

and

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

By step 3 we have

$$B_n - C_n \to 0 \quad a.s.,$$

so by step 1 the assertion of Theorem 1 follows from

$$\limsup_{n\to\infty}(-C_n) \leq 0 \quad a.s. \tag{15}$$

Set

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

In step 4 we have shown

$$D_n \geq 0,$$

so

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

and (15) follows from

$$D_n - C_n \to 0 \quad a.s.$$

But this in turn is implied by (14), since

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

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 Kohler (2002).

In the *sixth (and final) step of the proof* we show (14). 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

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

where

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

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

and

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

For arbitrary $\epsilon > 0$ we get for n sufficiently large (because of

$$\begin{aligned}
|h(x,y)| &\le B\cdot\log\left(2\cdot\max\left\{(1/2)\cdot\left(\frac{g(x)}{\alpha}\right)^y e^{-g(x)+\alpha}, 1/2\right\}\right)\\
&\le B\cdot|y\cdot\log(g(x)/\alpha) - g(x) + \alpha|\\
&\le B\cdot(y\cdot\log(\beta_n/c_2) + c_3 + \beta_n) \le c_4\cdot y\cdot\log n \qquad (16)
\end{aligned}$$

for $x \in [0,1]^d$, $y \in \mathbb{N}$ and $h \in \mathcal{H}_n$, cf. (4)–(6)) by Markov inequality

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

Since

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

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

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

Similarly we get

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

So it remains to show

$$T_{1,n} \to 0 \quad a.s. \tag{17}$$

To do this, we apply Theorem 9.1 in Györfi et al. (2002) and Lemma 2 below. From these we get for an arbitrary $\epsilon > 0$

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

By the assumptions of Theorem 1 we have

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

Using this we get

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

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 $\epsilon > 0$. Application of the Borel-Cantelli lemma yields (17). 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 $x, u, v \in [0,1]^d$

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

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

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

from which we can conclude that the function

$$u \mapsto \frac{\hat{g}_x(u)^k e^{-\hat{g}_x(u)} + m(x)^k e^{-m(x)}}{2m(x)^k e^{-m(x)}} = \frac{\hat{g}_x(u)^k e^{-\hat{g}_x(u)}}{2m(x)^k e^{-m(x)}} + \frac{1}{2} \tag{18}$$

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 (18) 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

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

we can conclude that

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

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

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

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

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

With (5) we get the assertion. □

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

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

Let $\mathcal{N}(\epsilon, \mathcal{F}, x_1^n)$ denote the size k of the smallest ϵ–cover of $\mathcal{F}$ w.r.t. the distance d_1, and set $\mathcal{N}(\epsilon, \mathcal{F}, x_1^n) = \infty$ 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*

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

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* $\epsilon > 0$

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

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

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

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

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

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. (16)) for n sufficiently large, resp. By Lemma 16.5 in Györfi et al. (2002) we have

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

By the results of the eighth step in the proof of Theorem 2.1 in Kohler (2002) we have

$$\mathcal{N}\left(\frac{h_n^d \epsilon}{16 c_4 \cdot k_n \cdot \log n}, \mathcal{H}_{n,2}, (x,y)_1^n\right) \leq \left(\frac{c_{24} k_n \log n}{h_n^d \epsilon}\right)^{2(d+3)}.$$

Let $y \leq k_n$ 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 $y \leq k_n$ bounded in absolute value by

$$c_{25} \cdot k_n \cdot \beta_n^{2k_n}.$$

Since the log-function is on $[1/2, \infty)$ Lipschitz continuous with Lipschitz constant 2, we can conclude that ϕ is for $y \leq k_n$ on $[1/\beta_n, \beta_n] \times [c_2, c_3]$ Lipschitz continuous with Lipschitz constant

$$c_{26} \cdot k_n \cdot \beta_n^{2k_n}.$$

From this we get

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

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 $y \leq k_n$ the derivative of $\psi(z) = z^y e^{-z}/(y!)$ is on $[0, \beta_n]$ bounded in absolute value by some constant times $k_n \beta_n^{k_n}$, which implies

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

$$\leq \left(\frac{c_{29}\beta_n}{h_n^d \epsilon/(k_n^2 \cdot \beta_n^{3k_n})} \right)^{2(M+1)^d+2},$$

where the last inequality followed from monotonicity of the exponential function and Lemma 9.2, Theorem 9.4, Theorem 9.5 and Lemma 16.3 in Györfi et al. (2002).

Similarly we get

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

Putting together the above results we get the assertion. □

Acknowledgement

The authors wish to thank Jürgen Dippon and Jose Santos for several helpful discussions. 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] Algoet, P. (1999). Universal schemes for learning the best nonlinear predictor given the infinite past and side information. *IEEE Transactions on Information Theory*, **45**, pp. 1165–1185.

[2] Algoet, P. and Györfi, L. (1999). Strong universal pointwise consistency of some regression function estimation. *Journal of Multivariate Analysis*, **71**, pp. 125–144.

[3] Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **65**, pp. 181–237.

[4] Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, **97**, pp. 113–150.

[5] Climov, D., Hart, J. and Simar, L. (2002). Automatic smoothing and estimation in single index Poisson regression. *Journal of Nonparametric Statistics*, **14**, pp. 307–323.

[6] McCullagh, P. and Nelder, J. A. (1983). *Generalized linear models.* Monographs on Statistics and Applied Probability, Chapman & Hall, London.

[7] Devroye, L. (1987). *A Course in Density Estimation.* Birkhäuser.

[8] Devroye, L., Györfi, L., Krzyżak, A., and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, **22**, 1371–1385.

[9] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View.* John Wiley, New York.

[10] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman & Hall, London.

[11] Fan, J., Farmen, M. and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society, Series B,* **60**, pp. 591-608.

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

[13] Györfi, L. and Walk, H. (1996). On the strong universal consistency of a series type regression estimate. *Mathematical Methods of Statistics*, **5**, 332–342.

[14] Györfi, L. and Walk, H. (1997). On the strong universal consistency of a recursive regression estimate by Pál Révész. *Statistics and Probability Letters*, **31**, 177–183.

[15] Hannig, J. and Lee, T. C. M. (2003). On Poisson Signal Estimation under Kullback-Leibler Discrepancy and Squared Risk. Preprint, Colorado State University.

[16] Kohler, M. (2002). Universal consistency of local polynomial kernel regression estimates. *Annals of the Institute of Statistical Mathematics*, **54**, 879-899.

[17] Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory*, **47**, 3054–3058.

[18] Kozek, A. S., Leslie, J. R. and Schuster, E. F. (1998). On a universal strong law of large numbers for conditional expectations. *Bernoulli*, **4**, pp. 143–165.

[19] Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Annals of Mathematical Statistics*, **41**, pp. 802–828.

[20] Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, **1**, pp. 38–53.

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

[22] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications* **9**, pp. 141–142.

[23] O'Sullivan, F., Yandell, B. S., and Raynor, W. J., Jr. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association*, **81**, pp. 96–103.

[24] Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression. *Annals of Statistics*, **8**, 240–246.

[25] Stone, C.J. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, 595–645.

[26] van de Geer, S. (2000). *Empirical Processes in M–estimation.* Cambridge University Press.

[27] Walk, H. (2001). Strong universal pointwise consistency of recursive regression estimates. *Annals of the Institute of Statistical Mathematics*, **53**, pp. 691–707.

[28] Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, **26**, pp. 359–372.

[29] Yuan, M. (2003). Automatic Smoothing for Poisson Regression. Technical report no. 1083, Department of Statistics, University of Wisconsin.