Nonparametric estimation of conditional distributions *

László Györfi[1] and Michael Kohler[2]

[1] Department of Computer Science and Information Theory, Budapest University of Technology and Economics, 1521 Stoczek, U.2, Budapest, Hungary, email: gyorfi@szit.bme.hu

[2] Fachrichtung 6.1-Mathematik, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany, email: kohler@math.uni-sb.de

November 9, 2005

Abstract

Estimation of conditional distributions is considered. It is assumed that the conditional distribution is either discrete or that it has a density with respect to the Lebesgue-Borel-measure. Partitioning estimates of the conditional distribution are constructed and results concerning consistency and rate of convergence of the integrated total variation error of the estimates are presented.

Key words and phrases: conditional density, conditional distribution, confidence sets, partitioning estimate, Poisson regression, rate of convergence, universal consistency.

*Running title: *Estimation of conditional distributions*

Please send correspondence and proofs to: Michael Kohler, Fachrichtung 6.1-Mathematik, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany, email: kohler@math.uni-sb.de, phone +49-681-3022435, fax +49-681-3026583, e-mail: kohler@math.uni-sb.de

1 Introduction

One of the main tasks in statistics is to estimate a distribution from a given sample. Let μ be a probability distribution on $\mathbb{R}^d$ and let X_1, X_2, ... be independent and identically distributed random variables with distribution μ. A simple but powerful estimate of μ is the empirical distribution

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(X_i),$$

where I_A denotes the indicator function of the set A. By the strong law of large numbers we have

$$\mu_n(A) \to \mu(A) \quad a.s. \tag{1}$$

for each Borel set A. If we want to make some statistical inference about μ it is not enough to have (1) for each set individually, instead we need convergence of μ_n to μ uniformly over classes of sets. By the Glivenko-Cantelli theorem the empirical distribution satisfies

$$\sup_{x\in\mathbb{R}^d} |\mu_n((-\infty,x]) - \mu((-\infty,x])| \to 0 \quad a.s., \tag{2}$$

where $(-\infty,x] = (-\infty,x^{(1)}] \times \ldots \times (-\infty,x^{(d)}]$ for $x = (x^{(1)},\ldots,x^{(d)}) \in \mathbb{R}^d$. This is great in case that we want to make some statistical inference about intervals, but for more general investigations it would be much nicer if we are able to control the error in total variation defined as

$$\sup_{B\in\mathcal{B}_d} |\mu_n(B) - \mu(B)|, \tag{3}$$

where $\mathcal{B}_d$ are the Borel-sets in $\mathbb{R}^d$. Clearly, for the empirical distribution the error (3) does not converge to zero in general, since if μ has a continuous distribution function we have $\mu(\{X_1,\ldots,X_n\}) = 0$ and $\mu_n(\{X_1,\ldots,X_n\}) = 1$.

If we are able to construct estimates $\hat{\mu}_n$ of μ such that

$$\sup_{B\in\mathcal{B}_d} |\hat{\mu}_n(B) - \mu(B)| \to 0 \quad a.s., \tag{4}$$

then it is easy to construct confidence sets $\hat{B}_n$ for the values of X_1 such that they have asymptotically level α for given $\alpha \in (0,1)$, i.e. such that

$$\liminf_{n\to\infty} \mu(\hat{B}_n) \geq 1-\alpha \quad a.s.$$

Indeed, any set $\hat{B}_n$ with

$$\hat{\mu}_n(\hat{B}_n) \geq 1-\alpha$$

has this property since

$$\begin{aligned} \mu(\hat{B}_n) &= \mu_n(\hat{B}_n) - (\mu_n(\hat{B}_n) - \mu(\hat{B}_n)) \\ &\geq 1-\alpha - \sup_{B\in\mathcal{B}_d} |\mu_n(B) - \mu(B)|. \end{aligned}$$

Unfortunatley, as was shown in Devroye and Györfi (1990), it is impossible to construct estimates $\hat{\mu}_n$ such that (4) holds for all distributions μ. However, it follows from Barron et al. (1992) that in case we restrict ourselves to distributions where the nonatomic part is absolutely continuous with respect to a known dominating measure, it is possible to construct estimates such that (4) holds for all such distributions. Special cases include discrete measures (where we assume for notational convenience that $\mu(\mathbb{N}_0) = 1$) and measures which have a density with respect to the Lebesgue-Borel-measure. By Scheffe's theorem it suffices in these cases to construct estimates $(\hat{\mu}_n(\{k\}))_{k\in\mathbb{N}_0}$ of $(\mu(\{k\}))_{k\in\mathbb{N}_0}$ and estimates $\hat{f}_n$ of the density f of μ, resp., which satisfy

$$\sum_{k=0}^{\infty} |\hat{\mu}_n(\{k\}) - \mu(\{k\})| \to 0 \quad a.s. \tag{5}$$

and

$$\int |f_n(x) - f(x)| \lambda(dx) \to 0 \quad a.s., \tag{6}$$

where λ denotes the Lebesgue-Borel-measure. Here one estimates $\mu(B)$ by

$$\hat{\mu}_n(B) = \sum_{k\in\mathbb{N}_0} \hat{\mu}_n(\{k\}) \quad \text{and} \quad \hat{\mu}_n(B) = \int_B \hat{f}_n(x)\, dx, \quad \text{resp.}$$

Many estimates which satisfy (6) universally for all densities are constructed in Devroye and Györfi (1985a).

In this paper we want to apply the above ideas in the regression context. Here we have given independent and identically distributed random vectors (X, Y), (X_1, Y_1), ... with values in $\mathbb{R}^d \times \mathbb{R}^{d'}$. Given the sample

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

of the distribution of (X, Y) we want to construct estimates $\hat{\mathbf{P}}_n\{B|x\}$ of the conditional distribution $\mathbf{P}\{Y \in B|X = x\}$ of Y given X such that

$$\int \sup_{B \in \mathcal{B}_d} \left| \hat{\mathbf{P}}_n\{B|x\} - \mathbf{P}\{Y \in B|X = x\} \right| \mu(dx) \to 0 \quad a.s., \tag{7}$$

where μ denotes again the distribution of X. In contrast to standard regression, where $d' = 1$ and where only the mean $\mathbf{E}\{Y|X = x\}$ of the conditional distribution is estimated (cf., e.g., Györfi et al. (2002)), we can use estimates with the property (7) not only for prediction of the value of Y for given value of X, but also to construct confidence regions for the value of Y given the value of X. Indeed, similarly as above one gets that (7) implies that any set $C_n(x)$ with

$$\hat{\mathbf{P}}_n\{C_n(x)|x\} \geq 1 - \alpha$$

satisfies

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

since we have with $\mathbf{P}^*\{\cdot\} = \mathbf{P}\{\cdot|\mathcal{D}_n\}$

$$\begin{aligned}
&\mathbf{P}\{Y \in C_n(X)|\mathcal{D}_n\} \\
&= \int \mathbf{P}^*\{Y \in C_n(x)|X = x\}\mu(dx) \\
&\geq \int \hat{\mathbf{P}}_n\{C_n(x)|x\}\mu(dx) - \int \left| \hat{\mathbf{P}}_n\{C_n(x)|x\} - \mathbf{P}^*\{Y \in C_n(x)|X = x\} \right| \mu(dx) \\
&\geq 1 - \alpha - \int \sup_{B \in \mathcal{B}_n} \left| \hat{\mathbf{P}}_n(B|x) - \mathbf{P}\{Y \in B|X = x\} \right| \mu(dx).
\end{aligned}$$

In order to construct estimates with the property (7), we consider two special cases: In the first case the conditional distribution of Y given X is discrete (and for notational convenience we assume again that the support is contained in $\mathbb{N}_0$). In the second case the conditional distribution of Y given $X = x$ has a density $f(\cdot|x)$ with respect to the Lebesgue-Borel-measure. In both cases Scheffe's theorem implies that in order to have (7) we have to construct estimates of $\mathbf{P}\{Y = k|X = x\}$ and $f(\cdot|x)$ such that

$$\int \sum_{k=0}^{\infty} \left| \hat{\mathbf{P}}_n\{k|x\} - \mathbf{P}\{Y = k|X = x\} \right| \mu(dx) \to 0 \quad a.s. \tag{8}$$

and

$$\int \int |f_n(y|x) - f(y|x)| \, \lambda(dy)\mu(dx) \to 0 \quad a.s., \tag{9}$$

resp.

In order to construct in the first case estimates with the property (8) we use two different approaches: In the first approach we consider for each $y \in \mathbb{N}_0$

$$\mathbf{P}\{Y = y|X = x\} = \mathbf{E}\{I_{\{Y=y\}}|X = x\}$$

as a regression function and estimate it by applying a partitioning estimate to a sample of $(X, I_{\{Y=y\}})$. In the second approach we consider Poisson regression, i.e., we make a parametric assumption on the way the conditional distribution of Y given $X = x$ depends on $m(x)$ and assume that

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

for some $m : \mathbb{R}^d \to (0, \infty)$, where m is completely unknown. In this case we estimate $m(x) = \mathbf{E}\{Y|X = x\}$ by a partitioning estimate $m_n(x)$ applied to a sample of (X, Y), and consider the plug-in estimate

$$\hat{\mathbf{P}}_n\{Y = y|X = x\} = \frac{m_n(x)^y}{y!} \cdot e^{-m_n(x)} \quad (y \in \mathbb{N}_0).$$

In both approaches we present results concerning universal consistency, i.e. we show (8) for all corresponding discrete conditional distributions, and we analyze the rate of convergence of the estimates.

Estimates of the conditional density in the second case are defined as suitable partitioning estimates. We present results concerning universal consistency, i.e., we show (9) for all conditional distributions with density, and we analyze the rate of convergence under regularity assumptions on the smoothness of the conditional density.

The paper is organized as follows: Our main results concerning estimation of discrete conditional distributions and conditional densities are described in Section 2 and 3, resp. The proofs are given in Section 4.

2 The estimation of discrete conditional distributions

In this section we study partitioning estimates of discrete conditional distributions. In our first two theorems each conditional probability $\mathbf{P}\{Y = y|X = x\}$ is estimated separately. We have the following result concerning consistency of the estimate.

Theorem 1 *Let $\mathcal{P}_n = \{A_{n,j} : j\}$ be a partition of $\mathbb{R}^d$ and for $x \in \mathbb{R}^d$ denote by $A_n(x)$ that cell $A_{n,j}$ of $\mathcal{P}_n$ that contains x. Let*

$$\hat{\mathbf{P}}_n\{y|x\} = \frac{\sum_{i=1}^n I_{A_n(x)}(X_i) \cdot I_{\{Y_i=y\}}}{\sum_{j=1}^n I_{A_n(x)}(X_j)}$$

be the partitioning estimate of $\mathbf{P}\{Y = y|X = x\}$. Assume that the underlying partitioning $\mathcal{P}_n = \{A_{n,j} : j\}$ satisfies for each sphere S centered at the origin

$$\lim_{n\to\infty} \max_{j:A_{n,j}\cap S} diam(A_{n,j}) = 0 \tag{10}$$

and

$$\lim_{n\to\infty} \frac{|\{j : A_{n,j} \cap S \neq \emptyset\}|}{n} = 0, \tag{11}$$

where diam(A) denotes the diameter of the set A. Then

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

Next we consider the rate of convergence of the above partitioning estimate. It is well-known that in order to derive non-trivial rate of convergence results in nonparametric regression one needs smoothness assumption on the underlying regression function (cf., Devroye (1982)). In our next result we assume that the conditional probabilities are locally Lipschitz continuous, such that the integral over the sum of the Lipschitz constant is finite.

Theorem 2 *Assume X is bounded a.s.,*

$$|\mathbf{P}\{Y = y|X = x\} - \mathbf{P}\{Y = y|X = z\}| \leq C_y(x) \cdot \|x - z\|$$

for all x, z from the bounded support of X and for some local Lipschitz constants $C_y(x)$ satisfying

$$\int \sum_{y=0}^{\infty} C_y(x)\mu(dx) = C^* < \infty,$$

and assume

$$\sum_{y=0}^{\infty} \sqrt{\mathbf{P}\{Y = y\}} < \infty.$$

Let $\hat{\mathbf{P}}_n\{y|x\}$ be the partitioning estimate of $\mathbf{P}\{Y = y|X = x\}$ with respect to a partition of $\mathbb{R}^d$ consisting of cubes with side-length h_n. Then

$$\mathbf{E} \int \sum_{y=0}^{\infty} \left| \hat{\mathbf{P}}_n\{y|x\} - \mathbf{P}\{Y = y|X = x\} \right| \mu(dx)$$
$$\leq c_1 \left(1 + \sum_{y=0}^{\infty} \sqrt{\mathbf{P}\{Y = y\}} \right) \cdot \frac{1}{\sqrt{n \cdot h_n^d}} + \sqrt{d} \cdot C^* \cdot h_n,$$

so for

$$h_n = c_2 \cdot n^{-1/(d+2)}$$

we get

$$\mathbf{E}\int \sum_{y=0}^{\infty} \left|\hat{\mathbf{P}}_n\{y|x\} - \mathbf{P}\{Y=y|X=x\}\right| \mu(dx) \le c_3 \cdot n^{-\frac{1}{d+2}}.$$

In the next theorem we consider Poisson regression. Here the conditional distribution of Y given X is given by

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

for some $m : \mathbb{R}^d \to (0,\infty)$. Because of $m(x) = \mathbf{E}\{Y|X=x\}$ we can estimate it by applying a partitioning estimate to $\mathcal{D}_n$ and use a plug-in estimate

$$\hat{\mathbf{P}}_n\{y|x\} = \frac{m_n(x)^y}{y!} \cdot e^{-m_n(x)} \quad (y \in \mathbb{N}_0)$$

to estimate the conditional distribution of Y given X. For this estimate we have the following result.

Theorem 3 *Assume that* $\mathbf{E}\{Y\} < \infty$ *and*

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

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

$$m_n(x) = \begin{cases} \frac{\sum_{i=1}^n I_{A_n(x)}(X_i)\cdot Y_i}{\sum_{i=1}^n I_{A_n(x)}(X_i)} & \text{if} \quad \sum_{i=1}^n I_{A_n(x)}(X_i) > \log n \\ 0 & \text{otherwise.} \end{cases}$$

be the (modified) partitioning estimate of m *with partition* $\mathcal{P}_n = \{A_{n,j} : j\}$ *and set*

$$\hat{\mathbf{P}}_n\{y|x\} = \frac{m_n(x)^y}{y!} \cdot e^{-m_n(x)} \quad (y \in \mathbb{N}_0).$$

a) *Assume that the underlying partition* $\mathcal{P}_n$ *satisfies (10) and for each sphere* S *centered at the origin*

$$\lim_{n\to\infty} \frac{|\{j : A_{n,j} \cap S \neq \emptyset\}| \log n}{n} = 0. \tag{12}$$

Then

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

b) *Assume X is bounded a.s. and assume that $\mathbf{E}\{Y^2\} < \infty$ and m is Lipschitz continuous, i.e.*

$$|m(x) - m(z)| \leq C \cdot \|x - z\|$$

for some constant $C \in \mathbb{R}_+$. Choose the underlying partition such that it consists of cubes of side-length h_n. Then

$$\mathbf{E} \int \sum_{y=0}^{\infty} \left| \hat{\mathbf{P}}_n\{y|x\} - \mathbf{P}\{Y = y|X = x\} \right| \mu(dx) \qquad \leq \frac{c_4}{\sqrt{n \cdot h_n^d}} + c_5 \cdot h_n,$$

so for

$$h_n = c_6 \cdot n^{-1/(d+2)}$$

we get

$$\mathbf{E} \int \sum_{y=0}^{\infty} \left| \hat{\mathbf{P}}_n\{y|x\} - \mathbf{P}\{Y = y|X = x\} \right| \mu(dx) \leq c_7 \cdot n^{-\frac{1}{d+2}}.$$

Remark 1. Assume that the assumptions of Theorem 3 b) hold. The function $f(u) = u^y e^{-u}/(y!)$ satisfies for $u \in [0, B]$

$$|f'(u)| = \left| \frac{y \cdot u^{y-1}}{y!} \cdot e^{-u} - \frac{u^y}{y!} \cdot e^{-u} \right| \leq (B+1) \cdot \frac{B^{y-1}}{(y-1)!},$$

so by boundedness of the Lipschitz continuous regression function m we get for $y > 0$

$$|\mathbf{P}\{Y = y|X = x\} - \mathbf{P}\{Y = y|X = z\}| \leq (B+1) \cdot \frac{B^{y-1}}{(y-1)!} \cdot C \cdot \|x - z\|.$$

This implies that the conditional probabilities are Lipschitz continuous and that the integral over the sum of the Lipschitz constant is bounded by

$$\left(1 + \sum_{y=1}^{\infty} (B+1) \cdot \frac{B^{y-1}}{(y-1)!} \right) \cdot C = \left(1 + (B+1) \cdot e^B \right) \cdot C,$$

hence under the assumption of Theorem 3 b) the estimate in Theorem 2 achieves the same rate of convergence although it does not depend on the particular form of the conditional distribution.

Remark 2. Under more restrictive regularity assumptions on the underlying distribution consistency of a localized log-likelihood Poisson regression estimate was shown in Kohler and Krzyżak (2005).

3 The estimation of conditional densities

In this section assume that Y takes values in $\mathbb{R}^{d'}$. Our aim is to estimate the conditional distribution of Y given X consistently in total variation. We assume that Y has absolutely continuous distribution and the conditional density of Y given X is denoted by

$$f(y|x).$$

For estimating $f(y|x)$, introduce a histogram estimate. Let $\mathcal{Q}_n = \{B_{n,j} \ : \ j\}$ be a partition of $\mathbb{R}^{d'}$, such that the Lebesgue measure λ of each cell is positive and finite. Let $B_n(y)$ be the cell of $\mathcal{Q}_n$ into which y falls. As before let $\mathcal{P}_n = \{A_{n,j} : j\}$ be a partition of $\mathbb{R}^d$ and denote the cell into which x falls by $A_n(x)$.

Put

$$\nu_n(A, B) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \in A, Y_i \in B\}},$$

then the histogram estimate is as follows:

$$f_n(y|x) = \frac{\nu_n(A_n(x), B_n(y))}{\mu_n(A_n(x)) \cdot \lambda(B_n(y))}.$$

We will use the following conditions: assume that for each sphere S centered at the origin we have

$$\lim_{n\to\infty} \max_{j: B_{n,j} \cap S} diam(B_{n,j}) = 0 \tag{13}$$

and

$$\lim_{n\to\infty} \frac{|\{j \ : \ B_{n,j} \cap S \neq \emptyset\}|}{n} = 0. \tag{14}$$

The next theorem extends the density-free strong consistency result of Abou-Jaoude (1976) to conditional density estimation.

Theorem 4 *Assume that the partitions $\mathcal{P}_n$ and $\mathcal{Q}_n$ satisfy (10), (11), (13) and (14), resp. Then*

$$\int\int |f_n(y|x) - f(y|x)|\, \lambda(dy)\mu(dx) \to 0 \quad a.s.$$

Devroye and Györfi (1985a), and Beirlant and Györfi (1998) calculated the rate of convergence of the expected L_1 error of the histogram. Next we extend these results to the estimates of conditional densities.

Theorem 5 *Assume X and Y are bounded a.s., and*

$$|f(u|x) - f(y|x)| \le C_1(x) \cdot \|u - y\|$$

and

$$|f(y|z) - f(y|x)| \le C_2(y) \cdot \|x - z\|$$

for all x, z from the bounded support of X and for all y, u from the bounded support of Y such that

$$\int C_1(z)\mu(dz) < \infty$$

and

$$\int C_2(y)\lambda(dy) < \infty.$$

Let $f_n(y|x)$ be the histogram estimate of $f(y|x)$ with respect to a partitions $\mathcal{P}_n$ and $\mathcal{Q}_n$ consisiting of cubes with side-lengths h_n and H_n, resp. Then

$$\begin{aligned}&\mathbf{E}\int\int |f_n(y|x) - f(y|x)|\, \lambda(dy)\mu(dx)\\ &\le \sqrt{\frac{c_8}{n \cdot h_n^d}} + \sqrt{\frac{c_9}{n \cdot h_n^d \cdot H_n^{d'}}} + \sqrt{d} \cdot c_{10} \cdot h_n + \sqrt{d'} \cdot c_{11} \cdot H_n,\end{aligned}$$

so for

$$h_n = c_{12} \cdot n^{-1/(d+d'+2)} \quad \textit{and} \quad H_n = c_{13} \cdot n^{-1/(d+d'+2)}$$

we get

$$\mathbf{E}\int\int |f_n(y|x) - f(y|x)|\, \lambda(dy)\mu(dx) \le c_{14} \cdot n^{-\frac{1}{d+d'+2}}.$$

4 Proofs

4.1 Proof of Theorem 1

Using

$$|a-b| = 2(b-a)_+ + (a-b)$$

(where $x_+ = \max\{x, 0\}$) we get

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

Using the Cauchy-Schwarz inequality and Theorem 23.1 in Györfi et al. (2002) we get for each fixed $y \in \mathbb{N}_0$

$$\begin{aligned}
&\int \left(\mathbf{P}\{Y=y|X=x\} - \hat{\mathbf{P}}_n\{y|x\} \right)_+ \mathbf{P}_X(dx) \\
&\leq \int \left| \hat{\mathbf{P}}_n\{y|x\} - \mathbf{P}\{Y=y|X=x\} \right| \mu(dx) \\
&\leq \sqrt{\int \left(\hat{\mathbf{P}}_n\{y|x\} - \mathbf{P}\{Y=y|X=x\} \right)^2 \mu(dx)} \to 0 \quad a.s.,
\end{aligned}$$

which implies together with the dominated convergence theorem, that the first term on the right–hand side above converges to zero.

Concerning the second term we observe

$$\begin{aligned}
&\int \sum_{y=0}^{\infty} \hat{\mathbf{P}}_n\{y|x\} \mu(dx) - \int \sum_{y=0}^{\infty} \mathbf{P}\{Y=y|X=x\} \mu(dx) \\
&= \int \left(\sum_{y=0}^{\infty} \frac{\sum_{i=1}^n I_{A_n(x)}(X_i) \cdot I_{\{Y_i=y\}}}{\sum_{j=1}^n I_{A_n(x)}(X_j)} - 1 \right) \mu(dx) \\
&= \int \left(I_{\left\{\sum_{j=1}^n I_{A_n(x)}(X_j) > 0\right\}} - 1 \right) \mu(dx) \\
&= -\sum_{j=0}^{\infty} I_{\left\{\sum_{i=1}^n I_{A_{n,j}}(X_i) = 0\right\}} \cdot \mu\{A_{n,j}\}.
\end{aligned}$$

Together with (11), it implies that

$$\begin{aligned}
&\left| \int \sum_{y=0}^{\infty} \hat{\mathbf{P}}_n\{y|x\}\mu(dx) - \int \sum_{y=0}^{\infty} \mathbf{P}\{Y=y|X=x\}\mu(dx) \right| \\
&\leq \sum_{j=0}^{\infty} |\mu\{A_{n,j}\} - \mu_n\{A_{n,j}\}| \\
&\to 0
\end{aligned}$$

a.s. (cf. Lemma 1 in Devroye and Györfi (1985b) or, with better constant in the exponential upper bound, cf. the proof of Lemma 23.2 in Györfi et al. (2002)). □

4.2 Proof of Theorem 2

In the sequel we use the notation

$$\nu_{y,n}(A) = \frac{1}{n} \sum_{i=1}^{n} I_{\{Y_i=y, X_i \in A\}},$$

and with this notation the partition estimate is given by

$$\hat{\mathbf{P}}_n\{y|x\} = \frac{\nu_{y,n}(A_n(x))}{\mu_n(A_n(x))}.$$

Thus,

$$\begin{aligned}
&\int \sum_{y=0}^{\infty} |\hat{\mathbf{P}}_n\{y|x\} - \mathbf{P}\{Y=y|X=x\}|\mu(dx) \\
&= \sum_{y=0}^{\infty} \int \left| \frac{\nu_{y,n}(A_n(x))}{\mu_n(A_n(x))} - \mathbf{P}\{Y=y|X=x\} \right| \mu(dx) \\
&= \sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \int_A \left| \frac{\nu_{y,n}(A)}{\mu_n(A)} - \mathbf{P}\{Y=y|X=x\} \right| \mu(dx) \\
&\leq \sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \int_A \left| \frac{\nu_{y,n}(A)}{\mu_n(A)} - \frac{\nu_{y,n}(A)}{\mu(A)} \right| \mu(dx) \\
&\quad + \sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \int_A \left| \frac{\nu_{y,n}(A)}{\mu(A)} - \frac{\mathbf{P}\{Y=y, X \in A\}}{\mu(A)} \right| \mu(dx) \\
&\quad + \sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \int_A \left| \frac{\mathbf{P}\{Y=y, X \in A\}}{\mu(A)} - \mathbf{P}\{Y=y|X=x\} \right| \mu(dx)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{y=0}^{\infty} \sum_{A\in\mathcal{P}_n} \left| \frac{\nu_{y,n}(A)}{\mu_n(A)} - \frac{\nu_{y,n}(A)}{\mu(A)} \right| \mu(A) \\
&\quad + \sum_{y=0}^{\infty} \sum_{A\in\mathcal{P}_n} \left| \frac{\nu_{y,n}(A)}{\mu(A)} - \frac{\mathbf{P}\{Y=y, X\in A\}}{\mu(A)} \right| \mu(A) \\
&\quad + \sum_{y=0}^{\infty} \sum_{A\in\mathcal{P}_n} \int_A \left| \frac{\mathbf{P}\{Y=y, X\in A\}}{\mu(A)} - \mathbf{P}\{Y=y|X=x\} \right| \mu(dx) \\
&= \sum_{A\in\mathcal{P}_n} \sum_{y=0}^{\infty} \nu_{y,n}(A) \cdot \left| \frac{1}{\mu_n(A)} - \frac{1}{\mu(A)} \right| \mu(A) \\
&\quad + \sum_{y=0}^{\infty} \sum_{A\in\mathcal{P}_n} |\nu_{y,n}(A) - \mathbf{P}\{Y=y, X\in A\}| \\
&\quad + \sum_{y=0}^{\infty} \sum_{A\in\mathcal{P}_n} \int_A \left| \frac{\mathbf{P}\{Y=y, X\in A\}}{\mu(A)} - \mathbf{P}\{Y=y|X=x\} \right| \mu(dx) \\
&\leq \sum_{A\in\mathcal{P}_n} |\mu_n(A) - \mu(A)| \\
&\quad + \sum_{y=0}^{\infty} \sum_{A\in\mathcal{P}_n} |\nu_{y,n}(A) - \mathbf{P}\{Y=y, X\in A\}| \\
&\quad + \sum_{y=0}^{\infty} \sum_{A\in\mathcal{P}_n} \int_A \left| \frac{\mathbf{P}\{Y=y, X\in A\}}{\mu(A)} - \mathbf{P}\{Y=y|X=x\} \right| \mu(dx),
\end{aligned}
$$

where we have used for the last inequality that

$$
\sum_{y=0}^{\infty} \nu_{y,n}(A) = \mu_n(A).
$$

Since $n\cdot\mu_n(A)$ is binomially distributed with parameters n and $\mu(A)$ we get by Cauchy-Schwarz inequality

$$
\begin{aligned}
\sum_{A\in\mathcal{P}_n} \mathbf{E}\{|\mu_n(A) - \mu(A)|\} &\leq \sum_{A\in\mathcal{P}_n} \sqrt{\mathbf{E}\{(\mu_n(A) - \mu(A))^2\}} \\
&\leq \sum_{A\in\mathcal{P}_n} \sqrt{\frac{\mu(A)}{n}}.
\end{aligned}
$$

By Jensen inequality we have

$$
\left(\frac{a_1 + \ldots + a_l}{l} \right)^2 \leq \frac{a_1^2 + \ldots a_l^2}{l},
$$

which implies

$$a_1 + \ldots + a_l \leq \sqrt{l \cdot (a_1^2 + \ldots a_l^2)}.$$

Using this inequality in the sum above for the c_{15}/h_n^d many cells $A \in \mathcal{P}_n$ contained in the bounded support of X (which are the only ones with $\mu(A) \neq 0$) we conclude

$$\begin{aligned}
\sum_{A \in \mathcal{P}_n} \mathbf{E}\{|\mu_n(A) - \mu(A)|\} &\leq \sqrt{\frac{c_{15}}{h_n^d} \cdot \sum_{A \in \mathcal{P}_n} \left(\sqrt{\mu(A)/n}\right)^2} \\
&= \sqrt{\frac{c_{15}}{n \cdot h_n^d} \cdot \sum_{A \in \mathcal{P}_n} \mu(A)} \\
&\leq \sqrt{\frac{c_{15}}{n \cdot h_n^d}}.
\end{aligned}$$

Similarly we get

$$\begin{aligned}
&\sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \mathbf{E}\{|\nu_{y,n}(A) - \mathbf{P}\{Y = y, X \in A\}|\} \\
&\leq \sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \sqrt{\mathbf{E}\{(\nu_{y,n}(A) - \mathbf{P}\{Y = y, X \in A\})^2\}} \\
&\leq \sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \sqrt{\frac{\mathbf{P}\{Y = y, X \in A\}}{n}} \\
&\leq \sum_{y=0}^{\infty} \sqrt{\frac{c_{15} \sum_{A \in \mathcal{P}_n} \mathbf{P}\{Y = y, X \in A\}}{n \cdot h_n^d}} \\
&= \sum_{y=0}^{\infty} \sqrt{\frac{c_{15} \mathbf{P}\{Y = y\}}{n \cdot h_n^d}} \\
&= \sqrt{\frac{c_{15}}{n \cdot h_n^d}} \sum_{y=0}^{\infty} \sqrt{\mathbf{P}\{Y = y\}}.
\end{aligned}$$

Finally

$$\begin{aligned}
&\sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \int_A \left| \frac{\mathbf{P}\{Y = y, X \in A\}}{\mu(A)} - \mathbf{P}\{Y = y | X = x\} \right| \mu(dx) \\
&\leq \sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \int_A \left| \frac{\int_A \mathbf{P}\{Y = y | X = z\} \mu(dz)}{\mu(A)} - \frac{\int_A \mathbf{P}\{Y = y | X = x\} \mu(dz)}{\mu(A)} \right| \mu(dx) \\
&\leq \sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \int_A \frac{\int_A |\mathbf{P}\{Y = y | X = z\} - \mathbf{P}\{Y = y | X = x\}| \mu(dz)}{\mu(A)} \mu(dx)
\end{aligned}$$

$$\leq \sum_{y=0}^{\infty} \sum_{A \in \mathcal{P}_n} \int_A \frac{C_y(x) \cdot diam(A) \cdot \mu(A)}{\mu(A)} \mu(dx)$$
$$\leq \sqrt{d} \cdot h_n \cdot \sum_{y=0}^{\infty} \int C_y(x) \mu(dx)$$
$$\leq \sqrt{d} \cdot h_n \cdot C^*.$$

Summarizing the above results, the assertion follows. □

4.3 Proof of Theorem 3

In the proof we will use the following lemma.

Lemma 1 *For arbitrary $u, v \in \mathbb{R}_+$ we have*

$$\sum_{j=0}^{\infty} \left| \frac{u^j}{j!} \cdot e^{-u} - \frac{v^j}{j!} \cdot e^{-v} \right| \leq 2 \cdot |u - v|.$$

Proof. W.l.o.g. assume $u < v$. Then

$$\begin{aligned}
&\sum_{j=0}^{\infty} \left| \frac{u^j}{j!} e^{-u} - \frac{v^j}{j!} e^{-v} \right| \\
&\leq \sum_{j=0}^{\infty} \left| \frac{u^j}{j!} e^{-u} - \frac{u^j}{j!} e^{-v} \right| + \sum_{j=0}^{\infty} \left| \frac{u^j}{j!} e^{-v} - \frac{v^j}{j!} e^{-v} \right| \\
&= \sum_{j=0}^{\infty} \left(\frac{u^j}{j!} e^{-u} - \frac{u^j}{j!} e^{-v} \right) + \sum_{j=0}^{\infty} \left(\frac{v^j}{j!} e^{-v} - \frac{u^j}{j!} e^{-v} \right) \\
&= e^u \left(e^{-u} - e^{-v} \right) + e^v e^{-v} - e^u e^{-v} \\
&= 2 \left(1 - e^{-(v-u)} \right) \\
&\leq 2|v - u|,
\end{aligned}$$

since $1 + x \leq e^x$ $(x \in \mathbb{R})$. □

Proof of Theorem 3. Proof of a): By Lemma 1 we get

$$\begin{aligned}
&\int \sum_{y=0}^{\infty} \left| \hat{\mathbf{P}}_n\{y|x\} - \mathbf{P}\{Y = y|X = x\} \right| \mu(dx) \\
&= \int \sum_{y=0}^{\infty} \left| \frac{m_n(x)^y}{y!} \cdot e^{-m_n(x)} - \frac{m(x)^y}{y!} \cdot e^{-m(x)} \right| \mu(dx)
\end{aligned}$$

$$\leq 2 \cdot \int |m_n(x) - m(x)| \mu(dx) \tag{15}$$
$$\to 0$$

a.s. by Györfi (1991) (see also Theorems 23.3 in Györfi et al. (2002)).

Proof of Part b): Using (15),

$$\mathbf{E}\left\{ \int \sum_{y=0}^{\infty} \left| \hat{\mathbf{P}}_n\{y|x\} - \mathbf{P}\{Y = y|X = x\} \right| \mu(dx) \right\} \leq 2 \cdot \sqrt{\mathbf{E}\left\{ \int |m_n(x) - m(x)|^2 \mu(dx) \right\}}$$
$$\leq \frac{c_4}{\sqrt{n \cdot h_n^d}} + c_5 \cdot h_n,$$

where the last step can be done in a similar way as the proof of Theorem 4.3 in Györfi et al. (2002). □

4.4 Proof of Theorem 4

Introduce the notation

$$\nu(A, B) = \mathbf{E}\{\nu_n(A, B)\} = \mathbf{P}\{X \in A, Y \in B\},$$

then

$$\int \int |f_n(y|x) - f(y|x)| \, \lambda(dy)\mu(dx)$$
$$= \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} \int_A \int_B \left| \frac{\nu_n(A, B)}{\mu_n(A) \cdot \lambda(B)} - f(y|x) \right| \lambda(dy)\mu(dx)$$
$$\leq \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} \int_A \int_B \left| \frac{\nu_n(A, B)}{\mu_n(A) \cdot \lambda(B)} - \frac{\nu_n(A, B)}{\mu(A) \cdot \lambda(B)} \right| \lambda(dy)\mu(dx)$$
$$+ \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} \int_A \int_B \left| \frac{\nu_n(A, B)}{\mu(A) \cdot \lambda(B)} - \frac{\nu(A, B)}{\mu(A) \cdot \lambda(B)} \right| \lambda(dy)\mu(dx)$$
$$+ \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} \int_A \int_B \left| \frac{\nu(A, B)}{\mu(A) \cdot \lambda(B)} - f(y|x) \right| \lambda(dy)\mu(dx)$$
$$= \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} \left| \frac{\nu_n(A, B)}{\mu_n(A) \cdot \lambda(B)} - \frac{\nu_n(A, B)}{\mu(A) \cdot \lambda(B)} \right| \mu(A)\lambda(B)$$

$$+\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\left|\frac{\nu_n(A,B)}{\mu(A)\cdot\lambda(B)}-\frac{\nu(A,B)}{\mu(A)\cdot\lambda(B)}\right|\mu(A)\lambda(B)$$
$$+\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B\left|\frac{\nu(A,B)}{\mu(A)\cdot\lambda(B)}-f(y|x)\right|\lambda(dy)\mu(dx),$$

therefore

$$\int\int|f_n(y|x)-f(y|x)|\,\lambda(dy)\mu(dx)$$

$$\leq\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\nu_n(A,B)\left|\frac{1}{\mu_n(A)}-\frac{1}{\mu(A)}\right|\mu(A)$$
$$+\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}|\nu_n(A,B)-\nu(A,B)|$$
$$+\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B\left|\frac{\nu(A,B)}{\mu(A)\cdot\lambda(B)}-f(y|x)\right|\lambda(dy)\mu(dx)$$

$$\leq\sum_{A\in\mathcal{P}_n}|\mu_n(A)-\mu(A)| \tag{16}$$
$$+\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}|\nu_n(A,B)-\nu(A,B)| \tag{17}$$
$$+\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B\left|\frac{\nu(A,B)}{\mu(A)\cdot\lambda(B)}-f(y|x)\right|\lambda(dy)\mu(dx), \tag{18}$$

where we have used for the last inequality that

$$\sum_{B\in\mathcal{Q}_n}\nu_n(A,B)=\mu_n(A).$$

Because of (11), (16) tends to 0 a.s., while (11) and (14) imply that (17) tends to 0 a.s. (cf. Lemma 1 in Devroye and Györfi (1985b)).

Concerning the convergence of the bias term (18), introduce the notation

$$\bar{f}_n(y|x)=\frac{\int_{A_n(x)}\int_{B_n(y)}f(u|z)\lambda(du)\mu(dz)}{\mu(A_n(x))\cdot\lambda(B_n(y))}$$

then

$$\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B\left|\frac{\nu(A,B)}{\mu(A)\cdot\lambda(B)}-f(y|x)\right|\lambda(dy)\mu(dx)$$

$$=\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B\left|\frac{\int_A\int_B f(u|z)\lambda(du)\mu(dz)}{\mu(A)\cdot\lambda(B)}-f(y|x)\right|\lambda(dy)\mu(dx)$$

$$= \int \int \left| \bar{f}_n(y|x) - f(y|x) \right| \lambda(dy)\mu(dx)$$

$$\to 0,$$

because of the conditions (10) and (13). This convergence is obvious if $f(y|x)$ is continuous and has compact support. In general, we use that $f(y|x) \in L_1(\mu \times \lambda)$, and refer to the denseness result such that the set of continuous functions in $L_1(\mu \times \lambda)$ with compact support is dense in $L_1(\mu \times \lambda)$ (cf., e.g., Devroye and Györfi (2002)). An alternative technique would be the Lebesgue density theorem (cf., e.g., Lemma 24.5 in Györfi et al. (2002)), which is a pointwise convergence, and together with the Scheefe theorem and the dominated convergence theorem we are ready. □

4.5 Proof of Theorem 5

Because of the proof of Theorem 4,

$$\begin{aligned}
&\mathbf{E}\left\{ \int \int |f_n(y|x) - f(y|x)| \, \lambda(dy)\mu(dx) \right\} \\
&\leq \sum_{A \in \mathcal{P}_n} \mathbf{E}\left\{ |\mu_n(A) - \mu(A)| \right\} \\
&\quad + \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} \mathbf{E}\left\{ |\nu_n(A,B) - \nu(A,B)| \right\} \\
&\quad + \int \int \left| \frac{\nu(A_n(x), B_n(y))}{\mu(A_n(x)) \cdot \lambda(B_n(y))} - f(y|x) \right| \lambda(dy)\mu(dx).
\end{aligned}$$

According to the proof of Theorem 2, the condition that X is bounded implies that

$$\sum_{A \in \mathcal{P}_n} \mathbf{E}\left\{ |\mu_n(A) - \mu(A)| \right\} \leq \sqrt{\frac{c_{15}}{n \cdot h_n^d}},$$

and, similarly, using X and Y are bounded we can show

$$\sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} \mathbf{E}\left\{ |\nu_n(A,B) - \nu(A,B)| \right\} \leq \sqrt{\frac{c_{16}}{n \cdot h_n^d \cdot H_n^{d'}}}.$$

Concerning the rate of convergence of the bias term we observe

$$\int \int \left| \frac{\nu(A_n(x), B_n(y))}{\mu(A_n(x)) \cdot \lambda(B_n(y))} - f(y|x) \right| \lambda(dy)\mu(dx)$$

$$= \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B \left|\frac{\nu(A,B)}{\mu(A)\cdot\lambda(B)} - f(y|x)\right| \lambda(dy)\mu(dx)$$

$$= \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B \left|\frac{\int_A\int_B f(u|z)\lambda(du)\mu(dz)}{\mu(A)\cdot\lambda(B)} - f(y|x)\right| \lambda(dy)\mu(dx)$$

$$\leq \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B \frac{\int_A\int_B |f(u|z)-f(y|x)|\lambda(du)\mu(dz)}{\mu(A)\cdot\lambda(B)}\lambda(dy)\mu(dx)$$

$$\leq \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B \frac{\int_A\int_B |f(u|z)-f(y|z)|\lambda(du)\mu(dz)}{\mu(A)\cdot\lambda(B)}\lambda(dy)\mu(dx)$$
$$+ \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B \frac{\int_A\int_B |f(y|z)-f(y|x)|\lambda(du)\mu(dz)}{\mu(A)\cdot\lambda(B)}\lambda(dy)\mu(dx).$$

Applying the conditions the theorem we get that

$$\int\int \left|\frac{\nu(A_n(x),B_n(y))}{\mu(A_n(x))\cdot\lambda(B_n(y))} - f(y|x)\right| \mu(dx)\lambda(dy)$$

$$\leq \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B \frac{\int_A\int_B C_1(z)\cdot\sqrt{d'}\cdot H_n\lambda(du)\mu(dz)}{\mu(A)\cdot\lambda(B)}\lambda(dy)\mu(dx)$$
$$+ \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\int_A\int_B \frac{\int_A\int_B C_2(y)\cdot\sqrt{d}\cdot h_n\lambda(du)\mu(dz)}{\mu(A)\cdot\lambda(B)}\lambda(dy)\mu(dx)$$

$$= \int C_1(z)\mu(dz)\lambda(S_Y)\cdot\sqrt{d'}\cdot H_n + \int C_2(y)\lambda(dy)\cdot\sqrt{d}\cdot h_n,$$

where S_Y is the bounded support of Y. □

References

[1] Abou-Jaoude, S. (1976). Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité. *Annales de l'Institut Henri Poincaré* , XII, 213-231.

[2] Barron, A. R., Györfi, L. and van der Meulen, E. C. (1992). Distribution estimation consistent in total variation and two types of information divergence. *IEEE Trans. Information Theory* **38**, pp. 1437-1454.

[3] Beirlant, J. and Györfi, L. (1998). On the L_1 error in histogram density estimation: the multidimensional case. *Nonparametric Statistics* **9**, pp. 197-216.

[4] Devroye, L. (1982). Any discrimination rule can have arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **4**, 154–157.

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

[6] Devroye, L. and Györfi, L. (1985b). Distribution-free exponential bound for the L_1 error of partitioning estimates of a regression function. In *Probability and Statistical Decision Theory*, F. Konecny, J. Mogyoródi, W. Wertz, Eds., D. Reidel, pp. 67-76.

[7] Devroye, L. and Györfi, L. (1990). No empirical measure can converge in the total variation sense for all distribution. *Annals of Statistics* **18**, pp.1496-1499.

[8] Devroye, L. and Györfi, L. (2002). Distribution and density estimation. In *Principles of Nonparametric Learning*, L. Györfi (Ed.), Springer-Verlag, Wien, pp. 223-286.

[9] Györfi, L. (1991). Universal consistency of a regression estimate for unbounded regression functions, *Nonparametric functional estimation and related topics* (ed. G. Roussas), 329–338, NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

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

[11] Kohler, M. and Krzyżak, A. Asymptotic confidence intervals for Poisson regression. Submitted for publication, 2005.