ESAIM: PS 23 (2019) 350–386
https://doi.org/10.1051/ps/2018026

ESAIM: Probability and Statistics
www.esaim-ps.org

STATISTICAL ESTIMATION OF CONDITIONAL SHANNON ENTROPY

Alexander Bulinski[*] and Alexey Kozhevin

Abstract. The new estimates of the conditional Shannon entropy are introduced in the framework of the model describing a discrete response variable depending on a vector of d factors having a density w.r.t. the Lebesgue measure in $\mathbb{R}^d$. Namely, the mixed-pair model (X, Y) is considered where X and Y take values in $\mathbb{R}^d$ and an arbitrary finite set, respectively. Such models include, for instance, the famous logistic regression. In contrast to the well-known Kozachenko–Leonenko estimates of unconditional entropy the proposed estimates are constructed by means of the certain spacial order statistics (or k-nearest neighbor statistics where $k = k_n$ depends on amount of observations n) and a random number of i.i.d. observations contained in the balls of specified random radii. The asymptotic unbiasedness and L^2-consistency of the new estimates are established under simple conditions. The obtained results can be applied to the feature selection problem which is important, *e.g.*, for medical and biological investigations.

Mathematics Subject Classification. 60F25, 62G20, 62H12.

Received June 28, 2018. Accepted November 28, 2018.

1. Introduction

The entropy concept plays a prominent role in physics and mathematics, see, *e.g.*, [3]. On various approaches to the entropy definition we refer to the deep works by Boltzmann, Gibbs, Plank, Shannon, Kolmogorov, Sinai, Renyi, Tsallis, Holevo. There are important problems where one employs statistical estimates of due entropy constructed by means of i.i.d. observations. For example, such estimates are useful in feature selection theory [35] and in detection of texture inhomogeneities [1]. We leave apart many other domains where entropy estimates are applied, see, *e.g.*, [33]. There are a number of various approaches to the entropy estimation, we refer, *e.g.*, to [2, 7, 11, 19, 28, 34, 36, 41–43].

The main goal of the paper is to introduce new statistical estimates of conditional Shannon entropy for models where a discrete response variable, taking values in an arbitrary finite set, depends on a vector of factors (features) having density w.r.t. the Lebesgue measure in $\mathbb{R}^d$. These models include the famous logistic regression (see, *e.g.*, [21, 23]). The proposed estimates involve the k-nearest neighbor statistics where $k = k_n$ depends on a number of observations n (on the k-nearest neighbor statistics see, *e.g.*, the recent book [6]). Note that our estimates do not employ the well-known Kozachenko-Leonenko statistics [24] used for estimation of the unconditional Shannon entropy. Under simple assumptions (*cf.*, *e.g.*, [5, 10, 14, 18, 40]) we establish the

Keywords and phrases: Shannon entropy, conditional entropy estimates, asymptotic unbiasedness, L^2-consistency, logistic regression, Gaussian model.

Steklov Mathematical Institute of Russian Academy of Sciences, Moscow, Russia.

[*] Corresponding author: bulinski@yandex.ru

Article published by EDP Sciences

asymptotic unbiasedness and L^2-consistency of our estimates when the sample size tends to infinity. An interest in the study of conditional entropy is explained as follows. The mutual information of two random vectors is represented by means of conditional entropy of one of them and unconditional entropy of another. That information characteristic of two random vectors facilitates the identification of relevant factors having impact on a response variable under consideration (see, *e.g.*, [4, 16, 17, 32, 44]). Such analysis is useful in medical and biological studies. Thus statistical estimates of the mutual information involving new estimates will be valuable for feature selection.

We stipulate that all the random variables and random vectors are defined on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Recall that the Shannon entropy (see [38]) of a discrete random variable Y taking values in a finite set M with probabilities $P(y) := \mathsf{P}(Y = y)$, $y \in M$, and a (differential) entropy of a random vector X in $\mathbb{R}^d$ having density $f(\cdot)$, $x \in \mathbb{R}^d$, w.r.t. the Lebesgue measure μ are introduced by the following respective formulas

$$H(Y) := -\mathsf{E}\log P(Y) = -\sum_{y\in M} P(y)\log P(y), \tag{1.1}$$

$$H(X) := -\mathsf{E}\log f(X) = -\int_{\mathbb{R}^d} f(x)\log f(x)\,\mu(\mathrm{d}x). \tag{1.2}$$

Clearly, one can view the entropy as a function of a probability distribution since the above formulas involve the laws of X and Y. Note that the probability distribution discretization techniques for a random variable having a density (w.r.t. the Lebesgue measure) and evaluation of the Shannon entropy for thus arising random variables do not lead to the differential entropy as the mesh of the discretization tends to zero (see, *e.g.*, Thm. 8.3.1 in [13] and [29]). More generally, when a measure σ is fixed on a measure space $(S, \mathcal{B})$, one can define the notion of the entropy of a probability measure ν given on the same space and absolutely continuous w.r.t. σ. Namely, whenever the following integral is well defined (and can take infinite values),

$$H_\sigma(\nu) := -\int_S \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\sigma}\right)\mathrm{d}\nu \tag{1.3}$$

where $\frac{\mathrm{d}\nu}{\mathrm{d}\sigma}$ is the Radon–Nikodym derivative.

If Y has a law ν on $(M, 2^M)$ then (1.1) is a particular case of (1.3) where $S = M$, $\mathcal{B} = 2^M$ and σ is a counting measure on M. If X has a law ν on $(S, \mathcal{B})$ then (1.3) leads to (1.2) when $S = \mathbb{R}^d$, $\mathcal{B} = \mathcal{B}(\mathbb{R}^d)$ and $\sigma = \mu$. The definition of the Kullback–Leibler (see, *e.g.*, [13], p. 19, 251) relative entropy (or divergence) for two probability measures is closely related to (1.3). We refer to [37] where various kinds of f-divergences are compared.

Consider a random vector (X, Y) such that $X : \Omega \to \mathbb{R}^d$ $(d \in \mathbb{N})$ and $Y : \Omega \to M$. Here M is an arbitrary finite set. We assume that $\mathsf{P}(Y = y) > 0$ for each $y \in M$. Suppose that there exists a measurable function $f_{X,Y} : \mathbb{R}^d \times M \to \mathbb{R}_+$ such that, for any $B \in \mathcal{B}(\mathbb{R}^d)$ and $y \in M$,

$$\mathsf{P}(X \in B, Y = y) = \int_B f_{X,Y}(x, y)\,\mu(\mathrm{d}x). \tag{1.4}$$

In other words, $f_{X,Y}$ is a density of a random vector (X, Y) w.r.t. measure $\sigma := \mu \otimes \lambda$ on $\mathcal{B}(\mathbb{R}^d) \otimes 2^M$. For $x \in \mathbb{R}^d$ and $y \in M$, let us define the following functions:

$$f_X(x) := \sum_{y\in M} f_{X,Y}(x, y),$$

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x, y)}{\mathsf{P}(Y = y)},$$

$$f_{Y|X}(y|x) := \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)}, & f_X(x) > 0, \\ 0, & f_X(x) = 0. \end{cases} \tag{1.5}$$

Note that f_X is a density of X, $f_{X|Y}$ is a conditional density of X given Y, and $f_{Y|X}$ provides a conditional distribution of Y given X. To simplify notation we will write $\mathrm{d}x$ instead of $\mu(\mathrm{d}x)$ and set $f(x,y) := f_{X,Y}(x,y)$, $f(y|x) := f_{Y|X}(y|x)$.

According to (1.3) (see also [31]) the entropy of a vector (X,Y) in the framework of model (1.4) is given by the formula

$$H(X,Y) := -\mathsf{E}\log f(X,Y) = -\sum_{y\in M}\int_{\mathbb{R}^d} f(x,y)\log f(x,y)\,\mathrm{d}x.$$

Introduce the conditional entropy of Y given X

$$H(Y|X) := -\mathsf{E}\log f(Y|X) = -\sum_{y\in M}\int_{\mathbb{R}^d} f(x,y)\log f(y|x)\,\mathrm{d}x. \tag{1.6}$$

One can verify that this conditional entropy $H(Y|X)$ is always finite.

The mutual information of X and Y is defined as $I(X,Y) := D_{KL}(P_{X,Y}||P_X\otimes P_Y)$, *i.e.* it is the Kullback–Leibler divergence D_{KL} between the distribution of a vector (X,Y) and a product of X and Y laws. Thus

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{1.7}$$

whenever all these expressions are well-defined (*i.e.* the expressions are excluded where one adds infinities with different signs). Note that $H(Y)$ and $H(Y|X)$ are finite for the model under consideration and therefore $I(X,Y)$ is also finite. It is well-known that $I(X,Y)\geq 0$. Moreover, $I(X,Y)=0$ if and only if X and Y are independent. The latter statement is applied to the information approach for the identification of relevant factors having an impact on a random response. Mention in passing that extension of (1.7) to the case of n random vectors is fruitful as well (see, *e.g.*, [15]). There are a number of papers devoted to various estimates of (unconditional) entropy. In this regard we indicate the recent work [10] where the estimates of the Shannon differential entropy are studied and where one can find further references.

The scheme (1.4) under consideration comprises the famous logistic model widely used in the classification problems (see, *e.g.*, [26]). Namely, let $M=\{1,2\}$ and

$$\mathsf{P}(Y=1|X=x) = \frac{1}{1+\exp\{-(w,x)-b\}},\quad x\in\mathbb{R}^d,\quad w\in\mathbb{R}^d,\quad b\in\mathbb{R}, \tag{1.8}$$

where $(\cdot,\cdot)$ is a scalar product in $\mathbb{R}^d$ and $\mathsf{P}(Y=2|X=x) = 1-\mathsf{P}(Y=1|X=x)$. Let f_X be a vector X density. Then

$$\begin{aligned} f_{X,Y}(x,1) &= \mathsf{P}(Y=1|X=x)f_X(x),\\ f_{X,Y}(x,2) &= f_X(x) - f_{X,Y}(x,1). \end{aligned}$$

Note that there exist generalizations of logistic regression where a response variable Y takes more than two different values.

To conclude the introduction we mention that in Section 2 statistical estimates of $H(Y|X)$ are introduced and two principle results are formulated. The proposed estimates are constructed by means of the certain k-nearest neighbor statistics (where $k=k_n$ depends on a number n of i.i.d. observations $(X_1,Y_1),\dots,(X_n,Y_n)$) and a random number of observations contained in the balls of specified random radii. Under wide conditions the asymptotic unbiasedness and L^2-consistency of our estimates are proved in Sections 4 and 5, respectively, whereas in Section 3 some auxiliary results are provided. Their proofs and that of corollary are given in Appendix. The applications to the feature selection problems along with simulations will be considered separately. In particular, for considered vectors (X, Y) our estimate of the conditional entropy $H(Y|X)$ has advantages over estimates constructed as differences of statistical estimates of $H(X, Y)$ and $H(X)$. Note also that other estimates of mutual information for discrete-continuous mixtures models based on the Kraskov–Stögbauer–Grassberger [25] approach were studied in [12, 18] under different conditions. Also it is worth to emphasize that our estimates construction does not suppose the existence of any topological structure on a set M (thus we do not use the distances between Y_i and Y_j, $i, j = 1, \dots, n$).

2. Main results

Let $Z_1, Z_2, \dots$ be a sequence of i.i.d. random vectors $Z_i = (X_i, Y_i)$, $i \in \mathbb{N}$, such that a distribution of Z_1 coincides with one of the vector (X, Y) described by model (1.4). Introduce the estimate $H(Y|X)$ constructed by a sample $Z_1, \dots, Z_n$ as follows

$$\widehat{H}_{n,k} = \frac{1}{n} \sum_{i=1}^{n} \widehat{H}_{n,k,i}. \tag{2.1}$$

Here $n \in \mathbb{N}$, $n > 1$, $k = k(n) \in \{1, \dots, n-1\}$,

$$\widehat{H}_{n,k,i} = -\log(\xi_{n,k,i}(Z_1, \dots, Z_n) + 1) + \log k, \tag{2.2}$$

$$\xi_{n,k,i}(Z_1, \dots, Z_n) := \sharp\{j \in \{1, \dots, n\} \setminus \{i\} \colon Y_j = Y_i, \|X_i - X_j\| \le \rho_{n,k,i}(X_1, \dots, X_n)\}, \tag{2.3}$$

$\sharp$ stands for the cardinality of a finite set, $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$ and

$$\rho_{n,k,i}(X_1, \dots, X_n) := \|X_i - X_{i,(k)}\|, \tag{2.4}$$

$X_{i,(k)}$ being the kth nearest neighbor of X_i in the sample $\{X_1, \dots, X_n\} \setminus \{X_i\}$ *i.e.* $\rho_{n,k,i}(X_1, \dots, X_n)$ is the Eulidean distance from X_i to its kth nearest neighbor. Clearly, the random variable $\xi_{n,k,i}(Z_1, \dots, Z_n)$ takes values $0, 1, \dots, k$. Observe that with probability one the points $X_1, \dots, X_n$ do not pair-wise coincide as the vector X has a density.

Thus, in contrast to the well-known Kozachenko–Leonenko [24] estimate of the Shannon differential entropy of a random vector, along with the distance to the kth nearest neighbor of X_i in the sample $X_1, \dots, X_n$ (without point X_i) the principle role is played by random variables $\xi_{n,k,i}$, $i = 1, \dots, n$. Namely, at first we find a random set $J \subset \{1, \dots, n\}$, consisting of all the indexes $j \in \{1, \dots, n\} \setminus \{i\}$ such that X_j belongs to the ball $B(X_i, \rho_{n,k,i})$ with a random center and a random radius. Then from the collection $\{(X_j, Y_j), j \in J\}$ we take $\{(X_j, Y_j), j \in I_i\}$ where $I_i := \{j \in J : Y_j = Y_i\}$. The collection of random variables $\{(X_j, Y_j), j \in I_i\}$ arises where I_i is also a random set. The cardinality of this set I_i, *i.e.* $\sharp I_i$, equals to the random variable $\xi_{n,k,i}$, $i = 1, \dots, n$.

Definition 2.1. A function $g \colon \mathbb{R}^d \to \mathbb{R}$ is called locally constricted at a point x in $\mathbb{R}^d$ if there exist strictly positive $R_0(x)$ and $C_0(x)$ such that

$$\left| g(x) - \frac{1}{|B(x, R)|} \int_{B(x,R)} g(v) \, \mathrm{d}v \right| \le C_0(x) R \quad \text{for} \quad R \in (0, R_0(x)) \tag{2.5}$$

where $|B(x, R)|$ is a ball $B(x, R) := \{v \in \mathbb{R}^d : \|v - x\| \le R\}$ volume, *i.e.* $|B(x, R)| = \mu(B(x, R))$. A function g is C_0-constricted if it is locally constricted for μ-almost all points $x \in \mathbb{R}^d$ and, moreover, for such x one has $C_0(x) \le C_0$ and $R_0(x) \ge R_0$ where C_0 and R_0 are strictly positive constants.

Remark 2.2. If a function $g : \mathbb{R}^d \to \mathbb{R}$ satisfies the Lipschitz condition at $x \in \mathbb{R}^d$ with a factor $C(x)$, that is $|g(v) - g(x)| \le C(x)\|v - x\|$ for all $v \in \mathbb{R}^d$, then (2.5) is valid for any $R_0(x) > 0$. It is easily seen that if $g(x)$, $x \in \mathbb{R}^d$, is a density of non-degenerate Gaussian law then this function is C_0-constricted.

Theorem 2.3. *Let in the framework of model* (1.4) *the following conditions be satisfied. For each fixed $y \in M$ and μ-almost all $x \in \mathbb{R}^d$, a function $f(x, y)$, i.e. $f(\cdot, y)$, is strictly positive and C_0-constricted,*

$$k = k_n \propto n^\alpha \tag{2.6}$$

for some $\alpha \in (0, 1)$, and, for some $\varepsilon > 0$,

$$\mathsf{E}|\log f_X(X)|^{1+\varepsilon} < \infty \tag{2.7}$$

where $f_X(\cdot)$ is a density of X. Then

$$\mathsf{E}\widehat{H}_{n,k} \to H(Y|X), \ \ n \to \infty, \tag{2.8}$$

i.e. $\widehat{H}_{n,k}$ is an asymptotically unbiased estimate of $H(Y|X)$.

Theorem 2.4. *Let the condition* (2.7) *of Theorem 2.3 be replaced by the following one. For some $\varepsilon > 0$,*

$$\mathsf{E}|\log f_X(X)|^{2+\varepsilon} < \infty. \tag{2.9}$$

Then

$$\mathsf{E}(\widehat{H}_{n,k} - H(Y|X))^2 \to 0 \ \ \textit{as} \ \ n \to \infty,$$

i.e. $\widehat{H}_{n,k}$ is an L_2-consistent estimate of $H(Y|X)$.

Corollary 2.5. *Let in the framework of model* (1.4)*, for each $y \in M$, the function $f(\cdot, y)$ be a density of non-degenerate Gaussian law in $\mathbb{R}^d$ (with mean vector and covariance matrix depending on y). Then $\widehat{H}_{n,k}$, where $k(n)$ satisfies* (2.6)*, are asymptotically unbiased and L^2-consistent estimates of $H(Y|X)$ as $n \to \infty$.*

3. Auxiliary results

In this section, as previously, we consider i.i.d. vectors $Z_1, Z_2, \ldots$, having the same distribution as (X, Y) in the framework of model (1.4). Using notation introduced in Section 2, for $x \in \mathbb{R}^d$, $y \in M$, $k \in \{1, \ldots, n-1\}$ and $n > 2$, set

$$\rho_{n,k,1}(x) := \rho_{n,k,1}(x, X_2, \ldots, X_n), \ \ \xi_{n,k,1}(x, y) := \xi_{n,k,1}((x, y), Z_2, \ldots, Z_n).$$

These random variables depend, respectively, on $x, X_2, \ldots, X_n$ and $(x, y), Z_2, \ldots, Z_n$. To simplify notation we omit the random arguments of these functions.

Now, we formulate two auxiliary results concerning conditional distributions of random variables playing an essential role in the asymptotic behavior analysis of the conditional entropy estimates. It turns out surprisingly that the mentioned conditional distributions are specified mixtures of certain binomial laws with explicitly indicated weight coefficients. The proofs of these results are provided in Appendix. We write P_η for distribution of a random vector (or variable) η.

Lemma 3.1. *For any $y \in M$, $x \in \mathbb{R}^d$, $r = 0, 1, \dots, k$ and $\mathsf{P}_{\rho_{n,k,1}(x)}$-almost all $t \in (0, \infty)$, the following relation holds:*

$$\begin{aligned}\mathsf{P}(\xi_{n,k,1}(x,y) = r|\rho_{n,k,1}(x) = t)&\\ = \binom{k-1}{r}\mathsf{P}(Y = y|\|X-x\| \le t)^r(1-\mathsf{P}(Y = y|\|X-x\| \le t)^{k-1-r}\alpha(x,y,t)&\\ +\binom{k-1}{r-1}\mathsf{P}(Y = y|\|X-x\| \le t)^{r-1}(1-\mathsf{P}(Y = y|\|X-x\| \le t)^{k-r}(1-\alpha(x,y,t))&\end{aligned}$$

where $\alpha(x,y,t) = \mathsf{P}(Y \neq y|\|X-x\| = t)$ and $\binom{N}{m} := 0$ for $m < 0$ and $m > N$ $(N \in \mathbb{N},\ m \in \mathbb{Z})$.

Remark 3.2. As usual, for random vectors $\eta : \Omega \to \mathbb{R}^q$, $\zeta : \Omega \to \mathbb{R}^s$, and for $B \in \mathcal{B}(\mathbb{R}^q)$, $x \in \mathbb{R}^s$, the notation $\mathsf{P}(\eta \in B|\zeta = x) = \varphi(x)$ means that one takes a Borel function $\varphi(x)$, $x \in \mathbb{R}^s$, such that $\mathsf{P}(\eta \in B|\zeta) = \varphi(\zeta)$. The function φ is defined uniquely P_ζ-almost sure (see, *e.g.*, [39], v.1, Chap. II, Sect. 5).

Let $z_j = (x_j, y_j)$ where $x_j \in \mathbb{R}^d$, $y_j \in M$, $j \in \{1, 2\}$. For $n > 2$, let us define the random variables

$$\rho_{n,k,j}(x_1, x_2) := \rho_{n,k,j}(x_1, x_2, X_3, \dots, X_n), \tag{3.1}$$

$$\xi_{n,k,j}(z_1, z_2) := \xi_{n,k,j}(z_1, z_2, Z_3, \dots, Z_n). \tag{3.2}$$

Again we omit the random arguments of these functions.

Lemma 3.3. *Let $z_j = (x_j, y_j)$ where $x_j \in \mathbb{R}^d$, $y_j \in M$, $j = 1, 2$, $x_1 \neq x_2$. Introduce a random vector $\zeta := (\rho_{n,k,1}(x_1, x_2), \rho_{n,k,2}(x_1, x_2))$. Then, for any $n > 2$, $k \in \{0, 1, \dots, [(n-2)/2]\}$, where $[\cdot]$ is the integer part of a number, any $r_1, r_2 \in \{0, \dots, k\}$ and P_ζ-almost all (t_1, t_2) such that $t_1 > 0$, $t_2 > 0$ and $t_1, t_2 < |x_1 - x_2|/2$ the following relation is valid:*

$$\mathsf{P}(\xi_{n,k,1}(z_1, z_2) = r_1, \xi_{n,k,2}(z_1, z_2) = r_2|\zeta = (t_1, t_2)) = \prod_{j=1}^{2}\mathsf{P}(\xi_{n,k,j}(z_1, z_2) = r_j|\zeta = (t_1, t_2)). \tag{3.3}$$

Moreover,

$$\begin{aligned}\mathsf{P}\left(\xi_{n,k,j}(z_1, z_2) = r_j|\zeta = (t_1, t_2)\right)&\\ = \binom{k-1}{r_j}p_j^{r_j}(1-p_j)^{k-1-r_j}\alpha(x_j, y_j, t_j) + \binom{k-1}{r_j-1}p_j^{r_j-1}(1-p_j)^{k-r_j}(1-\alpha(x_j, y_j, t_j))&\end{aligned} \tag{3.4}$$

where $p_j = \mathsf{P}(Y = y_j|X \in B(x_j, t_j))$, $j = 1, 2$, and $\alpha(x, y, t)$ is the same as in Lemma 3.1.

We will also employ the following elementary results.

Lemma 3.4. *Let W be a random variable having finite $\mathsf{E}W$, and V be a random vector with values in $\mathbb{R}^m$ such that $\mathsf{P}(V \in B) > 0$ where $B \in \mathcal{B}(\mathbb{R}^m)$. Then*

$$\mathsf{E}(W|V \in B) = \int_B \mathsf{E}(W|V = y)\,\widetilde{\mathsf{P}}_{V,B}(\mathrm{d}y),$$

where $\mathsf{E}(W|V \in B) := \frac{1}{\mathsf{P}(V \in B)}\mathsf{E}(W\mathbb{I}\{V \in B\})$ and $\widetilde{\mathsf{P}}_{V,B}(A) := \mathsf{P}(V \in A|V \in B)$, $A \in \mathcal{B}(\mathbb{R}^m)$.

Lemma 3.5. *Let ξ, η be some random variables and $\mathsf{E}|\xi| < \infty$. Assume that a random variable ζ takes values in a finite or countable set S. Then, for P_η-almost all t, one has*

$$\mathsf{E}(\xi|\eta = t) = \sum_{r\in S} \mathsf{E}(\xi|\zeta = r, \eta = t)\mathsf{P}(\zeta = r|\eta = t).$$

4. Proof of Theorem 2.3

Observations $Z_1, Z_2, \dots$ have identical distribution. Thus $\mathsf{E}\widehat{H}_{n,k} = -\mathsf{E}\log\left(\frac{\xi_{n,k,1}+1}{k}\right)$ and one has to prove that

$$-\mathsf{E}\log\left(\frac{\xi_{n,k,1}+1}{k}\right) \to H(Y|X), \quad n \to \infty. \tag{4.1}$$

Taking into account the independence of Z_1 and $Z_2, \dots, Z_n$, we get

$$\mathsf{E}\left(\log\left(\frac{\xi_{n,k,1}+1}{k}\right)\middle| X_1 = x, Y_1 = y\right) = \mathsf{E}\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right) := h_{n,k}(x,y).$$

Consequently,

$$\mathsf{E}\log\left(\frac{\xi_{n,k,1}+1}{k}\right) = \sum_{y\in M}\int_{\mathbb{R}^d} h_{n,k}(x,y)f(x,y)\mathrm{d}x. \tag{4.2}$$

To prove (4.1) we study the right-hand side of formula (4.2). For each $n \in \mathbb{N}$, we take the specified partition of $\mathbb{R}^d$ into subsets $B_{1,n}$ and $B_{2,n}$ depending on auxiliary parameters $\theta, \nu > 0$. Then we consider integrals over these sets and show that the integrals over $B_{1,n}$ determine, for each $y \in M$, the asymptotic behavior (as $n \to \infty$) of the integrals appearing in (4.2). We use simultaneously two averaging procedures. The first one is given by (4.2). The second one means that in (4.2) instead of $h_{n,k}(x,y)$ we write $\mathsf{E}(\mathsf{E}(\log(\frac{\xi_{n,k,1}(x,y)+1}{k})|n^\beta\rho_{n,k,1}(x))$ where $\beta > 0$ is a parameter. Then we analyze the combination of integrals with the help of Lemma 3.1 providing the formula for conditional law of $\xi_{n,k,1}(x,y)$ given $\rho_{n,k,1}(x)$. Here, for $\delta > 0$, we also use the partition of the values taken by $n^\beta\rho_{n,k,1}(x)$ into sets $(0,\delta]$, (δ,∞) and evaluate the contribution of each integral to their sum. The appropriate choice of parameters θ, ν, δ and β leads to the desired result. To simplify the proof we divide it into three steps.

Step 1. Introduce parameters $\theta, \nu > 0$. In the sequel we will make an appropriate choice of these parameters. Demonstrate that one can consider only such $x \in \mathbb{R}^d$ that $f_X(x) > n^{-\theta}$ and, for each $y \in M$, $f(y|x) > n^{-\nu}$.

Due to (2.7), for $n > 2$ and $y \in M$, we come to relations

$$\int\limits_{\{x: f_X(x)\le n^{-\theta}\}} f_X(x)\,\mathrm{d}x \le \int\limits_{\{x: f_X(x)\le n^{-\theta}\}} \frac{|\log f_X(x)|^{1+\varepsilon}}{|\log n^{-\theta}|^{1+\varepsilon}} f_X(x)\,\mathrm{d}x \le \frac{1}{(\theta\log n)^{1+\varepsilon}}\mathsf{E}|\log f_X(X)|^{1+\varepsilon}, \tag{4.3}$$

$$\int\limits_{\{x: f(y|x)\le n^{-\nu}\}} f(x,y)\,\mathrm{d}x = \int\limits_{\{x: f(y|x)\le n^{-\nu}\}} f(y|x)f_X(x)\,\mathrm{d}x \le n^{-\nu}\int\limits_{\mathbb{R}^d} f_X(x)\,\mathrm{d}x = n^{-\nu}. \tag{4.4}$$

For $n \in \mathbb{N}$, take $\theta_n := n^{-\theta}$, $\nu_n := n^{-\nu}$ and consider the sets

$$B_{1,n} := \bigcap_{y\in M}\{x \in \mathbb{R}^d\colon f(y|x) > \nu_n\} \cap \{x \in \mathbb{R}^d\colon f_X(x) > \theta_n\}, \quad B_{2,n} := \mathbb{R}^d \setminus B_{1,n}. \tag{4.5}$$

One can write $\mathsf{E}\log\left(\frac{\xi_{n,k,1}+1}{k}\right) = I_1(n,k) + I_2(n,k)$ where

$$I_j(n,k) := \sum_{y\in M}\int\limits_{B_{j,n}} h_{n,k}(x,y)f(x,y)\mathrm{d}x, \quad j=1,2.$$

For $k>1$, all $x\in\mathbb{R}^d$ and $y\in M$, the inequality $|h_{n,k}(x,y)|\le\log k$ is valid because $\xi_{n,k,1}(x,y)$ takes values $0,1,\dots,k$. Thus

$$|I_2(n,k)| \le \log k\left(\int\limits_{\{x:f_X(x)\le\theta_n\}} f_X(x)\,\mathrm{d}x + \sum_{y\in M}\int\limits_{\{x:f(y|x)\le\nu_n\}} f(x,y)\,\mathrm{d}x\right). \tag{4.6}$$

According to (4.3) and (4.4) we infer that $I_2(n,k)\to 0$, $n\to\infty$, since $k\propto n^\alpha$.

Fix parameter $\beta>0$ and note that

$$\mathsf{E}\left(\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right)\middle| n^\beta\rho_{n,k,1}(x)\right) = \sum_{r=0}^{k}\log\left(\frac{r+1}{k}\right)\mathsf{P}(\xi_{n,k,1}(x,y)=r|n^\beta\rho_{n,k,1}(x)).$$

Hence,

$$h_{n,k}(x,y) = \sum_{r=0}^{k}\int_{(0,\infty)}\log\left(\frac{r+1}{k}\right)\mathsf{P}(\xi_{n,k,1}(x,y)=r|n^\beta\rho_{n,k,1}(x)=u)f^{(k)}_{n,x,\beta}(u)\mathrm{d}u$$

where $f^{(k)}_{n,x,\beta}(\cdot)$ is a density of a positive random variable $n^\beta\rho_{n,k,1}(x)$, and a density of a random variable $\rho_{n,k,1}(x)$ is indicated in the proof of Lemma 3.1 (see Appendix, Eq. (A.6)), here du stands for the Lebesgue measure on $\mathbb{R}$.

Now we fix an arbitrary $\delta>0$ and write $I_1(n,k)=S_1(n,k)+S_2(n,k)$ where, for $V_1=(\delta,\infty)$, $V_2=(0,\delta]$ and $j=1,2$,

$$S_j(n,k):=\sum_{y\in M}\int_{B_{1,n}}\sum_{r=0}^{k}\int_{V_j}\log\left(\frac{r+1}{k}\right)\mathsf{P}(\xi_{n,k,1}(x,y)=r|\rho_{n,k,1}(x)=un^{-\beta})f^{(k)}_{n,x,\beta}(u)\mathrm{d}uf(x,y)\mathrm{d}x.$$

The rest of the proof is divided into two parts.

Step 2. Let us show that $S_1(n,k)\to 0$ as $n\to\infty$ when parameters β and δ are taken appropriately. We find an upper bound for $|S_1(n,k)|$. For all $n\in\mathbb{N}$, $1\le k\le n$ and $x,y\in\mathbb{R}^d$, the variable $\xi_{n,k,1}(x,y)$ takes values $0,\dots,k$. Therefore, for $k>1$, one has

$$\sum_{r=0}^{k}\left|\log\left(\frac{r+1}{k}\right)\right|\mathsf{P}(\xi_{n,k,1}(x,y)=r|\rho_{n,k,1}(x)=un^{-\beta})\le\log k.$$

Thus

$$\begin{aligned}|S_1(n,k)| &\le \log k\sum_{y=1}^{m}\int\limits_{B_{1,n}}\int_{(\delta,\infty)} f^{(k)}_{n,x,\beta}(u)f(x,y)\,\mathrm{d}u\,\mathrm{d}x\\ &= \log k\int\limits_{B_{1,n}}\mathsf{P}\left(\rho_{n,k,1}(x)>\delta n^{-\beta}\right)f_X(x)\,\mathrm{d}x = \log k\int\limits_{B_{1,n}}\mathsf{P}(\eta_n(\beta,\delta,x)\le k-1)f_X(x)\,\mathrm{d}x,\end{aligned} \tag{4.7}$$

here $\eta_n(\beta,u,x) \sim Bin(n-1, p_n(\beta,u,x))$, *i.e.* $\eta_n(\beta,u,x)$ has a binomial law with parameters $n-1$ and $p_n(\beta,u,x)$ where

$$\begin{aligned} p_n(\beta,u,x) &:= \mathsf{P}(X \in U_n(\beta,u,x)) = \int_{U_n(\beta,u,x)} f_X(v)\,\mathrm{d}v, \\ U_n(\beta,u,x) &= \{v \in \mathbb{R}^d\colon \|v-x\| \le un^{-\beta}\}, \quad u>0, \quad x \in \mathbb{R}^d. \end{aligned}$$

Indeed, as $f^{(k)}_{n,x,\beta}(\cdot)$ is a density of a variable $\rho_{n,k}(x)n^{\beta}$, we can write

$$\int_{(\delta,\infty)} f^{(k)}_{n,x,\beta}(u)\,\mathrm{d}u = \mathsf{P}(\rho_{n,k,1}(x) > \delta n^{-\beta}). \tag{4.8}$$

The event $\{\omega : \rho_{n,k,1}(x) > \delta n^{-\beta}\}$ means that in a ball $B(x, \delta n^{-\beta})$ one can find no more than $k-1$ point among $\{X_i\}_{i=2}^n$. The independence of the observations yields $\mathsf{P}(\rho_{n,k,1}(x) > \delta n^{-\beta}) = \mathsf{P}(\eta_n(\beta,\delta,x) \le k-1)$. According to the inequality for binomial sums proved in [46], for any $n>1$, $k = 0,1,\dots,n-1$ and all considered values β, δ and x, the following bound holds

$$\begin{aligned} &\mathsf{P}(\eta_n(\beta,\delta,x) \le k-1) \\ &\qquad \le \Phi\left(\operatorname{sgn}\left(\frac{k}{n-1} - p_n(\beta,\delta,x)\right)\sqrt{2(n-1)h\left(\frac{k}{n-1}, p_n(\beta,\delta,x)\right)}\right) \end{aligned} \tag{4.9}$$

where $\Phi(\cdot)$ is the distribution function of a standard normal random variable,

$$\begin{aligned} \operatorname{sgn}(t) &= \begin{cases} 1, & t>0, \\ 0, & t=0, \\ -1, & t<0, \end{cases} \\ h(t,s) &= t\log\left(\frac{t}{s}\right) + (1-t)\log\left(\frac{1-t}{1-s}\right), \quad s,t \in (0,1). \end{aligned}$$

For each $y \in M$, the function $f(\cdot,y)$ is C_0-constricted, therefore, for μ-almost all $x \in \mathbb{R}^d$, each $u \in (0,\delta]$ and any n large enough,

$$\left|\frac{\mathsf{P}(X \in U_n(\beta,u,x), Y=y)}{|U_n(\beta,u,x)|} - f(x,y)\right| \le C_0 u n^{-\beta}. \tag{4.10}$$

Since $f_X(x) = \sum_{y \in M} f(x,y)$, for x, u and β under consideration, we get

$$\left|f_X(x) - \frac{p_n(\beta,u,x)}{|U_n(\beta,u,x)|}\right| \le \sharp M C_0 u n^{-\beta} = Cun^{-\beta} \tag{4.11}$$

where $C = \sharp M C_0$. This implies that, for arbitrary $\delta>0$, $\beta>0$, μ-almost all $x \in B_{1,n}$ and for any $n \ge N_0$, where $N_0 = N_0(\delta,\beta)$, the following inequality is satisfied

$$p_n(\beta,\delta,x) \ge V_d\delta^d n^{-d\beta}(f_X(x) - C\delta n^{-\beta})) \ge V_d\delta^d n^{-d\beta}(\theta_n - C\delta n^{-\beta}). \tag{4.12}$$

Recall that $V_d = |B(0,1)|$, $B(0,1) \subset \mathbb{R}^d$. Now we can obtain the upper bound for the argument of a function sgn in formula (4.9). In view of (4.12) one has

$$\frac{k}{n-1} - p_n(\beta,\delta,x) \le \frac{k}{n-1} - V_d \delta^d n^{-d\beta}(\theta_n - C\delta n^{-\beta}).$$

Take $\beta > \theta$. Then $n^{-\beta} = o(\theta_n)$ as $n \to \infty$. According to (2.6) we get $\frac{k}{n-1} \propto n^{\alpha-1}$, $n \to \infty$. Let parameters β and θ be such that

$$\alpha - 1 < -d\beta - \theta. \tag{4.13}$$

For (4.13) validity it is sufficient that $(d+1)\theta < 1-\alpha$, because we can choose $\beta > \theta$ arbitrary close to θ. Then $n^{\alpha-1} = o(n^{-d\beta-\theta})$, $n \to \infty$. Thus there exists $N \in \mathbb{N}$ (where $N = N(C,d,\alpha,\delta,\beta,\theta)$) such that if $n > N$, then $\frac{k}{n-1} - p_n(\beta,\delta,x) < 0$ for μ-almost all $x \in B_{1,n}$, which yields

$$\operatorname{sgn}\left(\frac{k}{n-1} - p_n(\beta,\delta,x)\right) = -1. \tag{4.14}$$

For $s,t \in (0,1)$ introduce the functions

$$L_{1,n}(t,s) = 2(n-1)(1-t)\log\left(\frac{1-t}{1-s}\right), \quad L_{2,n}(t,s) = 2(n-1)t\log\left(\frac{t}{s}\right).$$

Then $2(n-1)h(t,s) = L_{1,n}(t,s) + L_{2,n}(t,s)$. Now, for $t = \frac{k}{n-1}$ and $s = p_n(\beta,\delta,x)$ consider the behavior of the functions $L_{1,n}(t,s)$ and $L_{2,n}(t,s)$ (for μ-almost all $x \in B_{1,n}$) as $n \to \infty$. Applying (4.12), for $n \ge N_0$, we come to the bound

$$\begin{aligned} L_{1,n}(t,s) &= 2(n-1-k)\left(\log\left(1 - \frac{k}{n-1}\right) - \log(1 - p_n(\beta,\delta,x))\right) \\ &\ge 2(n-1-k)\left(\log\left(1 - \frac{k}{n-1}\right) - \log(1 - V_d\delta^d n^{-d\beta}(\theta_n - C\delta n^{-\beta}))\right) := \mathcal{L}_{1,n}(k). \end{aligned} \tag{4.15}$$

Evidently, $\mathcal{L}_{1,n}(k)$ depends not only on n and k, but also on a collection of parameters appearing in (4.15). Note that $\log(1+z) = z + o(z)$ as $z \to 0$. Hence, in view of (4.13) and since $\beta > \theta$ we get $\mathcal{L}_{1,n}(k) \propto n^{1-d\beta-\theta}$, as $n \to \infty$. For the same t, s and $x \in B_{1,n}$, taking into account that $0 < p_n(\beta,\delta,x) \le 1$, we obtain

$$L_{2,n}(t,s) = 2k\left(\log\left(\frac{k}{n-1}\right) - \log p_n(\beta,\delta,x)\right) \ge 2k\log\left(\frac{k}{n-1}\right) := \mathcal{L}_{2,n}(k).$$

Therefore, $\mathcal{L}_{2,n}(k) \propto n^\alpha \log n$, $n \to \infty$. Thus according to (4.13) we conclude that, for all n large enough, and for μ-almost all $x \in B_{1,n}$,

$$2(n-1)H\left(\frac{k}{n-1}, p_n(\beta,\delta,x)\right) \ge \mathcal{L}_{1,n}(k) + \mathcal{L}_{2,n}(k) := R_n \propto n^{1-d\beta-\theta} \tag{4.16}$$

where $k = k_n \propto n^\alpha$. Here R_n depends not only on n, but also on $\alpha, \beta,\ \delta,\ \theta$ and d. Therefore, (4.16) gives, for all n large enough, an estimate

$$\Phi\left(\operatorname{sgn}\left(\frac{k}{n-1} - p_n(\beta,\delta,x)\right)\sqrt{2(n-1)h\left(\frac{k}{n-1}, p_n(\beta,\delta,x)\right)}\right) \leq \Phi(-\sqrt{R_n}).$$

Since $\Phi(-z) \leq \frac{1}{\sqrt{2\pi}z}e^{-\frac{z^2}{2}}$ for $z > 1$, taking into account inequalities (4.7), (4.9) and (4.16) we get, for $k = k_n \propto n^\alpha$, $x \in B_{1,n}$ and all n large enough,

$$\int_{(\delta,\infty)} f^{(k)}_{n,x,\beta}(u)\,\mathrm{d}u \leq \frac{1}{\sqrt{2\pi R_n}}e^{-R_n/2}, \tag{4.17}$$

$$|S_1(n,k)| \leq \frac{\log k_n}{\sqrt{2\pi R_n}}e^{-R_n/2} \to 0,\quad n\to\infty.$$

Hence the proof of Step 2 is complete.

Step 3. We show that $S_2(n,k) \to -H(Y|X)$, $n \to \infty$. For $\mathsf{P}_{n^\beta \rho_{n,k,1}(x)}$-almost all u by virtue of Lemma 3.1

$$\begin{aligned}&\mathsf{E}\left(\log(\xi_{n,k,1}(x,y)+1)|\rho_{n,k,1}(x) = un^{-\beta}\right)\\ &\quad = \mathsf{P}(Y\neq y|\|X-x\| = un^{-\beta})\mathsf{E}\log(\mu_n+1) + \mathsf{P}(Y=y|\|X-x\| = un^{-\beta})\mathsf{E}\log(\mu_n+2)\end{aligned} \tag{4.18}$$

where the random variable μ_n does not depend on (X,Y) and

$$\mu_n = \mu_n(k,\beta,u,x,y) \sim Bin(k-1, P_n(\beta,u,x,y)), \tag{4.19}$$

$$P_n(\beta,u,x,y) := \mathsf{P}(Y=y|X\in U_n(\beta,u,x)). \tag{4.20}$$

Note that $\mathsf{P}(X \in B(x,t)) > 0$ for each $x \in \mathbb{R}^d$ and any $t > 0$ since $f_X(\cdot)$ is strictly positive μ-almost everywhere. At first we study, for $u \in (0,\delta]$, $y \in M$ and μ-almost all $x \in \mathbb{R}^d$ (such that $f_X(x) > 0$), the convergence rate of $P_n(\beta,u,x,y)$ to $f(y|x)$ as $n\to\infty$. One has

$$\begin{aligned}|P_n(\beta,u,x,y) - f(y|x)| = \frac{|U_n(\beta,u,x)|}{\mathsf{P}(X\in U_n(\beta,u,x))}&\left|\left\{\left(\frac{\mathsf{P}(X\in U_n(\beta,u,x), Y=y)}{|U_n(\beta,u,x)|} - f(x,y)\right)\right.\right.\\ &\left.\left. + f(y|x)\left(f_X(x) - \frac{\mathsf{P}(X\in U_n(\beta,u,x))}{|U_n(\beta,u,x)|}\right)\right\}\right|.\end{aligned}$$

For all $y \in M$ and μ-almost all $x \in \mathbb{R}^d$, the Lebesgue theorem on measures differentiation (see, *e.g.*, [45], Thm. 25.17) gives that

$$f(y|x) = \lim_{R\to 0+}\frac{\mathsf{P}(Y=y, X\in B(x,R))}{\mathsf{P}(X\in B(x,R))} \leq 1.$$

Hence, due to (4.10) and (4.11), for all n large enough, we obtain the inequality

$$|P_n(\beta,u,x,y) - f(y|x)| \leq C_0(\sharp M + 1)\delta n^{-\beta}\frac{|U_n(\beta,u,x)|}{\mathsf{P}(X\in U_n(\beta,u,x))}$$

for $u \in (0, \delta]$, μ-almost all $x \in \mathbb{R}^d$ and $y \in M$. In view of (4.11), for μ-almost all $x \in B_{1,n}$, $u \in (0, \delta]$ and $y \in M$,

$$\frac{\mathsf{P}(X \in U_n(\beta, u, x))}{|U_n(\beta, u, x)|} \geq f_X(x) - C_0(\sharp M)un^{-\delta} \geq \theta_n - C_0(\sharp M)\delta n^{-\beta} \geq \frac{1}{2}n^{-\theta}$$

if n is large enough ($n \geq N(\sharp M, C_0, \delta, \beta, \theta)$) and $\beta > \theta$. Thus, for such n and indicated u, x and y,

$$|P_n(\beta, u, x, y) - f(y|x)| \leq 2C_0\delta(\sharp M + 1)n^{-\beta+\theta}. \tag{4.21}$$

Note now that, for $k_n > 1$ and $P_n := P_n(\beta, u, x, y)$,

$$\mathsf{E}\log(\mu_n + 1) = \log((k_n - 1)P_n) + \mathsf{E}\log\left(\frac{\mu_n + 1}{(k_n - 1)P_n}\right).$$

Set $G_n := (0, \delta] \times B_{1,n} \times M$. We will demonstrate that

$$\sup_{(u,x,y)\in G_n}\left|\mathsf{E}\log\left(\frac{\mu_n + 1}{(k_n - 1)P_n}\right)\right| \to 0, \quad n \to \infty. \tag{4.22}$$

According to the Lyapunov inequality it is sufficient to prove that

$$\sup_{(u,x,y)\in G_n}\mathsf{E}\left(\log\left(\frac{\mu_n + 1}{(k_n - 1)P_n}\right)\right)^2 \to 0, \quad n \to \infty. \tag{4.23}$$

Introduce $\eta_n := \frac{\mu_n - (k_n - 1)P_n + 1}{(k_n - 1)P_n}$. Then

$$\begin{aligned}\mathsf{E}\left(\log\left(\frac{\mu_n + 1}{(k_n - 1)P_n}\right)\right)^2 &= \mathsf{E}\left(\log\left(1 + \eta_n\right)\right)^2 \\ &= \mathsf{E}\left((\log(1 + \eta_n))^2\mathbb{I}\left\{|\eta_n| < \frac{1}{2}\right\}\right) + \mathsf{E}\left((\log(1 + \eta_n))^2\mathbb{I}\left\{|\eta_n| \geq \frac{1}{2}\right\}\right) := T_1(n) + T_2(n)\end{aligned}$$

where $T_1(n) = T_1(n; k_n, u, x, y, \alpha)$, $T_2(n) = T_2(n; k_n, u, x, y, \alpha)$. For $k_n > 1$ and $(u, x, y) \in G_n$, $\frac{1}{(k_n-1)P_n} \leq \frac{\mu_n+1}{(k_n-1)P_n} \leq \frac{2}{P_n}$. Consequently,

$$\left|\log\left(\frac{\mu_n + 1}{(k_n - 1)P_n}\right)\right| \leq \max\left\{\left|\log\left(\frac{1}{(k_n - 1)P_n}\right)\right|, \left|\log\left(\frac{2}{P_n}\right)\right|\right\}.$$

Taking into account (4.21) and the bound $f(y|x) > n^{-\nu}$ for $(u, x, y) \in G_n$, we see that if $0 < \nu < \beta - \theta$ then $\frac{1}{2}n^{-\nu} \leq P_n \leq 1$ and $\frac{1}{4}n^{-\nu}k_n \leq (k_n - 1)P_n \leq k_n$ when n is large enough. Therefore, for all n large enough,

$$\left|\log\left(\frac{\mu_n + 1}{(k_n - 1)P_n}\right)\right| \leq b\log n$$

where $b = b(\alpha, \nu)$ does not depend on n. If $0 < \nu < \alpha$ then, for all n large enough,

$$T_2(n) \leq (b\log n)^2\mathsf{P}\left(|\eta_n| \geq \frac{1}{2}\right) \leq (b\log n)^2\mathsf{P}(|\mu_n - (k_n - 1)P_n| \geq \frac{1}{4}(k_n - 1)P_n).$$

For a random variable $H(m) \sim Bin(m,p)$, $p \in (0,1)$, $p = p_m$ and $\varepsilon = \varepsilon_m > 0$, the Hoeffding inequality (see, *e.g.*, [27], p. 22, and further generalizations there) yields

$$\mathsf{P}(|H(m) - mp| \geq m\varepsilon) \leq 2\exp\{-2m\varepsilon^2\}.$$

We employ this inequality for $m = (k_n - 1)$, $k_n \propto n^{\alpha}$, $p = P_n(\beta,u,x,y)$ and $\varepsilon_{k-1} = \frac{P_n}{4}$. Then

$$\sup_{(u,x,y)\in G_n} T_2(n) \leq 2(b\log n)^2 \exp\left\{-\frac{1}{8}(k_n-1)P_n^2\right\} \to 0, \quad n \to \infty,$$

whenever $\alpha - 2\nu > 0$ (we can take positive ν arbitrary small).

To get an upper bound for $T_1(n)$ we note that $|\log(1+z)| \leq 2|z|$ for $|z| < \frac{1}{2}$. Hence

$$T_1(n) = \mathsf{E}\left((\log(1+\eta_n))^2 \mathbb{I}\left\{|\eta_n| < \frac{1}{2}\right\}\right) \leq 4\mathsf{E}\eta_n^2.$$

It holds

$$\mathsf{E}\eta_n^2 = \frac{(k_n-1)P_n(1-P_n)+1}{((k_n-1)P_n)^2} \leq \frac{1}{(k_n-1)P_n} + \frac{1}{((k_n-1)P_n)^2}. \tag{4.24}$$

We have seen that, for $(u,x,y) \in G_n$ and all n large enough, the following inequality takes place $(k_n-1)P_n \geq \frac{1}{4}n^{-\nu}k_n \to \infty$ if $0 < \nu < \alpha/2$ (we also assume that $\nu < \beta - \theta$). Therefore, the right-hand side of (4.24) tends to zero as $n \to \infty$. Thus we have verified that $\sup_{(u,x,y)\in G_n} T_1(n) \to 0$, $n \to \infty$. In such a way (4.23) and (4.22) are proved. Hence

$$\sup_{(u,x,y)\in G_n} |\mathsf{E}\log(\mu_n+1) - \log((k_n-1)P_n(\beta,u,x,y))| \to 0, \quad n \to \infty.$$

Introduce notation $F_n := \frac{P_n(\beta,u,x,y)}{f(y|x)}$ where $f(y|x) > 0$. Then

$$\sup_{(u,x,y)\in G_n} |F_n - 1| = \sup_{(u,x,y)\in G_n} \frac{|P_n(\beta,u,x,y) - f(y|x)|}{f(y|x)} \leq cn^{-\beta+\theta+\nu} \to 0, \quad n \to \infty,$$

if $0 < \nu < \beta - \theta$ and c is defined by means of (4.21) and does not depend on n. We see that $\sup_{(u,x,y)\in G_n} |F_n - 1| < \frac{1}{2}$ for all n large enough. Then

$$|\log((k_n-1)P_n(\beta,u,x,y)) - \log((k_n-1)f(y|x))| = |\log(1+(F_n-1))| \leq 2|F_n - 1|.$$

So we come to the relation $\sup_{(u,x,y)\in G_n} |\log F_n| \to 0, \;\; n \to \infty$. Thus

$$\sup_{(u,x,y)\in G_n} |\mathsf{E}\log(\mu_n+1) - \log((k_n-1)f(y|x)| \to 0, \quad n \to \infty. \tag{4.25}$$

In a similar way we verify that $\sup_{(u,x,y)\in G_n} |\mathsf{E}\log(\mu_n+2) - \log((k_n-1)f(y|x)| \to 0, \;\; n \to \infty$. Taking into account (4.18) we ascertain that

$$\sup_{(u,x,y)\in G_n} \left|\mathsf{E}\left(\log(\xi_{n,k,1}(x,y)+1)|\rho_{n,k,1}(x) = un^{-\beta}\right) - \log((k-1)f(y|x)\right|$$

$$\leq \sup_{(u,x,y)\in G_n} |\mathsf{E}\log(\mu_n+1) - \log((k-1)f(y|x)| + \sup_{(u,x,y)\in G_n} |\mathsf{E}\log(\mu_n+2) - \log((k-1)f(y|x)| \to 0$$

as $n\to\infty$. Consequently,

$$\begin{aligned}\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}&\left(\mathsf{E}\left(\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right)\Big|\rho_{n,k,1}(x)=un^{-\beta}\right)-\log\left(\frac{k-1}{k}\right)-\log f(y|x)\right)\\ &\times f^{(k)}_{n,x,\beta}(u)\mathrm{d}uf(x,y)\mathrm{d}x\\ &=S_2(n,k)-\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\left(\log\left(\frac{k-1}{k}\right)+\log f(y|x)\right)f^{(k)}_{n,x,\beta}(u)\mathrm{d}uf(x,y)\mathrm{d}x\to 0\end{aligned}$$

as $n\to\infty$ (recall that $k=k_n$). It remains to show that

$$-\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\left(\log\left(\frac{k-1}{k}\right)+\log f(y|x)\right)f^{(k)}_{n,x,\beta}(u)\mathrm{d}uf(x,y)\mathrm{d}x\to H(Y|X),\quad n\to\infty.$$

Firstly, we can write

$$\begin{aligned}0&\leq-\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\log\left(\frac{k-1}{k}\right)f^{(k)}_{n,x,\beta}(u)\mathrm{d}uf(x,y)\mathrm{d}x\\ &=-\log\left(1-\frac{1}{k}\right)\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}f^{(k)}_{n,x,\beta}(u)\mathrm{d}uf(x,y)\mathrm{d}x\leq-\log\left(1-\frac{1}{k}\right)\to 0,\quad n\to\infty.\end{aligned}$$

Secondly,

$$\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\log f(y|x)f^{(k)}_{n,x,\beta}(u)\mathrm{d}uf(x,y)\mathrm{d}x=\sum_{y\in M}\int_{B_{1,n}}\log f(y|x)\int_{(0,\delta]}f^{(k)}_{n,x,\beta}(u)\mathrm{d}uf(x,y)\mathrm{d}x.$$

Inequality (4.17) yields

$$0\leq\Delta_n:=\sup_{x\in B_{1,n}}\left(1-\int_{(0,\delta]}f^{(k)}_{n,x,\beta}(u)\mathrm{d}u\right)\to 0,\quad n\to\infty.$$

Thus

$$\begin{aligned}&\left|\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\log f(y|x)f^{(k)}_{n,x,\beta}(u)\mathrm{d}uf(x,y)\mathrm{d}x-\sum_{y\in M}\int_{B_{1,n}}\log f(y|x)f(x,y)\mathrm{d}x\right|\\ &\quad\leq\Delta_n\sum_{y\in M}\int_{\mathbb{R}^d}|\log f(y|x)|f(x,y)\mathrm{d}x\to 0,\quad n\to\infty.\end{aligned}$$

Note now that

$$\sum_{y\in M}\int_{B_{1,n}}\log f(y|x)f(x,y)\mathrm{d}x\to\sum_{y\in M}\int_{\mathbb{R}^d}\log f(y|x)f(x,y)\mathrm{d}x,\quad n\to\infty,$$

since $B_{1,n} \nearrow \{(x,y) : f_X(x) > 0, \cap_{y\in M}\{f(y|x) > 0\}\}$ and $\int_B h(x)\mathrm{d}x = 0$ when $h \in L^1(\mathbb{R})$ and B is a Borel subset of $\mathbb{R}^d$ such that $\mu(B) = 0$. Consequently,

$$-\sum_{y\in M}\int_{B_{1,n}}\int_0^{\delta}\log f(y|x)f^{(k)}_{n,x,\beta}(u)\mathrm{d}uf(x,y)\mathrm{d}x \to H(Y|X), \quad n\to\infty.$$

To prove Theorem 2.3 we have imposed on parameters $\beta > 0$, $\theta > 0$, and $\nu > 0$ the following conditions: $\beta > \theta$, $\nu < \beta - \theta$, $(d+1)\theta < 1-\alpha$, $\alpha - 2\nu > 0$. For each given $\alpha \in (0,1)$ we can guarantee the validity of the indicated inequalities. Namely, one can pick $\beta \in (0, \frac{1-\alpha}{d+1})$ and then take $\theta \in (0,\beta)$. After that it remains to fix $\nu \in (0, \beta-\theta)$ so that $\nu < \frac{1}{2}\alpha$.

Thus the proof of Theorem 2.3 is complete. □

Remark 4.1. The careful analysis of the proof of Theorem 2.3 shows that we have established the following relation

$$\mathsf{E}\widehat{H}_{n,k} - H = O((\log n)^{-\varepsilon}), \quad n\to\infty, \tag{4.26}$$

where ε is the same as in (2.7). The authors of quite recent works [20, 22] study the asymptotic behavior of the risk of certain estimators of the Shannon differential entropy under specified conditions imposed on a density f (*e.g.*, f belongs to some Hölder's ball of functions $f : [0,1]^d \to \mathbb{R}_+$). We employ other hypothesis to study the conditional Shannon's entropy in a mixed model. For bounded density f satisfying smoothness and other conditions the power-type convergence rate of the bias of truncated Kozachenko–Leonenko is obtained in [18]. Thus it seems natural that the convergence rate of the $\widehat{H}_{n,k}$ bias to zero is rather slow in our setting. It will be interesting in the future research to study $\widehat{H}_{n,k}$ behavior when f belongs to various functional classes.

Remark 4.2. One can modify the proposed estimator replacing $\log k$ by $\psi(k)$, where ψ is the Digamma function. Namely, introduce $\widetilde{H}_{n,k}$ according to (2.1) with

$$\widetilde{H}_{n,k,i} = -\log(\xi_{n,k,i}(Z_1,\ldots,Z_n)+1) + \psi(k)$$

instead of $\widehat{H}_{n,k,i}$. Such correction was used, for an arbitrary fixed k, to improve the bias of the Kozchenko–Leonenko entropy estimator and of the Kraskov–Stögbauer–Grassberger mutual information estimator. In our case one has $k = k(n)$, where $k(n) \propto n^{\alpha}$, $n\to\infty$, $\alpha \in (0,1)$. Since (see, *e.g.*, [30], Thm. 3.7)

$$\psi(z) = \log z - \frac{1}{z} + O\left(\log\left(1+\frac{1}{z}\right)\right), \quad z\to\infty,$$

the proof of Theorem 2.3 leads to (4.26) if we replace $\widehat{H}_{n,k}$ by $\widetilde{H}_{n,k}$.

5. Proof of Theorem 2.4

The proof is divided into several steps. Since $Z_1, Z_2, \ldots$ are i.i.d. observations, one has

$$\begin{aligned}\mathsf{E}(\widehat{H}_{n,k} - H(Y|X))^2 = {}&\frac{1}{n}\mathsf{E}(\widehat{H}_{n,k,1} - H(Y|X))^2\\ &+\left(1-\frac{1}{n}\right)\mathsf{E}(\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X)).\end{aligned}$$

We will see that the expectations in the right-hand side of the latter formula are finite. Moreover, we will verify that, for $n\to\infty$,

(A) $\mathsf{E}(\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X)) = o(1)$,

(B) $\mathsf{E}(\widehat{H}_{n,k,1} - H(Y|X))^2 = o(n)$.

The proof of (A) is rather long as we will employ evaluation of a number of integrals depending on several auxiliary parameters. These parameters will be taken in appropriate way. In contrast to the proof of Theorem 2.3 we use now the conditional expectation given the vector $\zeta = (\rho_{n,k,1}(x_1, x_2), \rho_{n,k,2}(x_1, x_2))$, where $x_1, x_2 \in \mathbb{R}^d$. Here an essential role is played by Lemma 3.3 describing the conditional law of a vector $(\xi_{n,k,1}(z_1, z_2), \xi_{n,k,2}(z_1, z_2))$ given ζ, where $z_i = (x_i, y_i) \in \mathbb{R}^d \times M$, $i = 1, 2$. The proof of (B) is obtained as a byproduct.

Step 1. We show that (A) establishing can be reduced to verification of two statements. We need some notation. Consider $n > 2$, $k \in \{1, \dots, n\}$, $z_j = (x_j, y_j) \in \mathbb{R}^d \times M$, $j \in \{1, 2\}$. Set

$$\widehat{H}_{n,k,j}(z_1, z_2) := -\log\left(\frac{\xi_{n,k,j}(z_1, z_2, Z_3, \dots, Z_n) + 1}{k}\right).$$

The independence of observations $Z_1, \dots, Z_n$ implies that

$$\begin{aligned}&\mathsf{E}\left((\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X))|Z_1 = z_1, Z_2 = z_2\right)\\ &\qquad = \mathsf{E}(\widehat{H}_{n,k,1}(z_1, z_2) - H(Y|X))(\widehat{H}_{n,k,2}(z_1, z_2) - H(Y|X)) =: \mathcal{H}_{n,k}(z_1, z_2).\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathsf{E}(\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X))\\ &\qquad = \sum_{y_1,y_2=1}^{m} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathcal{H}_{n,k}(z_1, z_2) f(x_1, y_1) f(x_2, y_2)\, \mathrm{d}x_1 \mathrm{d}x_2.\end{aligned}$$

Due to the de la Vallée Poussin theorem (see, *e.g.*, [9], p. 10), for establishing (A) it suffices to prove validity of the following two statements.
1) If $dQ(x_1, x_2) := f(x_1, y_1) f(x_2, y_2)\, \mathrm{d}x_1 \mathrm{d}x_2$, then for each $y_1, y_2 \in M$ and Q-a.s. $(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\mathcal{H}_{n,k}(z_1, z_2) \to h(x_1, y_1) h(x_2, y_2), \quad n \to \infty, \tag{5.1}$$

here $h(x, y) := -\log f(y|x) - H(Y|X)$, $x \in \mathbb{R}^d$, $y \in M$, $z_j = (x_j, y_j), j \in \{1, 2\}$.
2) For some $a > 0$,

$$\sup_n \sum_{y_1,y_2 \in M} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\mathcal{H}_{n,k}(z_1, z_2)|^{1+a} f(x_1, y_1) f(x_2, y_2)\, \mathrm{d}x_1 \mathrm{d}x_2 < \infty. \tag{5.2}$$

Indeed, (5.1) and (5.2) imply that

$$\begin{aligned}&\sum_{y_1,y_2 \in M} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathcal{H}_{n,k}(z_1, z_2) f(x_1, y_1) f(x_2, y_2)\, \mathrm{d}x_1 \mathrm{d}x_2\\ &\qquad \to \sum_{y_1,y_2 \in M} \int_{\mathbb{R}^d} h(x_1, y_1) f(x_1, y_1)\, \mathrm{d}x_1 \int_{\mathbb{R}^d} h(x_2, y_2) f(x_2, y_2)\, \mathrm{d}x_2 = 0, \quad n \to \infty.\end{aligned}$$

Step 2. Let us prove (5.2). In view of the Jensen conditional inequality it is easily seen that (5.2) holds if

$$\sup_n \mathsf{E}\left|(\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X))\right|^{1+a} < \infty.$$

The Cauchy–Schwartz inequality yields

$$\mathsf{E}\left|(\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X))\right|^{1+a} \leq \mathsf{E}|\widehat{H}_{n,k,1} - H(Y|X)|^{2+2a}.$$

Thus, as $a > 0$ can be taken arbitrary small, (5.2) holds if, for some $\varepsilon > 0$,

$$\sup_n \mathsf{E}|\widehat{H}_{n,k,1} - H(Y|X)|^{2+\varepsilon} < \infty. \tag{5.3}$$

On applying the Lyapunov moment inequality we observe that (5.3) guarantees validity of (B).

Employing the reasoning used to prove Theorem 2.3 one arrives at an expression

$$\mathsf{E}\left|-\log\left(\frac{\xi_{n,k,1}(Z_1,\ldots,Z_n)+1}{k}\right) - H(Y|X)\right|^{2+\varepsilon} \\ = \sum_{y\in M}\int_{\mathbb{R}^d} \mathsf{E}\left|\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon} f(x,y)\,\mathrm{d}x. \tag{5.4}$$

Fix an arbitrary $\beta > 0$ and, for $x \in \mathbb{R}^d$, $y \in M$ and $u > 0$, set

$$I_{n,k,\beta,\varepsilon}(x,y,u) := \mathsf{E}\left(\left|\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon}\middle| n^\beta \rho_{n,k,1} = u\right).$$

Then we can write

$$\mathsf{E}\left|\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon} = \int_0^\infty I_{n,k,\beta,\varepsilon}(x,y,u) f^{(k)}_{n,x,\beta}(u)\,\mathrm{d}u$$

where $f^{(k)}_{n,x,\beta}(\cdot)$ is a density of random variable $n^\beta \rho_{n,k,1}$ (see Appendix, Eq. (A.6)). Taking into account the finiteness of the set M it is sufficient to prove that, for each $y \in M$, one has $\sup_n |\mathcal{I}_{n,k,i}(y)| < \infty$ where

$$\mathcal{I}_{n,k,i}(y) := \int_{B_{i,n}} \int_0^\infty I_{n,k,\beta,\varepsilon}(x,y,u) f^{(k)}_{n,x,\beta}(u) f(x,y)\,\mathrm{d}u\mathrm{d}x, \quad i = 1, 2,$$

and $B_{i,n}$ were defined in (4.5).

Step 3. Now we evaluate the integral $\mathcal{I}_{n,k,2}(y)$. Clearly, if $k > 1$ then, for any $x \in \mathbb{R}^d$ and $y \in M$, the following is P-a.s. true

$$\left|\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right) + H(Y|X)\right| \leq \log k + |H(Y|X)|. \tag{5.5}$$

Hence, using (4.3) and (4.4) with $2+\varepsilon$ instead of $1+\varepsilon$, we come to the relation $\sup_n |\mathcal{I}_{n,k,2}| < \infty$.

Step 4. We write $\mathcal{I}_{n,k,1}(y)$ as a sum of two summands and consider the first one. Fix $\delta > 0$. One has $\mathcal{I}_{n,k,1}(y) = S_{n,k,1}(y) + S_{n,k,2}(y)$ where, for $V_1 = (\delta, \infty)$ and $V_2 = (0, \delta]$,

$$S_{n,k,j}(y) = \int\limits_{B_{1,n}} \int_{V_j} I_{n,k,\beta,\varepsilon}(x,y,u) f^{(k)}_{n,x,\beta}(u) f(x,y)\, \mathrm{d}u \mathrm{d}x, \quad j = 1,2.$$

In similarity to (4.7), for each $y \in M$ and all n large enough, basing on (5.5) we obtain that

$$\begin{aligned} 0 \le S_{n,k,1}(y) &\le (\log k + |H(Y|X)|)^{2+\varepsilon} \int\limits_{B_{1,n}} \int_{\delta}^{\infty} f^{(k)}_{n,x,\beta}(u) f(x,y)\, \mathrm{d}u\, \mathrm{d}x \\ &\le (\log k + |H(Y|X)|)^{2+\varepsilon} \int\limits_{B_{1,n}} \mathsf{P}(\eta_n(\delta,\beta,x) \le k-1) f_X(x)\, \mathrm{d}x. \end{aligned}$$

According to inequality (4.17) one has

$$0 \le S_{n,k,1}(y) \le \frac{(\log k + |H(Y|X)|)^{2+\varepsilon}}{\sqrt{2\pi R_n}} e^{-R_n/2} \to 0, \quad n \to \infty, \tag{5.6}$$

where R_n was introduced in (4.16).

Step 5. Now we turn to the estimation of $S_{n,k,2}(y)$. Lemma 3.1 yields that

$$\begin{aligned} I_{n,k,\beta,\varepsilon}(x,y,u) = \mathsf{E}\left|\log\left(\frac{\mu_n+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon} \mathsf{P}(Y \neq y | \|X-x\| = un^{-\beta}) \\ + \mathsf{E}\left|\log\left(\frac{\mu_n+2}{k}\right) + H(Y|X)\right|^{2+\varepsilon} \mathsf{P}(Y = y | \|X-x\| = un^{-\beta}), \end{aligned}$$

here the variable μ_n is defined in (4.19) and does not depend on (X,Y). We show that

$$\sup_{(u,x,y)\in G_n} \left| I_{n,k,\beta,\varepsilon}(x,y,u) - |\log f(y|x) + H(Y|X)|^{2+\varepsilon} \right| \to 0, \quad n \to \infty, \tag{5.7}$$

where $G_n = (0,\delta] \times B_{1,n} \times M$. It is sufficient to prove that, as $n \to \infty$,

$$\sup_{(u,x,y)\in G_n} \left| \mathsf{E}\left|\log\left(\frac{\mu_n+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon} - |\log f(y|x) + H(Y|X)|^{2+\varepsilon} \right| \to 0, \tag{5.8}$$

$$\sup_{(u,x,y)\in G_n} \left| \mathsf{E}\left|\log\left(\frac{\mu_n+2}{k}\right) + H(Y|X)\right|^{2+\varepsilon} - |\log f(y|x) + H(Y|X)|^{2+\varepsilon} \right| \to 0. \tag{5.9}$$

For $\nu \in (0,1/2)$ and $\gamma \in (0, 1/2-\nu)$ one has $\mathsf{E}\left|\log(\frac{\mu_n+1}{k}) + H(Y|X)\right|^{2+\varepsilon} = T_{n,1} + T_{n,2}$ where

$$T_{n,1} := \mathsf{E}\left(\left|\log\left(\frac{\mu_n+1}{k}\right)+H(Y|X)\right|^{2+\varepsilon}\mathbb{I}\{|\mu_n-\mathsf{E}\mu_n|\le(k-1)^{1/2+\gamma}\}\right),$$
$$T_{n,2} := \mathsf{E}\left(\left|\log\left(\frac{\mu_n+1}{k}\right)+H(Y|X)\right|^{2+\varepsilon}\mathbb{I}\{|\mu_n-\mathsf{E}\mu_n|>(k-1)^{1/2+\gamma}\}\right),$$

here $T_{n,j}=T_{n,j}(\beta,\varepsilon,\gamma,u,x,y)$, $j=1,2$. Note that by elementary properties of the binomial distribution $\mathsf{E}\mu_n=(k-1)P_n(\beta,u,x,y)$ with $P_n(\beta,u,x,y)$ introduced in (4.20).

In view of (5.5), for $(u,x,y)\in G_n$ and all n large enough, one can write

$$0\le\left|\log\left(\frac{\mu_n+1}{k}\right)+H(Y|X)\right|^{2+\varepsilon}\le(\log k+|H(Y|X)|)^{2+\varepsilon}.$$

Applying the Hoeffding inequality with $m=k-1$, $H(m)=\mu_n$, $p=P_n$, $\varepsilon=(k-1)^{-1/2+\gamma}$ we have

$$0\le T_{n,2}\le(\log k+|H(Y|X)|)^{2+\varepsilon}\mathsf{P}(|\mu_n-\mathsf{E}\mu_n|>(k-1)^{1/2+\gamma})$$
$$\le(\log k+|H(Y|X)|)^{2+\varepsilon}2e^{-2(k-1)^{2\gamma}}\to0,\quad n\to\infty. \tag{5.10}$$

Note also that

$$|\log f(y|x)+H(Y|X)|^{2+\varepsilon}\mathsf{P}(|\mu_n-\mathsf{E}\mu_n|>(k-1)^{1/2+\gamma})\to0,\quad n\to\infty. \tag{5.11}$$

Set $B_i(n,p)=\mathsf{P}(\mu_n=i)$, $i=0,\dots,k-1$, and write $T_{n,1}$ in the following way

$$T_{n,1}=\sum_{i:|i-\mathsf{E}\mu_n|\le(k-1)^{1/2+\gamma}}\left|\log\left(\frac{i+1}{k}\right)+H(Y|X)\right|^{2+\varepsilon}B_i(n,p). \tag{5.12}$$

To get the upper bound for $|T_{n,1}-|\log f(y|x)+H(Y|X)|^{2+\varepsilon}|$ we employ the Lagrange formula for a function $g(z)=|\log z+a|^{2+\varepsilon}$, $z>0$, $a\in\mathbb{R}$. For $z,z_0>0$, one has

$$g(z)-g(z_0)=g'(\xi)(z-z_0),\quad \xi=z_0+\lambda(z-z_0),\quad \lambda=\lambda(z,z_0)\in(0,1).$$

Take $z=\frac{i+1}{k}$, $z_0=f(y|x)$ and $a=H(Y|X)$. Then, for i belonging to the summation set in (5.12), in view of (4.21) we get

$$|z-z_0|=\left|\frac{i+1}{k}-f(y|x)\right|\le\left|\frac{i-(k-1)P_n}{k}\right|+\frac{1}{k}+\left|\frac{(k-1)P_n}{k}-P_n\right|+|P_n-f(y|x)|$$
$$\le(k-1)^{-1/2+\gamma}+2k^{-1}+2C\delta(\sharp M+1)n^{-\beta+\theta}:=\mathcal{Z}_n\propto n^{-\beta+\theta},$$

provided that $\beta-\theta<1/2-\gamma$. Note that $g'(z)=(2+\varepsilon)z^{-1}(\log z+a)|\log z+a|^{\varepsilon}$, $z>0$. Hence $|g'(\xi)|\le(2+\varepsilon)|\xi|^{-1}|\log\xi+a|^{1+\varepsilon}$. For all n large enough,

$$|\log\xi|\le\max\{|\log z_0|,|\log z|\}\le\max\{|\log\nu_n|,\log k\}\le c\log n$$

where $c = c(\alpha, \nu)$. Clearly, $|\xi| \ge \min\{z, z_0\}$. In our case $z_0 \ge n^{-\nu}$ and $z \ge P_n(k_n - 1) - (k_n - 1)^{\frac{1}{2}+\gamma}$. According to (4.21) we can write $P_n \ge \frac{1}{2}n^{-\nu}$ for all n large enough if $\nu < \beta - \theta$. Therefore, for all n large enough, one can see that $z \ge c_1 n^{-\nu+\alpha}$ if $\alpha - \nu > \frac{1}{2} + \gamma$ (here $c_1 = c_1(\alpha)$). The latter inequality holds if $\nu < 2\alpha$ (then we take positive γ which is small enough). Thus $|\xi|^{-1} \le n^{\nu}$ for all n large enough. Consequently, uniformly in i belonging to the summation set in definition of $T_{n,1}$,

$$|g(z) - g(z_0)| = |g'(\xi)||z - z_0| \le (2+\varepsilon)(\log n)^{1+\varepsilon} n^{\nu} \mathcal{Z}_n \propto (\log n)^{1+\varepsilon} n^{\nu-\beta+\theta} \to 0, \quad n \to \infty,$$

whenever $\nu < \beta - \theta$. Taking into account that $\sum_{i:|i-\mathsf{E}\mu_n| \le (k-1)^{1/2+\gamma}} B_i(n,p) \le 1$ we come to relation

$$\sup_{(u,x,y)\in G_n} |T_{n,1} - |\log f(y|x) + H(Y|X)|^{2+\varepsilon}| \to 0, \quad n \to \infty.$$

This formula, (5.10) and (5.11) yield (5.8). Relation (5.9) is proved analogously. In such a way we establish (5.7). Hence,

$$S_{n,k,2}(y) - \int_{B_{1,n}} \int_0^{\delta} |\log f(y|x) + H(Y|X)|^{2+\varepsilon} f^{(k)}_{n,x,\beta}(u) f(x,y)\, \mathrm{d}u \mathrm{d}x \to 0, \quad n \to \infty.$$

Obviously, for each $n \in \mathbb{N}$ and any $y \in M$, we get

$$\int_{B_{1,n}} \int_0^{\delta} |\log f(y|x) + H(Y|X)|^{2+\varepsilon} f^{(k)}_{n,x,\beta}(u) f(x,y)\, \mathrm{d}u \mathrm{d}x \\ \le \int_{\mathbb{R}^d \cap \{x: f(y|x)>0\}} |\log f(y|x) + H(Y|X)|^{2+\varepsilon} f(x,y)\, \mathrm{d}x < \infty.$$

Indeed, in view of the Minkowski inequality and since $f(y|x) = \mathsf{P}(Y = y|X = x) \le 1$ for each $y \in M$ and P_X-almost all x, it is enough to show that

$$\int_{\mathbb{R}^d \cap \{x: f(y|x)>0\}} \left(\log \frac{1}{f(y|x)}\right)^{2+\varepsilon} f(x,y)\, \mathrm{d}x < \infty, \quad y \in M.$$

For any $\varepsilon > 0$, there exists $\mathcal{T} = \mathcal{T}(\varepsilon) > 1$ such that $(\log t)^{2+\varepsilon} \le t$ whenever $t > \mathcal{T}$. Hence, for each $y \in M$, the latter integral can be written as follows

$$\int_{\{x: f(y|x) \ge 1/\mathcal{T}\}} \left(\log \frac{1}{f(y|x)}\right)^{2+\varepsilon} f(x,y)\, \mathrm{d}x + \int_{\{x: 0 < f(y|x) < 1/\mathcal{T}\}} \left(\log \frac{1}{f(y|x)}\right)^{2+\varepsilon} f(y|x) f_X(x)\, \mathrm{d}x \\ \le (\log \mathcal{T})^{2+\varepsilon} \mathsf{P}(Y = y) + \int_{\{x: f(y|x) < 1/T\}} (f(y|x))^{-1} f(y|x) f_X(x)\, \mathrm{d}x \le (\log \mathcal{T})^{2+\varepsilon} + 1 < \infty.$$

Consequently, for $y \in M$, one has $\sup_n S_{n,k,2}(y) < \infty$. Thus by applying (5.6) we ensure, for each $y \in M$, that $\sup_n |\mathcal{I}_{n,k,1}(y)| < \infty$. Relation (5.3) is established.

Step 6. Now we concentrate on the proof of (5.1). For $z_1, z_2 \in \mathbb{R}^d \times M$, $n > 2$ and $k = k_n$, introduce

$$R_{n,k}(z_1, z_2) := (\widehat{H}_{n,k,1}(z_1, z_2) - H(Y|X))(\widehat{H}_{n,k,2}(z_1, z_2) - H(Y|X)).$$

Write $\rho_{n,k,j} := \rho_{n,k,j}(x_1,x_2)$, $j=1,2$. Fix $\delta>0$ and set

$$\mathcal{U}_{n,k,1}(z_1,z_2) := \mathsf{E}(R_{n,k}(z_1,z_2)|\rho_{n,k,1}n^{\beta}\le\delta,\rho_{n,k,2}n^{\beta}\le\delta)\mathsf{P}(\rho_{n,k,1}n^{\beta}\le\delta,\rho_{n,k,2}n^{\beta}\le\delta),$$
$$\mathcal{U}_{n,k,2}(z_1,z_2) := \mathcal{H}_{n,k}(z_1,z_2)-\mathcal{U}_{n,k,1}(z_1,z_2)$$

where, for an integrable random variable ξ, one has $\mathsf{E}(\xi|A):=0$ if $\mathsf{P}(A)=0$. For all $n\ge N_1$, where $N_1=N_1(H(Y|X))$, and any $z_j\in\mathbb{R}^d\times M$, $j=1,2$, the inequality $|R_{n,k}(z_1,z_2)|\le 2(\log k)^2$ holds with probability one. Therefore

$$|\mathcal{U}_{n,k,2}(z_1,z_2)|\le 2(\log k)^2\mathsf{P}\left(\left\{\rho_{n,k,1}>\delta n^{-\beta}\right\}\cup\left\{\rho_{n,k,2}>\delta n^{-\beta}\right\}\right)\le 4(\log k)^2\mathsf{P}\left(\rho_{n,k,1}>\delta n^{-\beta}\right).$$

Due to (4.8) and (4.17) we know, for each $z_j\in\mathbb{R}^d\times M$, $j=1,2$, that $|\mathcal{U}_{n,k,2}(z_1,z_2)|\to 0$, $n\to\infty$.

Step 7. It remains to show that, for any $y_1,y_2\in M$ and Q-almost all $(x_1,x_2)\in\mathbb{R}^d\times\mathbb{R}^d$,

$$\mathcal{U}_{n,k,1}(z_1,z_2)\to(-\log f(y_1|x_1)-H(Y|X))(-\log f(y_2|x_2)-H(Y|X)),\quad n\to\infty. \tag{5.13}$$

In the proof of Lemma 3.3 (see the Appendix) it is shown that, for each $x_1,x_2\in\mathbb{R}^d$ $(x_1\ne x_2)$, a finite measure $G(B):=\mathsf{P}((\rho_{n,k,1},\rho_{n,k,2})\in B)$, defined for $B\in\mathcal{B}(\mathbb{R}^2)\cap D_{x_1,x_2}$, where $D_{x_1,x_2}=(0,|x_1-x_2|/3]\times(0,|x_1-x_2|/3]$ can be written in the following way

$$\mathsf{P}((\rho_{n,k,1},\rho_{n,k,2})\in B)=\int_B g_{n,k}(x_1,x_2,u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2.$$

Here $g_{n,k}(x_1,x_2,\cdot,\cdot)\colon D_{x_1,x_2}\to\mathbb{R}_+$ is a certain integrable function (w.r.t. the restriction to D_{x_1,x_2} of $\Lambda\otimes\Lambda$). Clearly, for $0<k<n-2$,

$$\mathsf{P}(\rho_{n,k,i}(x_1,x_2)=0)=\sum_{j=k}^{n-2}\binom{n-2}{j}\mathsf{P}(\|X-x_i\|=0)^j(1-\mathsf{P}(\|X-x_i\|=0))^{n-2-j}=0,\quad i=1,2.$$

since X has a density (and $k=k(n)\propto n^{\alpha}$, $\alpha\in(0,1)$).

Let us set $g_{n,k}(x_1,x_2,\cdot,\cdot)=0$ on $\overline{D}_{x_1,x_2}\setminus D_{x_1,x_2}$ where $\overline{D}_{x_1,x_2}$ stands for a closure of D_{x_1,x_2} in $\mathbb{R}^2$. Take $N(x_1,x_2,\beta,\delta)$ to ensure that $\delta n^{-\beta}\le|x_1-x_2|/6$ for all $n\ge N(x_1,x_2,\beta,\delta)$. Then $Q_{n,\beta,\delta}:=[0,\delta n^{-\beta}]\times[0,\delta n^{-\beta}]\subset\overline{D}_{x_1,x_2}$ and $B(x_1,u_1n^{-\beta})\cap B(x_2,u_1n^{-\beta})=\varnothing$ for $u_i\in[0,\delta]$, $i=1,2$. Thus, for any $B\in\mathcal{B}(\mathbb{R}^2)$ and $n\ge N(x_1,x_2,\beta,\delta)$,

$$\mathsf{P}(\zeta\in B|\zeta\in Q_{n,\delta,\beta})=\frac{\mathsf{P}(\zeta\in B\cap Q_{n,\delta,\beta})}{\mathsf{P}(\zeta\in Q_{n,\delta,\beta})}=\frac{1}{\mathsf{P}(\zeta\in Q_{n,\delta,\beta})}\int_{B\cap Q_{n,\delta,\beta}}g_{n,k}(x_1,x_2,u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2$$

since $B\cap Q_{n,\delta,\beta}\subset\overline{D}_{x_1,x_2}$. Note that $\mathsf{P}(\zeta\in Q_{n,\delta,\beta})>0$ for all n large enough as

$$\mathsf{P}(\zeta\in Q_{n,\delta,\beta})=\sum_{\substack{l_1\ge k,l_2\ge k,\\ l_1+l_2\le n-2}}\binom{n}{l_1}\binom{n-l_1}{l_2}p_{n,x_1,\delta,\beta}^{l_1}p_{n,x_2,\delta,\beta}^{l_2}(1-p_{n,x_1,\delta,\beta}-p_{n,x_2,\delta,\beta}))^{n-l_1-l_2}.$$

Here $p_{n,x_i,\delta,\beta}=\mathsf{P}(\|X-x_i\|\le\delta n^{-\beta})>0$ for all $n\in\mathbb{N}$ and $i=1,2$, as $\mathsf{P}_X(B(x,r))>0$ for any $x\in\mathbb{R}^d$ and $r>0$ (because $f_X(z)>0$ for μ-almost all $z\in\mathbb{R}^d$). We also take into account that $p_{n,x_i,\delta,\beta}\to 0$ as $n\to\infty$, $i=1,2$.

Therefore a function $\tilde{g}_{n,k}(x_1,x_2,\cdot,\cdot):\mathbb{R}^2\to\mathbb{R}_+$,

$$\tilde{g}_{n,k}(x_1,x_2,u_1,u_2)=\begin{cases}\frac{1}{\mathsf{P}(\zeta\in Q_{n,\delta,\beta})}g_{n,k}(x_1,x_2,u_1,u_2), & \text{if } (u_1,u_2)\in Q_{n,\delta,\beta},\\ 0, & \text{if } (u_1,u_2)\in\mathbb{R}^2\setminus Q_{n,\delta,\beta},\end{cases}$$

is a probability density of the measure $\mathsf{P}(\zeta\in\cdot|\zeta\in Q_{n,\delta,\beta})$ which is defined on $(\mathbb{R}^2,\mathcal{B}(\mathbb{R}^2))$. Thus the measure $\mathsf{P}(n^\beta\zeta\in\cdot|\zeta\in Q_{n,\delta,\beta}))$ on this space has a density (w.r.t. a restriction of $\Lambda\otimes\Lambda$ on $(\mathbb{R}^2,\mathcal{B}(\mathbb{R}^2))$) $f_{n,k,\beta,\delta}(x_1,x_2,\cdot,\cdot)\colon\mathbb{R}^2\to\mathbb{R}_+$.

Now we are going to employ Lemma 3.4 with

$$W:=R_{n,k}(z_1,z_2),\;\; V:=(n^\beta\rho_{n,k,1}(x_1,x_2),n^\beta\rho_{n,k,2}(x_1,x_2)),\;\; B:=\{(u_1,u_2)\in[0,\delta]\times[0,\delta]\}.$$

Note that $\{V\in B\}=\{\zeta\in Q_{n,\delta,\beta}\}$ and $\mathsf{E}W$ exists. Consequently, for considered x_1,x_2, β, δ and $n>N(x_1,x_2,\beta,\delta)$, the following formula is valid

$$\begin{aligned}\mathcal{U}_{n,k,1}(z_1,z_2)&=\mathsf{E}(W|V\in B)\,\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta)\\&=\int_0^\delta\!\!\int_0^\delta J_{n,k,\beta}(z_1,z_2,u_1,u_2)\,\mathrm{d}\mathsf{P}_{V,B}(u_1,u_2)\,\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta)\end{aligned}$$

where $J_{n,k,\beta}(z_1,z_2,u_1,u_2):=\mathsf{E}(R_{n,k}(z_1,z_2)|V=(u_1,u_2))$. It was shown that the measure $\mathsf{P}_{V,B}$ has a density $f_{n,k,\beta,\delta}(x_1,x_1,\cdot,\cdot)$, therefore

$$\mathcal{U}_{n,k,1}(z_1,z_2)=\int_0^\delta\!\!\int_0^\delta J_{n,k,\beta}(z_1,z_2,u_1,u_2)f_{n,k,\beta,\delta}(x_1,x_1,u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2\,\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta).$$

Step 8. Now we show that uniformly for $\mathsf{P}_{V,B}$-almost all $(u_1,u_2)\in(0,\delta]\times(0,\delta]$

$$|J_{n,k,\beta}(z_1,z_2,u_1,u_2)-(-\log f(y_1|x_1)-H(Y|X))(-\log f(y_2|x_2)-H(Y|X))|\to0,\;\; n\to\infty. \tag{5.14}$$

Set $\xi:=(\xi_{n,k,1},\xi_{n,k,2})$ where $\xi_{n,k,1}=\xi_{n,k,1}(z_1,z_2)$, $i=1,2$. Due to Lemma 3.5, for P_V-almost all (u_1,u_2),

$$J_{n,k,\beta}(z_1,z_2,u_1,u_2)=\sum_{r_1,r_2}\mathsf{E}(R_{n,k}(z_1,z_2)|V=(u_1,u_2),\xi=(r_1,r_2))\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2))$$

Note that, for $\mathsf{P}_{V,\xi}$-almost all (u_1,u_2,r_1,r_2),

$$\mathsf{E}(R_{n,k}(z_1,z_2)|V=(u_1,u_2),\xi=(r_1,r_2))\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2))$$

$$=\left(-\log\left(\frac{r_1+1}{k}\right)-H(Y|X)\right)\left(-\log\left(\frac{r_2+1}{k}\right)-H(Y|X)\right):=h(r_1,r_2) \tag{5.15}$$

because a random variable $R_{n,k}(z_1,z_2)$ is measurable w.r.t. σ-algebra $\sigma\{V,\xi\}$. A function $h(r_1,r_2)$ depends also on n,k,z_1,z_2.

Let $O\in\mathcal{B}(\mathbb{R}_+^2\times M^2)$ be the set consisting of (u_1,u_2,r_1,r_2) such that (5.15) holds. Then $\mathsf{P}_{V,\xi}(O)=1$. Since M^2 is a finite set, $O=\bigcup_{(r_1,r_2)\in M^2}O_{r_1,r_2}\times\{(r_1,r_2)\}$ where $O_{r_1,r_2}\in\mathcal{B}(\mathbb{R}_+^2)$. Note that at least one set O_{r_1,r_2} is not empty, otherwise $\mathsf{P}((V,\xi)\in O)\ne1$. Consequently, $\bigcup_{(r_1,r_2)\in M^2}O_{r_1,r_2}\ne\varnothing$. If $\bigcap_{(r_1,r_2)\in M^2}O_{r_1,r_2}\ne\varnothing$ then, for each $(u_1,u_2)\in\bigcap_{(r_1,r_2)\in M^2}O_{r_1,r_2}$,

$$J_{n,k,\beta}(z_1,z_2,u_1,u_2)=\sum_{r_1,r_2}h(r_1,r_2)\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2)). \tag{5.16}$$

Define the set $\widetilde{O}=(\bigcup_{(r_1,r_2)\in M}O_{r_1,r_2}\times M^2)\setminus O$. It also can be represented as

$$\widetilde{O}=\bigcup_{(r_1,r_2)\in M^2}\widetilde{O}_{r_1,r_2}\times\{(r_1,r_2)\},\ \ \widetilde{O}_{r_1,r_2}=\bigcup_{(s_1,s_2)\in M}O_{s_1,s_2}\setminus O_{r_1,r_2}.$$

Clearly, $\mathsf{P}_{V,\xi}(O)=1$ implies that $\mathsf{P}_V(\bigcup_{(r_1,r_2)\in M^2}O_{r_1,r_2})=1$. Indeed,

$$\mathsf{P}\left(V\in\bigcup_{(r_1,r_2)\in M}O_{r_1,r_2}\right)\geq\mathsf{P}\left(\bigcup_{(r_1,r_2)\in M}\{V\in O_{r_1,r_2},\xi=(r_1,r_2)\}\right)=\mathsf{P}((V,\xi)\in O)=1.$$

If the set $\widetilde{O}$ is empty then $O_{r_1,r_2}=O_{s_1,s_2}$ for $r_1\neq s_1,r_2\neq s_2$, so $\bigcap_{(r_1,r_2)\in M}O_{r_1,r_2}=\bigcup_{(r_1,r_2)\in M}O_{r_1,r_2}\neq\varnothing$, thus equality holds for P_V-almost all (u_1,u_2).

Let us consider the case where $\widetilde{O}\neq\varnothing$. Introduce $K:=\{(r_1,r_2)\in M^2:\widetilde{O}_{r_1,r_2}\neq\varnothing\}$. Hence $K\neq\varnothing$. If $L:=M^2\setminus K\neq\varnothing$ then, for each $(r_1,r_2)\in L$, one has $O_{r_1,r_2}=\bigcup_{(s_1,s_2)\in M}O_{s_1,s_2}$. We have seen that $\mathsf{P}_V(\cup_{(s_1,s_2)\in M^2}O_{s_1,s_2})=1$. Therefore, $\mathsf{P}_V(O_{r_1,r_2})=1$ for each $(r_1,r_2)\in L$. Introduce $K_1:=\{(r_1,r_2)\in K:\mathsf{P}_V(\widetilde{O}_{r_1,r_2})=0\}$ and $K_2:=K\setminus K_1$. If $(r_1,r_2)\in K_1$ then $\mathsf{P}_V(O_{r_1,r_2})=\mathsf{P}_V(\bigcup_{(s_1,s_2)\in M}O_{s_1,s_2}\setminus\widetilde{O}_{r_1,r_2})=\mathsf{P}_V(\bigcup_{(s_1,s_2)\in M}O_{s_1,s_2})-\mathsf{P}_V(\widetilde{O}_{r_1,r_2})=1$.

Now we will demonstrate that if $(r_1,r_2)\in K_2$ then $\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2))=0$ for all $(u_1,u_2)\in S_{r_1,r_2}\subset\widetilde{O}_{r_1,r_2}$, $S_{r_1,r_2}\in\mathcal{B}(\mathbb{R}_+^2)$ where $\mathsf{P}_V(S_{r_1,r_2})=\mathsf{P}_V(\widetilde{O}_{r_1,r_2})>0$. If the latter statement is true then (61) is valid since we come to the trivial relation $0=0$ and, consequently, we obtain the desired formula for any $(u_1,u_2)\in A:=(\cap_{(r_1,r_2)\in L\cup K_1}O_{r_1,r_2})\cap(\cap_{(r_1,r_2)\in K_2}(O_{r_1,r_2}\cup S_{r_1,r_2}))$ where $\mathsf{P}_V(A)=1$ because $\mathsf{P}_V(O_{r_1,r_2})=1$ for $(r_1,r_2)\in L\cup K_1$ and

$$\begin{aligned}\mathsf{P}_V(O_{r_1,r_2}\cup S_{r_1,r_2})&=\mathsf{P}_V(O_{r_1,r_2})+\mathsf{P}_V(S_{r_1,r_2})=\mathsf{P}_V(O_{r_1,r_2})+\mathsf{P}_V(\widetilde{O}_{r_1,r_2})\\&=\mathsf{P}_V(O_{r_1,r_2}\cup\widetilde{O}_{r_1,r_2})=\mathsf{P}_V(\cup_{(s_1,s_2)\in M}O_{s_1,s_2})=1.\end{aligned}$$

Here we take into account that $S_{r_1,r_2}\subset\widetilde{O}_{r_1,r_2}$ and $O_{r_1,r_2}\cap\widetilde{O}_{r_1,r_2}=\varnothing$.

For each $(r_1,r_2)\in K_2$

$$\mathsf{P}(V\in\widetilde{O}_{r_1,r_2},\xi=(r_1,r_2))=\int_{\widetilde{O}_{r_1,r_2}}\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2))\,\mathrm{d}\mathsf{P}_V(u_1,u_2)=0 \tag{5.17}$$

because $(\widetilde{O}_{r_1,r_2}\times\{r_1,r_2\})\cap O=\varnothing$ and $\mathsf{P}_{V,\xi}(O)=1$. Invoking that $\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2))\geq 0$ for P_V-almost all (u_1,u_2), from equation (5.17) we infer that $\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2))=0$ for some set $S_{r_1,r_2}\in\mathcal{B}(\mathbb{R}_+^2)$ such that $S_{r_1,r_2}\subset\widetilde{O}_{r_1,r_2}$ and $\mathsf{P}_V(S_{r_1,r_2})=\mathsf{P}_V(\widetilde{O}_{r_1,r_2})$. Accordingly, (5.16) holds for P_V-almost all (u_1,u_2).

For $n>N(x_1,x_2)$ and $u_i\in[0,\delta]$ one has $u_in^{-\beta}<|x_1-x_2|/2$, $i=1,2$, so Lemma 3.3 applies to $\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2))$. Then

$$J_{n,k,\beta}(z_1,z_2,u_1,u_2)=\sum_{r_1,r_2}h(r_1,r_2)\,\mathsf{P}(\xi_{n,k,1}=r_1|V=(u_1,u_2))\mathsf{P}(\xi_{n,k,2}=r_2|V=(u_1,u_2))=\mathcal{J}_{n,1}\mathcal{J}_{n,2}$$

where

$$\begin{aligned}
\mathcal{J}_{n,i} &:= \mathsf{E}\left(-\log\left(\frac{\xi_{n,k,i}+1}{k}\right) - H(Y|X)\Big| V=(u_1,u_2)\right)\\
&= \mathsf{E}\left(-\log\left(\frac{\mu_{n,i}+1}{k}\right) - H(Y|X)\right)\mathsf{P}(Y\neq y\big|\|X-x_i\| = u_i n^{-\beta})\\
&\quad + \mathsf{E}\left(-\log\left(\frac{\mu_{n,i}+2}{k}\right) - H(Y|X)\right)\mathsf{P}\left(Y = y\big|\|X-x_i\| = u_i n^{-\beta}\right)\\
&= \mathsf{E}\left(-\log\left(\frac{\mu_{n,i}+1}{k}\right) + \log f(y_i|x_i)\right)\mathsf{P}\left(Y\neq y\big|\|X-x_i\| = u_i n^{-\beta}\right)\\
&\quad + \mathsf{E}\left(-\log\left(\frac{\mu_{n,i}+2}{k}\right) + \log f(y_i|x_i)\right)\mathsf{P}\left(Y = y\big|\|X-x_i\| = u_i n^{-\beta}\right)\\
&\quad + (-\log f(y_i|x_i) - H(Y|X))
\end{aligned}$$

$\mu_{n,i} = \mu_n(k,\beta,u_i,x_i,y_i)$, $J_{n,i} = J_{n,i}(k,\beta,u_i,x_i,y_i)$ and $i=1,2$.

According to (4.25) and since $\log k - \log(k-1) \to 0$ as $n\to\infty$ we can write

$$\sup_{(u,x,y)\in G_n}\left|\mathsf{E}\left(\log\left(\frac{\mu_n(k,\beta,u,x,y)+1}{k}\right)\right) - \log f(y|x)\right| \to 0.$$

Note that $G_n \nearrow (0,\delta]\times\{x\in\mathbb{R}^d\colon f_X(x)>0\}\times M$ as $n\to\infty$. For a given version of f_X and any $x\in\mathbb{R}^d$ such that $f_X(x)>0$, one can find $N(x)\in\mathbb{N}$ to guarantee relation $x\in B_{1,n}$ when $n>N(x)$. Consider $x_i\in\mathbb{R}^d$ such that $f_X(x_i)>0$, $i=1,2$. Then, for $n\geq \max\{N(x_1),N(x_2)\}$ and $i=1,2$

$$\sup_{0<u_i\leq\delta}\left|\log\left(\frac{\mu_{n,i}+1}{k}\right) - \log f(y_i|x_i)\right| \leq \sup_{(u,x,y)\in G_n}\left|\log\left(\frac{\mu_n(k,\beta,u,x,y)+1}{k}\right) - \log f(y|x)\right| \to 0,$$

as $n\to\infty$. In a similar way

$$\sup_{0<u_i\leq\delta}\left|\log\left(\frac{\mu_{n,i}+2}{k}\right) - \log f(y_i|x_i)\right| \to 0,\quad n\to\infty,\quad i=1,2.$$

Therefore, for P_X-almost all $x_i\in\mathbb{R}^d$ and any $y_i\in M$ $(i=1,2)$,

$$\sup_{0<u_i\leq\delta}|\mathcal{J}_{n,i} - (-\log f(y_i|x_i) - H(Y|X))| \to 0,\quad n\to\infty. \tag{5.18}$$

Set $F_i := -\log f(y_i|x_i) - H(Y|X)$. Then one has

$$\sup_{0<u_1,u_2\leq\delta}|\mathcal{J}_{n,1}\mathcal{J}_{n,2} - F_1F_2| \leq \sup_{0<u_1\leq\delta}|\mathcal{J}_{n,1}|\sup_{0<u_2\leq\delta}|\mathcal{J}_{n,2}-F_2| + |F_2|\sup_{0<u_1\leq\delta}|\mathcal{J}_{n,1}-F_1|.$$

In view of (5.18), for $n\geq N(x_i,y_i,\delta)$, the following inequality holds $\sup_{0<u_1\leq\delta}|\mathcal{J}_{n,1}|\leq 2|F_1|$. Whence, for all $y_1,y_2\in M$ and $\mathsf{P}_X\otimes\mathsf{P}_X$-almost all $(x_1,x_2)\in\mathbb{R}^d\times\mathbb{R}^d$,

$$\sup_{u_1,u_2<\delta}|\mathcal{J}_{n,1}\mathcal{J}_{n,2} - F_1F_2| \to 0,\quad n\to\infty.$$

Thus (5.14) is proved.

Step 9. Now we come to the final part of the proof. Due to the previous step we get

$$\begin{aligned}
&|\mathcal{U}_{n,k,1}(z_1,z_2)-F_1F_2|\\
&\quad\le\Big|\int_0^\delta\int_0^\delta J_{n,k,\beta}(z_1,z_2,u_1,u_2)f_{n,k,\beta}(x_1,x_1,u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2-F_1F_2\Big|\\
&\qquad+\Big|\int_0^\delta\int_0^\delta J_{n,k,\beta}(z_1,z_2,u_1,u_2)f_{n,k,\beta}(x_1,x_1,u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2\Big|(1-\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta))\\
&\quad\le\sup_{0<u_1,u_2\le\delta}|\mathcal{J}_{n,1}\mathcal{J}_{n,2}-F_1F_2|\int_0^\delta\int_0^\delta f_{n,k,\beta}(x_1,x_2,u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2\\
&\qquad+|\log k+H(Y|X)|(1-\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta)).
\end{aligned}$$

Moreover,

$$1-\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta)\le\mathsf{P}\left(\{\rho_{n,k,1}>\delta n^{-\beta}\}\cup\{\rho_{n,k,2}>\delta n^{-\beta}\}\right)\le 2\mathsf{P}\left(\rho_{n,k,1}>\delta n^{-\beta}\right).$$

In view of (4.8) and (4.17) we get that, for any $z_j\in\mathbb{R}^d\times M$, $j=1,2$,

$$|\log k+H(Y|X)|(1-\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta))\to 0,\quad n\to\infty.$$

Therefore, $\mathcal{U}_{n,k,1}(z_1,z_2)-F_1F_2\to 0$ as $n\to\infty$. Hence (5.13) is established and the proof of Theorem 2.4 is complete. □

5.1. Proof of Corollary 2.5

Let $X\sim N(a,\Sigma)$ where $a\in\mathbb{R}^d$ and $\Sigma>0$. It is easily seen that, for any $\varepsilon>0$, one has $\mathsf{E}|\log f_X(X)|^\varepsilon<\infty$ (see, *e.g.*, [10]). Since, for $x\in\mathbb{R}^d$ and $y\in M$,

$$\begin{aligned}
f(x,y)&=\mathsf{P}(Y=y|X=x)f_X(x),\\
\mathsf{P}(Y=1|X=x)&=\frac{1}{1+\exp\{-(w,x)-b\}}>0,\\
f_X(x)&=\frac{1}{(2\pi)^{d/2}|\det\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-a)^T\Sigma^{-1}(x-a)\right\}>0,
\end{aligned}$$

we see that $f(\cdot,y)$ is μ-almost everywhere positive. We show that $f(\cdot,y)$ is C_0-constricting for any $y\in\{0,1\}$. According to Remark 2.2, it is sufficient to verify that this function is a Lipschitz one. Write

$$\begin{aligned}
|f(u,y)-f(v,y)|&=|\mathsf{P}(Y=y|X=u)f_X(u)-\mathsf{P}(Y=y|X=v)f_X(v)|\\
&\le\mathsf{P}(Y=y|X=u)|f_X(u)-f_X(v)|+f_X(v)|\mathsf{P}(Y=y|X=u)-\mathsf{P}(Y=y|X=v)|\\
&\le|f_X(u)-f_X(v)|+\max_x|f_X(x)|\,|\mathsf{P}(Y=y|X=u)-\mathsf{P}(Y=y|X=v)|.
\end{aligned}$$

Note that $\max_x|f_X(x)|<\infty$ and $f_X(\cdot)$ satisfies the Lipschitz condition. Thus it is enough to prove that the function $\mathsf{P}(Y=y|X=x)$ (as a function in x) satisfies the Lipschitz condition with a constant C. For any $x\in\mathbb{R}^d$

and $j \in \{1, \dots, d\}$,

$$\left|\frac{\partial}{\partial x_j}\mathsf{P}(Y=1|X=x)\right| = \frac{\exp\{-(w,x)-b\}}{(1+\exp\{-(w,x)-b\})^2}|w_j| \le \frac{1}{4}\|w\| < \infty.$$

Thus for $\mathsf{P}(Y=1|X=x)$ the desired property holds. Obviously, one has $\mathsf{P}(Y=0|X=x) = 1-\mathsf{P}(Y=1|X=x)$ and consequently $\mathsf{P}(Y=0|X=x)$ satisfies the Lipschitz condition as well. □

Appendix A.

A.1 Proof of Lemma 3.1

Fix $n>1$ and $k \in \{1, \dots, n-1\}$. For $t>0$, $x \in \mathbb{R}^d$ and $\delta \in (0,t)$, introduce an event

$$A_{n,k}(x,t,\delta) := \{t-\delta < \rho_{n,k,1}(x) \le t+\delta\}.$$

The event $\{\rho_{n,k,1}(x) > t-\delta\}$ means that there are less than k points (*i.e.* $0, 1, \dots, k-1$) among $X_2, \dots, X_n$ in the ball $B(x, t-\delta)$. The event $\{\rho_{n,k}(x) \le t+\delta\}$ signifies that at least k points (*i.e.* $k, k+1, \dots, n-1$) among $X_2, \dots, X_n$ are contained in the ball $B(x,t+\delta)$. For $x \in \mathbb{R}^d$ and $t>0$, consider the set

$$S_x(t,\delta) := \{z \in \mathbb{R}^d : t-\delta < \|z-x\| \le t+\delta\}.$$

Let $P_x(t,\delta) := \mathsf{P}(X \in S_x(t,\delta))$. Note that $P_x(t,\delta) \to 0$ as $\delta \to 0+$ since $\mu(S_x(t,\delta)) \to 0$ as $\delta \to 0+$, and because P_X is absolutely continuous w.r.t. the Lebesgue measure μ. For $2 \le i_1 < \ldots < i_q \le n$, taking into account the independence of $X_1, \dots, X_n$, one has

$$\mathsf{P}(X_{i_1} \in S_x(t,\delta), \dots, X_{i_q} \in S_x(t,\delta)) = P_x(t,\delta)^q.$$

Note that

$$A_{n,k}(x,t,\delta) = \bigcup_{(s,m)\in J} B_s D_m G_{n-1-(s+m)},$$

here $J := \{(s,m) : s \in \{0,\dots,k-1\}, m \in \{1,\dots,n-1\}, k \le s+m \le n-1\}$, $J = J(n,k)$,

$$\begin{aligned} B_s &:= \{s \text{ observations among } X_2,\dots,X_n \text{ are in } \ B(x,t-\delta)\},\\ D_m &:= \{m \text{ observations among } X_2,\dots,X_n \text{ belong to } S(t,\delta)\},\\ G_{n-1-(s+m)} &:= \{n-1-(s+m) \text{ observations among } X_2,\dots,X_n \text{ are in } \mathbb{R}^d \setminus B(x,t+\delta)\}. \end{aligned}$$

Clearly, the events B_s, D_m and $G_{n-1-(s+m)}$ depend on x,t,δ,n. We get

$$\mathsf{P}(A_{n,k}(x,t,\delta)) = \mathsf{P}(B_{k-1}D_1G_{n-1-k}) + O(P_x(t,\delta)^2), \quad \delta \to 0+, \tag{A.1}$$

as $\mathsf{P}(D_m) = \binom{n-1}{m}P_x(t,\delta)^m(1-P_x(t,\delta))^{n-1-m}$ and $\mathsf{P}(B_sD_mG_{n-1-(s+m)}) \le \mathsf{P}(D_m)$ for $(m,s) \in J$. Set $p_x(u) = \mathsf{P}(X \in B(x,u))$, $x \in \mathbb{R}^d$, $u>0$. Then

$$\mathsf{P}(B_{k-1}D_1G_{n-1-k}) = \binom{n-1}{k-1}(p_x(t-\delta))^{k-1}\binom{n-k}{1}P_x(t,\delta)(1-p_x(t+\delta))^{n-1-k}.$$

Indeed, there exist $\binom{n-1}{k-1}$ variants to choose $k-1$ points among $X_2,\dots,X_n$, contained in $B(x,t-\delta)$, and after that $n-1-(k-1)$ variants to choose one point contained in $S_x(t,\delta)$. Other points (their cardinality is $n-1-k$) belong to the complement to the ball $B(x,t+\delta)$.

Now note that, for $r=0,1,\dots,k-1$,

$$\mathsf{P}(\xi_{n,k,1}(x,y)=r,A_{n,k}(x,t,\delta))=\mathsf{P}(\xi_{n,k,1}(x,y)=r,B_{k-1}D_1G_{n-1-k})+O(P_x(t,\delta)^2),\quad \delta\to 0+,$$

and

$$\begin{aligned}&\mathsf{P}(\xi_{n,k,1}(x,y)=r,B_{k-1}D_1G_{n-1-k})\\ &\quad=\binom{n-1}{k-1}\binom{k-1}{r}\mathsf{P}(Y=y,X\in B(x,t-\delta))^r\mathsf{P}(Y\neq y,X\in B(x,t-\delta))^{k-1-r}\\ &\qquad\times\binom{n-k}{1}P(Y\neq y,X\in S_x(t,\delta))(1-p_x(t+\delta))^{n-1-k}\\ &\qquad+\binom{n-1}{k-1}\binom{k-1}{r-1}\mathsf{P}(Y=y,X\in B(x,t-\delta))^{r-1}\mathsf{P}(Y\neq y,X\in B(x,t-\delta))^{k-1-(r-1)}\\ &\qquad\times\binom{n-k}{1}\mathsf{P}(Y=y,X\in S_x(t,\delta))(1-p_x(t+\delta))^{n-1-k}.\end{aligned}\tag{A.2}$$

We take into account that there are $\binom{n-1}{k-1}$ variants to choose $k-1$ points among $X_2,\dots,X_n$ which lay in $B(x,t-\delta)$, and there exist $\binom{n-k}{1}$ variants to choose among other observations a point X_q belonging to $S_x(t,\delta)$. Further on there exist two possibilities.

1. If $Y_q\neq y$ then there are $\binom{k-1}{r}$ variants to choose among points, contained in $B(x,t-\delta)$, r points $X_{i_1},\dots,X_{i_r}$ such that $Y_{i_m}=y$, $m=1,\dots,r$. For other $k-1-r$ points $X_{j_1},\dots,X_{j_{k-1-r}}$, belonging to $B(x,t-\delta)$ one has $Y_{j_s}\neq y$, $s=1,\dots,k-1-r$.
2. If $Y_q=y$ then there are $\binom{k-1}{r-1}$ variants to choose among points, contained in $B(x,t-\delta)$, $r-1$ points $X_{i_1},\dots,X_{i_{r-1}}$ such that $Y_{i_m}=y$, $m=1,\dots,r-1$. For other $k-1-(r-1)$ points $X_{j_1},\dots,X_{j_{k-1-(r-1)}}$ belonging to $B(x,t-\delta)$ one has $Y_{j_s}\neq y$, $s=1,\dots,k-1-(r-1)$.

Other $n-1-k$ points have to be in the complement of the ball $B(x,t+\delta)$. The probability, for each observation X_m, to be in this complement is equal to $1-p_x(t+\delta)$.

For $r=k$, we get

$$\begin{aligned}&\mathsf{P}(\xi_{n,k,1}(x,y)=k,B_{k-1}D_1G_{n-1-k})\\ &\quad=\binom{n-1}{k-1}\binom{n-k}{1}\mathsf{P}(Y=y,X\in B(x,t-\delta))^{k-1}\mathsf{P}(Y=y,X\in S_x(t,\delta))(1-p_x(t+\delta))^{n-1-k}.\end{aligned}$$

In this case the reasoning is analogous to the previous one. The difference is the following. Not only for each $(k-1)$ points $X_{i_1},\dots,X_{i_{k-1}}$ (among $X_2,\dots,X_n$, belonging to $B(x,t-\delta)$), one has $Y_{i_1}=y,\dots,Y_{i_{k-1}}=y$, but also for X_q contained in $S_x(t,\delta)$ one has $Y_q=y$. The case $r=k$ is comprised by formula (A.2) since $\binom{k-1}{k}=0$.

If a random variable τ is such that $\tau\geq 0$ a.s., $\mathsf{E}\tau<\infty$ and a random vector ζ takes values in $\mathbb{R}^s$ then (see, *e.g.*, [39], Chap. II, Sect. 7.5) the function $\mathsf{E}(\tau|\zeta=t)$ can be defined in the following way. Set $G(B):=\mathsf{E}(\tau\mathbb{I}\{\zeta\in B\})$ where $B\in\mathcal{B}(\mathbb{R}^s)$. Evidently, G is a finite measure which is absolutely continuous w.r.t. P_ζ. Therefore there is a Borel function $\varphi : \mathbb{R}^s \to \mathbb{R}$ such that, for each $B \in \mathcal{B}(\mathbb{R}^s)$,

$$\mathsf{E}(\tau \mathbb{I}\{\zeta \in B\}) = \int_B \varphi(x) \mathsf{P}_\zeta(\mathrm{d}x).$$

In other words $\varphi(t)$ is the Radon - Nikodym derivative $\frac{\mathrm{d}G}{\mathrm{d}\mathsf{P}_\zeta}(t)$, $t \in \mathbb{R}^s$. Thus $\mathsf{E}(\tau|\zeta) = \varphi(\zeta)$. According to Theorem 5.8.8 [8] (we take into account that $G \ll \mathsf{P}_\zeta$) there exists

$$\lim_{\delta \to 0+} \frac{G(B(t,\delta))}{\mathsf{P}_\zeta(B(t,\delta))} = \frac{\mathrm{d}G}{\mathrm{d}\mathsf{P}_\zeta}(t), \quad t \in \mathbb{R}^s. \tag{A.3}$$

More precisely, this limit exists for P_ζ-almost all $t \in \mathbb{R}^s$ and is the Radon–Nikodym derivative of the measure G w.r.t. the measure P_ζ, that is a (version) of $\mathsf{E}(\tau|\zeta = t)$. We employ this result for $\tau = \mathbb{I}\{\xi_{n,k,1}(x,y) \in D\}$ where $D \in \mathcal{B}(\mathbb{R}_+)$, $\zeta = \rho_{n,k,1}(x)$, $x \in \mathbb{R}^d$, $y \in M$. Clearly, τ is an integrable random variable w.r.t. any finite measure. Formula (A.3) can be rewritten for $\mathsf{P}_{\rho_{n,k,1}(x)}$-almost all $t \in (0,\infty)$ as follows

$$\mathsf{P}(\xi_{n,k,1}(x,y) \in D | \rho_{n,k,1}(x) = t) = \lim_{\delta \to 0+} \frac{\mathsf{P}(\xi_{n,k,1}(x,y) \in D, \rho_{n,k,1}(x) \in B(t,\delta))}{\mathsf{P}(\rho_{n,k,1}(x) \in B(t,\delta))}. \tag{A.4}$$

Note that instead of $B(t,\delta) = [t-\delta, t+\delta]$, where $0 < \delta < t$, we can take a set $(t-\delta, t+\delta]$ since, for any $n \in \mathbb{N}$, $n > 1$, $k \in \{0, \dots, n-1\}$ and $x \in \mathbb{R}^d$, a random variable $\rho_{n,k,1}(x)$ has a density. Indeed, $\mathsf{P}(\rho_{n,k,1}(x) \le 0) = 0$ as there exists a density $f_X(\cdot)$ and, for $t > 0$,

$$\mathsf{P}(\rho_{n,k,1}(x) \le t) = \sum_{j=k}^{n-1} \binom{n-1}{j} p_x(t)^j (1 - p_x(t))^{n-1-j} \tag{A.5}$$

where $p_x(t) = \mathsf{P}(X \in B(x,t))$. Evidently, $p_x(t) = \mathsf{P}(\|X - x\| \le t)$ is a distribution function of a positive random variable $\|X - x\|$. At first we show that $p_x(\cdot)$ has a density $f_x(\cdot)$ w.r.t. the Lebesgue measure on $\mathcal{B}(\mathbb{R}_+)$. After that we prove that there exists a density of a random variable $\rho_{n,k,1}(x)$.

We know that $X = (X_1, \dots, X_d)$ has a density $f_X(\cdot)$ w.r.t. the Lebesgue measure μ in $\mathbb{R}^d$ (*i.e.* $\mathsf{P}_X \ll \mu$). On the other hand, since $f_X(x)$ is strictly positive for μ-almost all $x \in \mathbb{R}^d$, it is easily seen that $\mu \ll \mathsf{P}_X$. Consequently, $\mathsf{P}_X \sim \mu$.

Let μ_1 and μ_2 be some measures on a space $(S, \mathcal{B})$ and $h : S \to T$ be $\mathcal{B}|\mathcal{D}$-measurable function, where T is endowed with a σ-algebra $\mathcal{D}$. Introduce the measures $\nu_i := \mu_i h^{-1}$, $i = 1,2$. Then, obviously, $\mu_1 \ll \mu_2$ yields $\nu_1 \ll \nu_2$. If Q is a Gaussian measure on $\mathcal{B}(\mathbb{R}^d)$ having a density w.r.t. μ, then $Q \sim \mu$ as there exists a strictly positive version of $dQ/d\mu$ on $\mathbb{R}^d$. Consider $(S, \mathcal{B}) := (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $(T, \mathcal{D}) := (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, $h : \mathbb{R}^d \to \mathbb{R}_+$, where $h(x) = x_1^2 + \ldots + x_d^2$ for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. Let $\mu_1 = \mathsf{P}_X$ and μ_2 be a Gaussian law $N(0, I)$ in $\mathbb{R}^d$ with zero mean vector and the unit covariance matrix I. Then $\mu_1 \sim \mu_2$ since $\mu_1 \sim \mu$ and $\mu_2 \sim \mu$. Consequently, $\nu_1 \sim \nu_2$. Clearly, $\nu_2 = \mu_2 h^{-1}$ has the χ_d^2-distribution with a density w.r.t. the Lebesgue measure $\Lambda_{\mathbb{R}_+}$ on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, *i.e.*

$$\mathsf{P}_{\chi_d^2}(u) = \frac{u^{\frac{d}{2}-1} e^{-\frac{u}{2}}}{2^{\frac{d}{2}} \Gamma(\frac{d}{2})}, \quad u \ge 0.$$

This density is strictly positive on $(0,\infty)$ and therefore $\nu_2 \sim \Lambda_{\mathbb{R}_+}$. Thus $\mathsf{P}_X h^{-1} \sim \nu_2$, hence $\mathsf{P}_X h^{-1} \sim \Lambda_{\mathbb{R}_+}$. We proved that there exists the density g of a random variable $X_1^2 + \ldots + X_d^2$ w.r.t. the Lebesgue measure on $\mathbb{R}_+$.

Write $d(\mathsf{P}_X h^{-1})/d\Lambda_{\mathbb{R}_+} = g$. For any $B \in \mathcal{B}(\mathbb{R}_+)$, one has

$$\int_B g(t)\mathrm{d}t = \mathsf{P}_X h^{-1}(B).$$

If $\int_B g(t)\mathrm{d}t = 0$ then $\mathsf{P}_X h^{-1}(B) = 0$ and hence $\Lambda_{\mathbb{R}_+}(B) = 0$. Take $B := \{t \in \mathbb{R}_+ : g(t) = 0\}$. Then $\int_B g(t)\mathrm{d}t = 0$ and, therefore, $\Lambda_{\mathbb{R}_+}\{t : g(t) = 0\} = 0$. In other words, g is strictly positive $\Lambda_{\mathbb{R}_+}$-almost everywhere.

If a random vector V has a density (w.r.t. measure μ) $q(z)$, $z \in \mathbb{R}^d$, then, for $x \in \mathbb{R}^d$, the vector $V - x$ has a density $q_x(z) = q(z + x)$, $z \in \mathbb{R}^d$. Consequently, we can claim that, for each $x \in \mathbb{R}^d$, there exists a density of random variable $\|X - x\|^2$ w.r.t. the Lebesgue measure $\Lambda_{\mathbb{R}_+}$ on $\mathbb{R}_+$. This density is strictly positive w.r.t. $\Lambda_{\mathbb{R}_+}$ whenever $f_X(\cdot)$ is strictly positive w.r.t. μ. If a random variable $\xi_x \geq 0$ has a density $\gamma_x(u)$, $u \geq 0$ $(x \in \mathbb{R}^d)$, then the random variable $\sqrt{\xi_x}$ has a density $p_x(u) = 2u\gamma_x(u^2)$, $u \geq 0$. Thus there is a density $f_x(u)$, $u \geq 0$, of a random variable $\|X - x\|$, this density is strictly positive for $\Lambda_{\mathbb{R}_+}$-almost all $u \geq 0$ and $\mathsf{P}_{\|X-x\|} \sim \Lambda_{\mathbb{R}_+}$.

Now we can prove that the density (w.r.t. $\Lambda_{\mathbb{R}_+}$) of a random variable $\rho_{n,k,1}(x)$ has the form

$$h_{n-1,k,x}(u) = \sum_{j=k}^{n-1} \binom{n-1}{j} \left(j p_x(u)^{j-1}(1-p_x(u))^{n-1-j} - p_x(u)^j (n-j-1)(1-p_x(u))^{n-j-2}\right) f_x(u) \tag{A.6}$$

where $f_x(\cdot)$ is a density corresponding to the distribution function $p_x(\cdot)$, $x \in \mathbb{R}^d$.

It is worth to emphasize that we cannot differentiate the distribution function to find the density (as the celebrated Cantor function shows). Thus we have to employ the integral relations. Let F be a distribution function of a positive random variable with a density f (thus $F(0) = 0$ and $f(u) = 0$, $u < 0$). Then using the integration by parts (see, *e.g.*, [39], Chap. II, Sect. 6.12) and an induction one can prove that, for each $n \in \mathbb{N}$, a distribution function $F^n(u)$ has a density $nF^{n-1}(u)f(u)$, $u \in \mathbb{R}$. For $m \in \mathbb{N}$ and $j = 0, \ldots, m$, we can write

$$\begin{aligned} F(t)^j(1 - F(t))^{m-j} &= \sum_{r=0}^{m-j} \binom{m-j}{r} (-1)^{m-j-r} F(t)^{j+m-j-r} \\ &= \sum_{r=0}^{m-j} \binom{m-j}{r} (-1)^{m-j-r} \int_{(0,t]} (m-r)F^{m-r-1}(u)f(u)\mathrm{d}u. \end{aligned}$$

The latter formula and (A.5) lead, for $t \geq 0$, to the relation

$$\mathsf{P}(\rho_{n,k,1}(x) \leq t) = \int_0^t h_{n-1,k,x}(u)\mathrm{d}u \tag{A.7}$$

where $h_{n-1,k,1}$ is given in (A.6). We set $h_{n-1,k,1}(u) = 0$ for $u < 0$. Note that we come to (A.6) using the polynome of the degree $n - 1$ in a distribution function $p_x(\cdot)$. Hence $h_{n-1,k,x}$ is an integrable function (w.r.t. the Lebesgue measure Λ on $\mathbb{R}$). However, the mentioned polynome has positive and negative coefficients. Therefore, we have to clarify why $h_{n-1,k,x}$ is a probability density. We explain that if, for a distribution function F, one has

$$F(t) = \int_{(-\infty,t]} f(u)\mathrm{d}u, \quad t \in \mathbb{R},$$

where $f \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \Lambda)$, then f is a probability density. Clearly, the Lebesgue theorem on dominated convergence yields that $\int_{\mathbb{R}} f(u)\mathrm{d}u = 1$ as $\lim_{t\to\infty} F(t) = 1$. It remains to show that $f(u) \geq 0$ for Λ-almost all u.

Introduce a function

$$Q(B) := \int_B f(u)\mathrm{d}u, \quad B \in \mathcal{B}(\mathbb{R}).$$

Obviously, $Q((a,b]) = F(b) - F(a) \ge 0$ for $-\infty \le a \le b \le \infty$ (we set $F(-\infty) := 0$, $F(\infty) := 1$ and $(a,\infty] := (a,\infty)$). Let G be a probability measure on $\mathcal{B}(\mathbb{R})$ generated by a distribution function F, then $G((a,b]) := F(b) - F(a)$. We see that G and Q coincide on an algebra $\mathcal{A}$ consisting of the finite unions of the pair-wise disjoint intervals having the form $(a,b]$, $-\infty \le a \le b \le \infty$. Hence, Q is a finite nonnegative function on $\mathcal{A}$. Clearly, Q is a countably additive function on $\mathcal{B}(\mathbb{R})$ and $Q(\mathbb{R})$ is finite. It remains to note that, for any $B \in \mathcal{B}(\mathbb{R})$ and each $\varepsilon > 0$, there exists $A \in \mathcal{A}$ such that $|Q(B) - Q(A)| < \varepsilon$. Indeed,

$$|Q(B) - Q(A)| = \left|\int_{\mathbb{R}} (\mathbb{I}\{B\} - \mathbb{I}\{A\}) f(u)\mathrm{d}u\right| \le \int_{\mathbb{R}} \mathbb{I}\{B \triangle A\}|f(u)|\mathrm{d}u.$$

Consequently, for each $B \in \mathcal{B}(\mathbb{R})$, one can find $A_n \in \mathcal{A}$ ($n \in \mathbb{N}$) such that $Q(A_n) \to Q(B)$ as $n \to \infty$. Taking into account that $Q(A_n) \ge 0$ we get $Q(B) \ge 0$. Assume now that $\mu(E) > 0$ where $E = \{x : f(x) < 0\}$. Note that $E = \cup_{m=1}^{\infty}\{-\infty < f(x) \le -\frac{1}{m}\}$. Then in a standard way we come to the contradiction. Therefore, $\mu(E) = 0$. Thus formula (A.6) provides a probability density $h_{n-1,k,x}(\cdot)$ of the random variable $\rho_{n,k,1}(x)$ distribution where $x \in \mathbb{R}^d$.

Hence, for each $x \in \mathbb{R}^d$, $y \in M$, $r \in \{0, 1, \dots, k\}$ and $\mathsf{P}_{\rho_{n,k,1}(x)}$-almost all $t \in (0,\infty)$ in view of (A.4) one has

$$\mathsf{P}(\xi_{n,k,1}(x,y) = r | \rho_{n,k,1}(x) = t) = \lim_{\delta \to 0+} \frac{\mathsf{P}(\xi_{n,k,1}(x,y) = r, A_{n,k}(x,t,\delta))}{\mathsf{P}(A_{n,k}(x,t,\delta))}. \tag{A.8}$$

Applying the expressions obtained for nominator and denominator of the latter fraction in (A.8) and taking into account that a function $\mathsf{P}(X \in B(x,t))$ is continuous in $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$ (see, *e.g.*, [10], Lem. 2.1) we get, for each $n \in \mathbb{N}$ ($n > 1$), $k \in \{1, \dots, n-1\}$, $r = 0, \dots, k$ and $\mathsf{P}_{\rho_{n,k,1}(x)}$-almost all $t \in (0,\infty)$,

$$\begin{aligned} &\mathsf{P}(\xi_{n,k,1}(x,y) = r | \rho_{n,k,1}(x) = t) \\ &\quad = \binom{k-1}{r} p^r (1-p)^{k-1-r} \lim_{\delta \to 0+} \frac{\mathsf{P}(Y \ne y, X \in S_x(t,\delta))}{\mathsf{P}(X \in S_x(t,\delta))} \\ &\quad + \binom{k-1}{r-1} p^{r-1} (1-p)^{k-r} \lim_{\delta \to 0+} \frac{\mathsf{P}(Y = y, X \in S_x(t,\delta))}{\mathsf{P}(X \in S_x(t,\delta))} \end{aligned} \tag{A.9}$$

where $p := \mathsf{P}(Y = y | X \in B(x,t))$, $p = p(x,y,t)$. We used that $\mathsf{P}(X \in B(x,t)) > 0$, for μ-almost all $x \in \mathbb{R}^d$ and $t > 0$, since $f_X(z)$ is strictly positive for μ-almost all $z \in \mathbb{R}^d$. However, we have to explain the existence of limits in (A.9). Let us employ formula (A.3) for $\zeta := \|X - x\|$, $G(D) := \mathsf{E}(\mathbb{I}\{\zeta \in D\}\mathbb{I}\{Y = y\})$, $x \in \mathbb{R}^d$, $y \in M$ and $D \in \mathcal{B}(\mathbb{R})$. We can claim that, for $x \in \mathbb{R}^d$ and $y \in M$, the limits appearing in (A.9) exist for $\mathsf{P}_{\|X-x\|}$-almost all $t > 0$. Indeed,

$$\begin{aligned} \lim_{\delta \to 0+} \frac{\mathsf{P}(Y \ne y, X \in S_x(t,\delta))}{\mathsf{P}(X \in S_x(t,\delta))} &= \lim_{\delta \to 0+} \frac{\mathsf{P}(Y \ne y, t - \delta < \|X - x\| \le t + \delta)}{\mathsf{P}(t - \delta < \|X - x\| \le t + \delta)} \\ &= \mathsf{P}(Y \ne y | \|X - x\| = t) := \alpha(x,y,t). \end{aligned} \tag{A.10}$$

We have seen that $\mathsf{P}_{\|X-x\|} \sim \Lambda_{\mathbb{R}_+}$ for each $x \in \mathbb{R}^d$. Therefore $\mathsf{P}(X \in S_x(t,\delta)) > 0$ for all $x \in \mathbb{R}^d$ and $\delta > 0$. Moreover, for each y belonging to a finite set M and $x \in \mathbb{R}^d$, the limits in (A.9) exist for $\Lambda_{\mathbb{R}_+}$-almost all

$t \in (0,\infty)$ as $\mathsf{P}_{\|X-x\|} \sim \Lambda_{\mathbb{R}_+}$. Consequently, the measure $\mathsf{P}_{\rho_{n,k,1}(x)}$ of a set of points $t \in \mathbb{R}_+$ such that the limits in (A.9) do not exist equals to zero since $\mathsf{P}_{\rho_{n,k,1}(x)} \ll \Lambda_{\mathbb{R}_+}$. The proof is complete. □

A.2 Proof of Lemma 3.3

Fix arbitrary $z_j = (x_j, y_j) \in \mathbb{R}^d \times M$, $j \in \{1,2\}$, such that $x_1 \neq x_2$. Consider $n \in \mathbb{N}$, $n > 2$, and $k \in \{1,\dots,n-1\}$. Note that

$$\xi_{n,k,1}(z_1,z_2) = \sharp\{i : i \in \{3,\dots,n\}, Y_i = y_1, X_i \in B(x_1,R_1)\} + \mathbb{I}\{y_2 = y_1, x_2 \in B(x_1,R_1)\},$$
$$\xi_{n,k,2}(z_1,z_2) = \sharp\{i : i \in \{3,\dots,n\}, Y_i = y_2, X_i \in B(x_2,R_2)\} + \mathbb{I}\{y_1 = y_2, x_1 \in B(x_2,R_2)\},$$

here $R_j := \rho_{n,k,j}(x_1,x_2)$, $j = 1,2$. Recall that $\rho_{n,k,j}(x_1,x_2)$ and $\xi_{n,k,j}(z_1,z_2)$ are defined in (3.1) and (3.2), respectively. Take any $\varepsilon \in (0, |x_1-x_2|/2)$ and $t_j \in (0, |x_1-x_2|/2-\varepsilon)$, $j=1,2$. Then there exists $\delta > 0$ such that $\delta < t_j$, $j=1,2$, and $t_1+t_2+2\delta < |x_1-x_2|$. Introduce the events

$$A_{n,k} := \bigcap_{j=1}^{2}\{t_j - \delta < \rho_{n,k,j}(x_1,x_2) \le t_j + \delta\}.$$

Clearly, $A_{n,k} = A_{n,k}(x_1,x_2,t_1,t_2,\delta)$. Further on in the proof we consider $j \in \{1,2\}$ without mentioning. To simplify the exposition we use a notation similar in meaning to that employed for proving Lemma 3.1. However, we have to emphasize that now we use vectors with two components in contrast to random variables appearing in the proof of Lemma 3.1. For instance, $A_{n,k}$ is not the same as previously. An event $\{\rho_{n,k,j}(x_1,x_2) > t_j - \delta\}$ means that the closed ball $B(x_j, t_j-\delta)$ contains less than k (*i.e.* $0,1,\dots,k-1$) observations among $X_3,\dots,X_n$, since x_i, $i \in \{1,2\}\setminus\{j\}$, does not belong to this ball as $|x_1-x_2| > t_j - \delta$. An event $\{\rho_{n,k,j}(x_1,x_2) \le t_i + \delta\}$ signifies that in $B(x_j, t_j+\delta)$ there are at least k (*i.e.* $k, k+1,\dots,n-2$) points among $X_3,\dots,X_n$ because x_i, $i \in \{1,2\}\setminus\{j\}$, does not belong to this ball as $|x_1-x_2| > t_j+\delta$. One has

$$A_{n,k} = \bigcup_{(s_1,s_2,m_1,m_2)\in J_{n,k}} B_{1,s_1}D_{1,m_1}B_{2,s_2}D_{2,m_2}G_{n-2-(s_1+s_2+m_1+m_2)}$$

where $J_{n,k}$ consists of (s_1,s_2,m_1,m_2) such that $s_1,s_2 \in \{0,\dots,k-1\}$, $m_1,m_2 \in \{1,\dots,n-1\}$, $s_1+m_1 \ge k$, $s_2+m_2 \ge k$, $s_1+s_2+m_1+m_2 \le n-2$ and

$$B_{j,s} := \{s \text{ variables among } X_3,\dots,X_n \text{ belong to } B(x_j,t_j-\delta)\},$$

$$D_{j,m} := \{m \text{ variables among } X_3,\dots,X_n \text{ belong to } S_{x_j}(t_j,\delta)\},$$

$$G_l := \left\{ l \text{ variables among } X_3,\dots,X_n \text{ belong to } \mathbb{R}^d \setminus \bigcup_{j=1}^{2} B(x_j,t_j+\delta) \right\}.$$

Since $k \le [(n-2)/2]$ the set $J_{n,k}$ is nonempty as $(k-1,k-1,1,1) \in J_{n,k}$. More precisely, $B_{j,s} = B_{j,s}(x_j,t_j,\delta,n)$, $D_{j,m} = D_{j,m}(x_j,t_j,\delta,n)$ and $G_l = G_l(x_1,x_2,t_1,t_2,\delta)$. In a way similar to (A.12) one has

$$\mathsf{P}(A_{n,k}) = \mathsf{P}(B_{1,k-1}D_{1,1}B_{2,k-1}D_{2,1}G_{n-2-2k}) + O(P_{x_1}(t_1,\delta)^2 P_{x_2}(t_2,\delta)) + O(P_{x_1}(t_1,\delta)P_{x_2}(t_2,\delta)^2)$$

as $\delta \to 0+$ because, for $(s_1, s_2, m_1, m_2) \in J_{n,k}$,

$$\mathsf{P}(D_{1,m_1} D_{2,m_2}) = \binom{n-2}{m_1} P_{x_1}(t_1,\delta)^{m_1} \binom{n-2-m_1}{m_2} P_{x_2}(t_2,\delta)^{m_2} (1 - P_{x_1}(t_1,\delta) - P_{x_2}(t_2,\delta))^{n-2-m_1-m_2},$$

$$\mathsf{P}(B_{1,s_1} D_{1,m_1} B_{2,s_2} D_{2,m_2} G_{n-2-(s_1+s_2+m_1+m_2)}) \le \mathsf{P}(D_{1,m_1} D_{2,m_2}),$$

$P_x(t,\delta) = \mathsf{P}(X \in S_x(t,\delta))$, $x \in \mathbb{R}^d$, $t > 0$ and $\delta > 0$. It is easily seen that

$$\begin{aligned}&\mathsf{P}(B_{1,k-1} D_{1,1} B_{2,k-1} D_{2,1} G_{n-2-2k}) = Poly(k-1, 1, k-1, 1, n-2-2k)\\ &\quad \times (p_{x_1}(t_1 - \delta))^{k-1} P_{x_1}(t_1,\delta)(p_{x_2}(t_2 - \delta))^{k-1} P_{x_2}(t_2,\delta)(1 - p_{x_1}(t_1 + \delta) - p_{x_2}(t_2 + \delta))^{n-2-2k}\end{aligned}$$

where $p_x(t) = \mathsf{P}(X \in B(x,t))$, $x \in \mathbb{R}^d$, $t > 0$ and

$$Poly(k_1, \ldots, k_q) := \frac{(k_1 + \ldots + k_q)!}{k_1! \ldots k_q!}, \;\; k_i \in \mathbb{Z}_+, \; i = 1, \ldots, q.$$

Thus, for $r_1, r_2 \in \{0, \ldots, k\}$,

$$\begin{aligned}&\mathsf{P}(\xi_{n,k,1}(z_1,z_2) = r_1, \xi_{n,k,2}(z_1,z_2) = r_2, A_{n,k})\\ &\qquad = \mathsf{P}(\xi_{n,k,1}(z_1,z_2) = r_1, \xi_{n,k,2}(z_1,z_2) = r_2, B_{1,k-1} D_{1,1} B_{2,k-1} D_{2,1} G_{n-2-2k})\\ &\qquad\quad + O(P_{x_1}(t_1,\delta)^2 P_{x_2}(t_2,\delta)) + O(P_{x_1}(t_1,\delta) P_{x_2}(t_2,\delta)^2), \;\; \delta \to 0+.\end{aligned}$$

Introduce the auxiliary events. Let $B^l_{j,s}$ mean that s observations among $X_3, \ldots, X_n$ are contained in $B(x_j, t_j - \delta)$ while the rest are not, moreover, l points among X_i's contained in this ball, *i.e.* $X_{i_1}, \ldots, X_{i_l}$ are such that $Y_{i_m} = y_j$, $m = 1, \ldots, l$. Clearly, $B^l_{j,s} = B^l_{j,s}(x_j, y_j, t_j, \delta, n)$. Analogously one can define an event $D^l_{j,s}$ (namely, s points among $X_3, \ldots, X_n$ are in $S_{x_j}(t_j,\delta)$ and other ones do not belong to this set, moreover, l points among X_i's belonging to $S_{x_j}(t_j,\delta)$ are such that corresponding $Y_i = y_j$). Note that $D^l_{j,m} = D^l_{j,m}(x_j, y_j, t_j, \delta, n)$. Then, for $r_1, r_2 \in \{0, 1, \ldots, k\}$,

$$\begin{aligned}&\{\xi_{n,k,1}(z_1,z_2) = r_1, \xi_{n,k,2}(z_1,z_2) = r_2\} \cap B_{1,k-1} D_{1,1} B_{2,k-1} D_{2,1} G_{n-2-2k}\\ &= B^{r_1}_{1,k-1} D^0_{1,1} B^{r_2}_{2,k-1} D^0_{2,1} G_{n-2-2k} \cup B^{r_1-1}_{1,k-1} D^1_{1,1} B^{r_2}_{2,k-1} D^0_{2,1} G_{n-2-2k}\\ &\cup B^{r_1}_{1,k-1} D^0_{1,1} B^{r_2-1}_{2,k-1} D^1_{2,1} G_{n-2-2k} \cup B^{r_1-1}_{1,k-1} D^1_{1,1} B^{r_2-1}_{2,k-1} D^1_{2,1} G_{n-2-2k}.\end{aligned} \tag{A.11}$$

If $r = 0$ then $B^{r-1}_{j,k-1} := \varnothing$ (for $j \in \{1,2\}$). Evidently, four events appearing in the union in (A.11) are pair-wise disjoint. We evaluate their probabilities. One has

$$\begin{aligned}&\mathsf{P}(B^{r_1}_{1,k-1} D^0_{1,1} B^{r_2}_{2,k-1} D^0_{2,1} G_{n-2-2k})\\ &\quad = Poly(k-1, 1, k-1, 1, n-2-2k) \binom{k-1}{r_1} \binom{k-1}{r_2} (1 - p_{x_1}(t_1+\delta) - p_{x_2}(t_2+\delta))^{n-2-2k}\\ &\qquad \times \mathsf{P}(Y = y_1, X \in B(x_1, t_1 - \delta))^{r_1} \mathsf{P}(Y \ne y_1, X \in B(x_1, t_1 - \delta))^{k-1-r_1} P(Y \ne y_1, X \in S_{x_1}(t_1,\delta))\\ &\qquad \times \mathsf{P}(Y = y_2, X \in B(x_2, t_2 - \delta))^{r_2} \mathsf{P}(Y \ne y_2, X \in B(x_2, t_2 - \delta))^{k-1-r_2} P(Y \ne y_2, X \in S_{x_2}(t_2,\delta)).\end{aligned}$$

Indeed, there are $Poly(k-1, 1, k-1, 1, n-2-2k)$ variants for partitioning of $X_3, \ldots, X_n$ into groups belonging, correspondingly, to pair-wise disjoint (under conditions imposed on $t_1, t_2, |x_1 - x_2|$ and δ) sets $B(x_1, t_1 - \delta)$,

$S_{x_1}(t_1,\delta)$, $B(x_2,t_2-\delta)$, $S_{x_2}(t_2,\delta)$ and $\mathbb{R}^d\setminus\cup_{j=1}^2 B(x_j,t_j+\delta)$. We note that there exist $\binom{k-1}{r_1}$ variants to choose r_1 points X_i, $i\in I$, among $X_{q_1},\dots,X_{q_{k-1}}$ $(3\le q_1<\ldots<q_{k-1}\le n)$ belonging to $B(x_1,t_1-\delta)$ such that $Y_i=y_1$ for $i\in I$ and $Y_q\neq y_1$ for $q\in\{q_1,\dots,q_{k-1}\}\setminus I$, $\sharp I=r_1$. In a similar way one can explain the appearance of a factor $\binom{k-1}{r_2}$. For other three events their probabilities can be found analogously. As a result we obtain

$$
\begin{aligned}
&\mathsf{P}(\xi_{n,1}=r_1,\xi_{n,2}=r_2,B_{1,k-1}D_{1,1}B_{2,k-1}D_{2,1}G_{n-2-2k})\\
&\quad= Poly(k-1,1,k-1,1,n-2-2k)(1-p_{x_1}(t_1+\delta)-p_{x_2}(t_2+\delta))^{n-2-2k}\\
&\qquad\times\prod_{j=1}^{2}\Bigg(\binom{k-1}{r_j}\mathsf{P}(Y=y_j,F(x_j,t_j,\delta))^{r_j}\mathsf{P}(Y\neq y_j,F(x_j,t_j,\delta))^{k-1-r_j}P(Y\neq y_j,X\in S_{x_j}(t_j,\delta))\\
&\qquad+\binom{k-1}{r_j-1}\mathsf{P}(Y=y_j,F(x_j,t_j,\delta))^{r_j-1}\mathsf{P}(Y\neq y_j,F(x_j,t_j,\delta))^{k-r_j}P(Y=y_j,X\in S_{x_j}(t_j,\delta))\Bigg)
\end{aligned}
$$

where $F(x_j,t_j,\delta):=\{X\in B(x_j,t_j-\delta)\}$, $j=1,2$. Hence, in view of (A.10) and as, for each $x\in\mathbb{R}^d$, the distribution of $\|X-x\|$ is equivalent on $\mathbb{R}_+$ to the Lebesgue measure $\Lambda_{\mathbb{R}_+}$, we can state the following. For any $x_1,x_2\in\mathbb{R}^d$ $(x_1\neq x_2)$, and $\Lambda\otimes\Lambda$-almost all $t=(t_1,t_2)\in(0,\infty)\times(0,\infty)$ such that $t_1+t_2<|x_1-x_2|$, one has

$$
\lim_{\delta\to0+}\frac{\mathsf{P}(\xi_{n,k,1}(z_1,z_2)=r_1,\xi_{n,k,2}(z_1,z_2)=r_2,A_{n,k})}{\mathsf{P}(A_{n,k})}\tag{A.12}
$$

$$
\begin{aligned}
&=\lim_{\delta\to0+}\prod_{j=1}^{2}\Bigg\{\binom{k-1}{r_j}\left(\frac{\mathsf{P}(Y=y_j,X\in B(x_j,t_j-\delta))}{p_{x_1}(t_1-\delta)}\right)^{r_j}\left(\frac{\mathsf{P}(Y\neq y_j,X\in B(x_j,t_j-\delta))}{p_{x_1}(t_1-\delta)}\right)^{k-1-r_j}\\
&\quad\times\left(\frac{P(Y\neq y_j,X\in S_{x_j}(t_j,\delta))}{P(X\in S_{x_j}(t_j,\delta))}\right)\\
&\quad+\binom{k-1}{r_j-1}\left(\frac{\mathsf{P}(Y=y_j,X\in B(x_j,t_j-\delta))}{p_{x_1}(t_1-\delta)}\right)^{r_j-1}\left(\frac{\mathsf{P}(Y\neq y_j,X\in B(x_j,t_j-\delta))}{p_{x_1}(t_1-\delta)}\right)^{k-r_j}\\
&\quad\times\left(\frac{P(Y=y_j,X\in S_{x_j}(t_j,\delta))}{P(X\in S_{x_j}(t_j,\delta))}\right)\Bigg\}\\
&=\prod_{j=1}^{2}\Bigg\{\binom{k-1}{r_j}p_j^{r_j}(1-p_j)^{k-1-r_j}\lim_{\delta\to0+}\left(\frac{P(Y\neq y_j,X\in S_{x_j}(t_j,\delta))}{P(X\in S_{x_j}(t_j,\delta))}\right)\\
&\quad+\binom{k-1}{r_j-1}p_j^{r_j-1}(1-p_j)^{k-r_j}\lim_{\delta\to0+}\left(\frac{P(Y=y_j,X\in S_{x_j}(t_j,\delta))}{P(X\in S_{x_j}(t_j,\delta))}\right)\Bigg\}\\
&=\prod_{j=1}^{2}\Bigg\{\binom{k-1}{r_j}p_j^{r_j}(1-p_j)^{k-1-r_j}\alpha(x_j,y_j,t_j)+\binom{k-1}{r_j-1}p_j^{r_j-1}(1-p_j)^{k-r_j}(1-\alpha(x_j,y_j,t_j))\Bigg\}
\end{aligned}
$$

where $p_j:=p_j(x_j,y_j,t_j)=\mathsf{P}(Y=y_j|X\in B(x_j,t_j))$ and we use that a function $\mathsf{P}(X\in B(x,t))$ is continuous in $(x,t)\in\mathbb{R}^d\times\mathbb{R}_+$. We employ also that $\mathsf{P}(X\in B(x,t))>0$, for all $x\in\mathbb{R}^d$ and $t>0$, since $f_X(z)$ is strictly positive for μ-almost all $z\in\mathbb{R}^d$.

Note that the proof of Theorem 5.8.8 in [8] shows that formula (A.3) is also valid if we replace in it the balls $B(t,\delta)$ by the cubes $\widetilde{B}(t,\delta):=\prod_{i=1}^{s}[t_i-\delta,t_i+\delta]$ where $t=(t_1,\dots,t_s)\in\mathbb{R}^s$. Thus, for P_ζ-almost all $t\in\mathbb{R}^s$, one can write instead of (A.3) that

$$\lim_{\delta\to 0+}\frac{G(\tilde{B}(t,\delta))}{\mathsf{P}_{\zeta}(\tilde{B}(t,\delta))}=\mathsf{E}(\tau|\zeta=t)$$

Take now $s=2$, $\tau:=\mathbb{I}\{\xi_{n,k,1}(z_1,z_2)=r_1,\xi_{n,k,2}(z_1,z_2)=r_2\}$, $\zeta:=((\rho_{n,k,1}(x_1,x_2),\rho_{n,k,2}(x_1,x_2))$ where $n\in\mathbb{N}$, $n>2$, $k=0,\dots,n-1$, $z_1=(x_1,y_1)\in\mathbb{R}^d\times M$, $z_2=(x_2,y_2)\in\mathbb{R}^d\times M$ and $r_1,r_2\in\{0,\dots,k\}$. Then, for P_ζ-almost all $t=(t_1,t_2)\in(0,\infty)\times(0,\infty)$, we get

$$\begin{aligned}&\mathsf{P}(\xi_{n,k,1}(z_1,z_2)=r_1,\xi_{n,k,2}(z_1,z_2)=r_2|\rho_{n,k,1}(x_1,x_2)=t_1,\rho_{n,k,2}(x_1,x_2)=t_2)\\&\qquad=\lim_{\delta\to 0+}\frac{\mathsf{E}(\mathbb{I}\{\zeta\in\widetilde{B}(t,\delta)\}\tau)}{\mathsf{P}(\zeta\in\widetilde{B}(t,\delta))}.\end{aligned}\tag{A.13}$$

Let us show that, for each $x_1,x_2\in\mathbb{R}^d$ $(x_1\neq x_2)$ and P_ζ-almost all $t=(t_1,t_2)$ belonging to the set $B_{x_1,x_2}(\varepsilon)=\{t\in\mathbb{R}^2:0<t_j\le|x_1-x_2|/2-\varepsilon,\ j=1,2\}$, the latter limit coincides with the obtained value for $\lim_{\delta\to0+}\mathsf{P}(\xi_{n,k,1}(z_1,z_2)=r_1,\xi_{n,k,2}(z_1,z_2)=r_2|A_{n,k})$ in (A.12) where $A_{n,k}$ depends on x_1,x_2 and δ.

For this purpose we demonstrate that, for any $x_1,x_2\in\mathbb{R}^d$ $(x_1\neq x_2)$, there exists a positive measurable function $f_{n,k}(x_1,x_2,\cdot,\cdot)$ such that if $B\in\mathcal{B}(\mathbb{R}^2)$ and $B\subset B_{x_1,x_2}(\varepsilon)$ then

$$\mathsf{P}(\zeta\in B)=\int_B f_{n,k}(x_1,x_2,u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2.\tag{A.14}$$

For $(u_1,u_2)\in B_{x_1,x_2}(\varepsilon)$, one can write

$$\begin{aligned}&\mathsf{P}(\rho_{n,k,1}(x_1,x_2)\le u_1,\rho_{n,k,2}(x_1,x_2)\le u_2)\\&\qquad=\sum_{(l_1,l_2)\in J(n,k)}\binom{n}{l_1}\binom{n-l_1}{l_2}p_{x_1}^{l_1}(u_1)p_{x_2}^{l_2}(u_2)(1-p_{x_1}(u_1)-p_{x_2}(u_2)))^{n-l_1-l_2}\end{aligned}$$

where $J(n,k):=\{(l_1,l_2):l_1\ge k,l_2\ge k,l_1+l_2\le n-2\}$. This set is nonempty as $(k,k)\in J(n,k)$.

Consequently, we get a polynome in $p_{x_1}(u_1)$ and $p_{x_2}(u_2)$. In the proof of Lemma 3.1 we have seen that, for $x\in\mathbb{R}^d$, $u>0$ and $l\in\mathbb{N}$, the distribution function $p_x(u)^l$ has a density $lp_x^{l-1}(u)f_x(u)$ where $f_x(\cdot)$ is a density of $p_x(\cdot)$ w.r.t. $\Lambda_{\mathbb{R}_+}$. Thus $p_{x_1}^{l_1}(u_1)p_{x_2}^{l_2}(u_2)$ is absolutely continuous w.r.t. $\Lambda_{\mathbb{R}_+}\otimes\Lambda_{\mathbb{R}_+}$. Hence there is an integrable (w.r.t. the restriction of $\Lambda\otimes\Lambda$ on $B_{x_1,x_2}(\varepsilon)$) function $f_{n,k}(x_1,x_2,\cdot,\cdot)\colon B_{x_1,x_2}(\varepsilon)\to\mathbb{R}$ such that, for a set $(0,u_1]\times(0,u_2]$, $(u_1,u_2)\in B_{x_1,x_2}(\varepsilon)$, formula (A.14) takes place. The additivity of the integral implies the validity of the mentioned formula for any parallelepiped $(a_1,b_1]\times(a_2,b_2]$ where $a_i\le b_i$ $(i=1,2)$, $(a_1,a_2),(b_1,b_2)\in B_{x_1,x_2}(\varepsilon)$. Moreover, formula (A.14) holds for an algebra $\mathcal{E}$ of subsets of $B_{x_1,x_2}(\varepsilon)$ which can be represented as a finite union of such parallelepipeds. Thus we have seen that

$$Q(B):=\int_B f_{n,k}(x_1,x_2,u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2,\quad B\in\mathcal{B}(\mathbb{R}^2)\cap B_{x_1,x_2}(\varepsilon),$$

and $G(B):=\mathsf{P}(\zeta\in B)$ coincides on $\mathcal{E}$. In a similar way to the proof of Lemma 3.1 we get that $f_{n,k}(x_1,x_2,u_1,u_2)\ge0$ for $(\Lambda\otimes\Lambda)$-almost all $u=(u_1,u_2)\in B_{x_1,x_2}(\varepsilon)$. Therefore, the desired formula (A.14) is established.

Compare (A.12) and (A.13). We show now that, for $(t_1,t_2)\in B_{x_1,x_2}(\varepsilon)$ and all $\delta>0$ small enough (*i.e.* for $\delta<\Delta(x_1,x_2,t_1,t_2)$), one has $\mathsf{E}(\mathbb{I}\{\zeta\in C(t,\delta)\}\tau)=\mathsf{E}(\mathbb{I}\{\zeta\in\widetilde{B}(t,\delta)\tau)$ and $\mathsf{P}(\zeta\in C(t,\delta))=\mathsf{P}(\zeta\in\widetilde{B}(t,\delta))$ where $C(t,\delta):=(t_1-\delta,t_1+\delta]\times(t_2-\delta,t_2+\delta]$. Using the relation $C(t,\delta)\subset\widetilde{B}(t,\delta)\subset B_{x_1,x_2}(\varepsilon)$ for all $\delta\in(0,\Delta(x_1,x_2,t_1,t_2))$ and due to (A.14), we get $\mathsf{P}(\zeta\in\widetilde{B}(t,\delta)\setminus C(t,\delta))=0$ as $(\Lambda\otimes\Lambda)(\widetilde{B}(t,\delta)\setminus C(t,\delta))=0$. Taking into account that the restriction of P_ζ to $B_{x_1,x_2}(\varepsilon)$ is absolutely continuous w.r.t. the corresponding restriction of $\Lambda_{\mathbb{R}_+} \otimes \Lambda_{\mathbb{R}_+}$ we can claim that, for each $x_1, x_2 \in \mathbb{R}^d$ $(x_1 \neq x_2)$ and for P_ζ-almost all $(t_1, t_2) \in B_{x_1,x_2}(\varepsilon)$, formulas (3.3) and (3.4) are established.

Note that $B_{x_1,x_2}(\varepsilon) \uparrow (0, |x_1 - x_2|/2) \times (0, |x_1 - x_2|/2)$ as $\varepsilon \to 0$. Consequently, formulas (3.3) and (3.4) are valid P_ζ-a.s. for points of the set $(0, |x_1 - x_2|/2) \times (0, |x_1 - x_2|/2)$. □

A.3 Proof of Lemma 3.4

One has

$$\mathsf{E}(W\mathbb{I}\{V \in B\}) = \mathsf{E}(\mathsf{E}(W\mathbb{I}\{V \in B\})|V) = \mathsf{E}(\mathbb{I}\{V \in B\}\mathsf{E}(W|V))$$

$$= \int_{\mathbb{R}^m} \mathbb{I}\{t \in B\}\mathsf{E}(W|V = t)\mathsf{P}_V(\mathrm{d}t) = \int_B \mathsf{E}(W|V = t)\mathsf{P}_V(\mathrm{d}t).$$

Consequently,

$$\mathsf{E}(W|V \in B) = \frac{\mathsf{E}(W\mathbb{I}\{V \in B\})}{\mathsf{P}(V \in B)} = \frac{\int_B \mathsf{E}(W|V = t)\mathsf{P}_V(\mathrm{d}t)}{\mathsf{P}(V \in B)} = \int_B \mathsf{E}(W|V = t)\widetilde{\mathsf{P}}_{V,B}(\mathrm{d}t)$$

where $\widetilde{\mathsf{P}}_{V,B}(D) = \frac{\mathsf{P}(D\cap B)}{\mathsf{P}(B)}$, $D \in \mathcal{B}(\mathbb{R}^m)$. □

A.4 Proof of Lemma 3.5

It is enough to demonstrate that, for any set $B \in \mathcal{B}(\mathbb{R})$,

$$\int_B \mathsf{E}(\xi|\eta = t)\, \mathsf{P}_\eta(\mathrm{d}t) = \sum_{r\in S} \int_B \mathsf{E}(\xi|\zeta = r, \eta = t)\mathsf{P}(\zeta = r|\eta = t)\, \mathsf{P}_\eta(\mathrm{d}t).$$

Clearly,

$$\int_B \mathsf{E}(\xi|\eta = t)\, \mathsf{P}_\eta(\mathrm{d}t) = \mathsf{E}\xi\mathbb{I}(\eta \in B) = \sum_{r\in S} \mathsf{E}\xi\mathbb{I}(\eta \in B, \zeta = r)$$

$$= \sum_{r\in S} \int_{B\times\{r\}} \mathsf{E}(\xi|\eta = t, \zeta = v)\mathrm{d}\mathsf{P}_{\eta,\zeta}(t, v)$$

where $\mathrm{d}\mathsf{P}_{\eta,\zeta}(t, r)$ means the integration w.r.t. measure $\mathsf{P}(\eta \in \cdot, \zeta \in \cdot)$. Now we show that, for a measurable function $\varphi\colon \mathbb{R} \times S \to \mathbb{R}$ such that $\mathsf{E}|\varphi(\eta, \zeta)| < \infty$, the following relation holds

$$\int_{B\times\{r\}} \varphi(t, v)\mathrm{d}\mathsf{P}_{\eta,\zeta}(t, v) = \int_B \varphi(t, r)\mathsf{P}(\zeta = r|\eta = t)\mathsf{P}_\eta(\mathrm{d}t). \tag{A.15}$$

Indeed, for $A \in \mathcal{B}(\mathbb{R})$ and $s \in S$, consider $\varphi(t, v) := \mathbb{I}\{t \in A, v = s\}$, $t \in \mathbb{R}$ and $v \in S$. Obviously, if $s \neq r$ then (A.15) is true. If $s = r$ then

$$\begin{aligned}\int_{B\times\{r\}} \varphi(t,v)d\mathsf{P}_{\eta,\zeta}(t,v) &= \mathsf{P}_{\eta,\zeta}((A\cap B)\times\{r\}) = \mathsf{P}(\zeta=r,\eta\in A\cap B)\\ &= \int_{A\cap B}\mathsf{P}(\zeta=r|\eta=t)\mathsf{P}_\eta(\mathrm{d}t) = \int_B \mathbb{I}(t\in A, r=r)\mathsf{P}(\zeta=r|\eta=t)\mathsf{P}_\eta(\mathrm{d}t)\\ &= \int_B \varphi(t,r)\mathsf{P}(\zeta=r|\eta=t)\mathsf{P}_\eta(\mathrm{d}t).\end{aligned}$$

Hence (A.15) is valid for $\varphi(t,v)=\mathbb{I}\{t\in A, v\in E\}$ where one can take arbitrary $A\in\mathcal{B}(\mathbb{R})$ and $E\subset S$. Taking into account that any measurable function $\varphi:\mathbb{R}\times S\to\mathbb{R}$ can be approximated by finite linear combinations of the considered functions of the type $\mathbb{I}\{A\}\mathbb{I}\{E\}$ we come to desired statement (A.15). We also note that $\mathsf{E}|\mathsf{E}(\xi|\eta,\zeta)|\leq \mathsf{E}|\xi|<\infty$. □

The applications of obtained results to the feature selection involving the mutual information estimation will be provided in the forthcoming paper.

Acknowledgements. The work is supported by the Russian Science Foundation under grant 14-21-00162 and performed at the Steklov Mathematical Institute of Russian Academy of Sciences. The authors are grateful to Professor V.I.Bogachev for the helpful discussions. We also thank the Associate Editor and the Referee for valuable remarks and suggestions.

References

[1] P. Alonso-Ruiz and E. Spodarev, Entropy-based inhomogeneity detection in porous media. Preprint arXiv:1611.02241 (2016).
[2] E. Archer, I.M. Park and J.W. Pillow, Bayesian entropy estimation for countable discrete distributions. *J. Mach. Learn. Res.* **15** (2014) 2833–2868.
[3] L. Benguigui, The different paths to entropy. *Eur. J. Phys.* **34** (2013) 303–321.
[4] M. Bennasar, Y. Hicks and R. Setchi, Feature selection using joint mutual information maximisation. *Exp. Syst. Appl.* **42** (2014) 8520–8532.
[5] T.B. Berrett, R.J. Samworth and M. Yuan, Efficient multivariate entropy estimation via k-nearest neighbour distances. *J. Reine Angew. Math.* **673** (2012) 1–31.
[6] G. Biau and L. Devroye, Lectures of the Nearest Neighbor Method. Springer, Cham (2015).
[7] J. Beirlant. E.J. Dudewicz, L. Györfi and E.C. van der Meulen, Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.* **6** (1997) 17–39.
[8] V.I. Bogachev, Measure Theory. Springer-Verlag, Berlin (2007).
[9] V.S. Borkar, Probability Theory. An Advanced Course. Springer-Verlag, New York (1995).
[10] A. Bulinski and D. Dimitrov, Statistical estimation of the Shannon entropy. *Acta Math. Sin.* **35** (2019) 17–46.
[11] A. Charzyńska and A. Gambin, Improvement of the k-NN entropy estimator with applications in systems biology. *Entropy* **18** (2016) 13.
[12] F. Coelho, A.P. Braga and M. Verleysen, A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems. *Int. J. Comput. Intell. Syst.* **9** (2016) 726–733.
[13] T.M. Cover and J.A. Thomas, Elements of Information Theory, 2nd ed. Wiley–Interscience, New York (1991).
[14] S. Delattre and N. Fournier, On the Kozachenko–Leonenko entropy estimator. *J. Stat. Plan. Infer.* **185** (2017) 69–93.
[15] G. Doquire and M. Verleysen, A comparison of mutual information estimators for feature selection, in *Proc. of the 1st International Conference on Pattern Recognition Applications and Methods* (2012) 176–185.
[16] D. Evans, A computationally efficient estimator for mutual information. *Proc. R. Soc. Lond. Ser. A* **464** (2008) 1203–1215.
[17] F. Fleuret, Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **4** (2004) 1531–1555.
[18] W. Gao, S. Kannan, S. Oh and P. Viswanath, Estimating Mutual Information for Discrete-Continuous Mixtures. Preprint arXiv:1709.06212v2 (2017).
[19] P. Hall and S.C. Morton, On the estimation of entropy. *Ann. Inst. Stat. Math.* **45** (1993) 69–88.
[20] Y. Han, J. Jiao, T. Wissman and Y. Wu, Optimal rates of entropy estimation over Lipschitz balls. Preprint arXiv:1711.02141 (2017).
[21] J.M. Hilbe, Practical Guide to Logistic Regression. CRC Press, Boca Raton (2015).
[22] J. Jiao, J.W. Gao and Y. Han, The nearest neighbor information estimator is adaptively near minimax rate-optimal. Preprint arXiv:1711.08824v3 (2017).
[23] D.G. Kleinbaum and M. Klein, Logistic Regression. A Self-Learning Text, 3rd ed. with contributions by E.R.Pryor. Springer, New York (2010).
[24] L.F. Kozachenko and N.N. Leonenko, Sample estimate of the entropy of a random vector. *Probl. Inf. Trans.* **23** (1987) 95–101.
[25] A. Kraskov, H. Stögbauer and P. Grassberger, Estimating mutual information. *Phys. Rev. E* **69** (2004) 066138.
[26] L. Massaron and A. Boschetti, Regression Analysis with Python. Packt Publishing Ltd., Birmingham (2016).

[27] P. Massart, Concentration inequalities and model selection, in *École d'Été de Probabilités de Saint-Flour XXXIII – 2003.* Springer–Verlag, Berlin (2007).

[28] E.G. Miller, A new class of entropy estimators for multidimensional densities, in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP03)*, Hong Kong, China (April 06–10, 2003) 297–300.

[29] J. Montalvão, R. Attux and D. Silva, A pragmatic entropy and differential entropy estimator for small datasets. *J. Commun. Inf. Syst.* **29** (2014) 29–36.

[30] I. Muqattash and M. Yahdi, Infinite family of approximations of the Digamma function. *Math. Comput. Model.* **43** (2006) 1329–1336.

[31] C. Nair, B. Prabhakar and D. Shah, On entropy for mixtures of discrete and continuous variables. Preprint arXiv:cs/0607075v2 (2007).

[32] J. Novovičová, P. Somol and P. Pudil, Conditional mutual information based feature selection for classification task, in *CIARP 2007,* in Vol. 4756 of *Lect. Notes Comput. Sci.*, edited by L. Rueda, D. Mery, J. Kittler. Springer-Verlag, Berlin, Heidelberg (2007) 417–426.

[33] D. Pál, B. Póczos and C. Szepesvári, Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs, in *Proc. of the 23rd International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada* (2010) 1849–1857.

[34] L. Paninski, Estimation of entropy and mutual information. *Neural Comput.* **15** (2003) 1191–1253.

[35] H. Peng, F. Long and C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27** (2005) 1226–1238.

[36] M.D. Penrose and J.E. Yukich, Limit theory for point processes in manifolds. *Ann. Appl. Prob.* **23** (2013) 2161–2211.

[37] I. Sason and S. Verdú, F-divergence inequalities. *IEEE Trans. Inf. Theory* **62** (2016) 5973–6006.

[38] C.E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27** (1948) 379–423; 623–656.

[39] A.N. Shiryaev, Probability – 1, 3rd ed. Springer, New York (2016).

[40] S. Singh and B. Pószoc, Analysis of k-nearest neighbor distances with application to entropy estimation. Preprint arXiv:1603.08578 (2016).

[41] K. Sricharan, D. Wei and A.O. Hero, Ensemble estimators for multivariate entropy estimation. *IEEE Trans. Inf. Theory* **59** (2013) 4374–4388.

[42] D. Stowell and M.D. Plumbley, Fast multidimensional entropy estimation by k-d partitioning. *IEEE Signal Process. Lett.* **16** (2009) 537–540.

[43] A.B. Tsybakov and E.C. van der Meulen, Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Stat.* **23** (1996) 75-83.

[44] J.R. Vergara and P.A. Estévez, A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **24** (2014) 175–186.

[45] J. Yeh, Real Analysis: Theory of Measure and Integration, 3rd ed. World Scientific, Singapore (2014).

[46] A.M. Zubkov and A.A. Serov, A complete proof of universal inequalities for distribution function of binomial law. *Theory Probab. Appl.* **57** (2013) 539–544.