Noname manuscript No.
(will be inserted by the editor)

Consistency of the k-nearest neighbor classifier for spatially dependent data

Ahmad Younso · Ziad Kanaya · Nour Azhari

the date of receipt and acceptance should be inserted later

Abstract The purpose of this paper is to investigate the k-nearest neighbor classification rule for spatially dependent data. Some spatial mixing conditions are considered, and under such spatial structures, the well known k-nearest neighbor rule is suggested to classify spatial data. We established consistency and strong consistency of the classifier under mild assumptions. Our main results extend the consistency result in the i.i.d. case to the spatial case.

Keywords Bayes rule · spatial data · training data · k-nearest neighbor rule · mixing condition · consistency.

Mathematics Subject Classification (2010) 62H11 · 62G08 · 62G20

1 Introduction

Analysis of spatial data arises in various areas of research including agricultural field trials, astronomy, econometrics, epidemiology, environmental science, geology, hydrology, image analysis, meteorology, ecology, oceanography and many others in which the data of interest are collected across space. One of the most fundamental issues in spatial analysis is classification and pattern

---

A. Younso
(a) MISTEA, Université Montpellier, INRAE, Institut Agro, Montpellier, France. Tel.: +33-760666251
E-mail: ahmad.younso@inrae.fr
(b) Department of Mathematical Statistics, Faculty of Sciences, Damascus University, Damascus, Syria

Z. Kanaya
Department of Mathematics, Faculty of Sciences, Tishreen University, Lattakia, Syria E-mail: ziad.kanaya@tishreen.edu.sy

N. Azhari
Department of Mathematics, Faculty of Sciences, Tishreen University, Lattakia, Syria E-mail: n.azhari@tishreen.edu.sy

recognition. For example, in remote sensing technology or digital geography information, we need somehow to classify spatial data into patterns or images into types. Recently, [22] propose a novel probabilistic model for classification, that incorporates a network's structure into the classical logistic regression model. This model is mostly used to classify data produced by social network analysis taking into account the connection between nodes, but without any influence of the spatial coordinates. [18,17,19,20,21] deal with kernel-based rules to classify temporally and spatially dependent data, and study asymptotic properties of classifiers. The aim of the present paper is to investigate whether the classical k-nearest neighbor classifier can be extended to classify spatial data. To the best of our knowledge this work is the first one dealing with spatial data. The k-nearest neighbor method for estimating density and regression or data classification has been widely used and studied for many years in the i.i.d. case. Key references on this topic are: [4], [3], [5], [6] and [2]. The use of the k-nearest neighbor method in the spatial case is due to [15] for density estimation. The real interest in the k-nearest neighbor method comes from the nature of the smoothing parameter. Indeed, in the traditional kernel method, the smoothing parameter is the bandwidth, which is a real positive number. Here, the number of neighbors k is the smoothing parameter and it takes its values in a discrete set. As we said previously, the other very important aspect of this method is that it allows the construction of a neighborhood adapted to the local structure of the data. The main difficulties with the kernel method appear when data are sparse; choosing the number of neighbors allows to avoid this problem and is adapted to the concentration of the data. Consistency of kernel-based rules on temporally or spatially dependent data has recently been investigated by [18,17,19,20,21] in finite and infinite-dimensional space. In this paper, we will establish the (strong) consistency of the k-nearest neighbor classifier for spatially dependent data. Let $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ be a random field defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathbb{R}^d \times \{0,1\}$. In the problem of classification, for each $\mathbf{i} \in \mathbb{Z}^N$, $X_{\mathbf{i}}$ is a vector of features and $Y_{\mathbf{i}}$ is the label (class) of $X_{\mathbf{i}}$. A point $\mathbf{i} = (i_1, ..., i_N) \in \mathbb{Z}^N$ will be referred to as a site. For $\mathbf{n} = (n_1, ..., n_N) \in (\mathbb{N}^*)^N$, we define the rectangular region $\mathcal{I}_{\mathbf{n}}$ by $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{Z}^N : \ 1 \le i_l \le n_l, \ \forall l = 1, ..., N\}$. We will write $\mathbf{n} \to \infty$ if $\min_{1\le l\le N} n_l \to \infty$. Define $\hat{\mathbf{n}} = n_1 \times ... \times n_N = \operatorname{card}(\mathcal{I}_{\mathbf{n}})$. We wish to predict the label $Y_{\mathbf{j}}$ of a new observation $X_{\mathbf{j}}$. The pair $(X_{\mathbf{j}}, Y_{\mathbf{j}})$ may be described by μ, the probability measure for $X_{\mathbf{j}}$, and $\eta(x) = \mathbb{E}(Y_{\mathbf{j}}/X_{\mathbf{j}} = x)$, the regression of $Y_{\mathbf{j}}$ on $X_{\mathbf{j}} = x$. Assume that for each $\mathbf{i} \in \mathbb{Z}^N$, $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ has the same distribution as the pair (X, Y). We create a classifier $g : \mathbb{R}^d \longrightarrow \{0,1\}$ mapping $X_{\mathbf{j}}$ into the predicted label of $X_{\mathbf{j}}$. The error rate, or risk, of a rule g is $L(g) = \mathbb{P}\{g(X_{\mathbf{j}}) \ne Y_{\mathbf{j}}\}$. This is minimized by the rule

$$g^*(x) = \begin{cases} 0 \text{ if } \mathbb{P}\{Y_{\mathbf{j}} = 0 | X_{\mathbf{j}} = x\} \ge \mathbb{P}\{Y_{\mathbf{j}} = 1 | X_{\mathbf{j}} = x\} \\ 1 \text{ otherwise,} \end{cases}$$

whose error rate $L^* = L(g^*)$ is called the Bayes risk and g^* is called the Bayes rule. This optimal rule depends on the distribution of $(X_{\mathbf{j}}, Y_{\mathbf{j}})$ which is generally unknown. we use the data $D_{\mathbf{n}} = \{(X_{\mathbf{i}}, Y_{\mathbf{i}}) : \mathbf{i} \in \mathcal{I}_{\mathbf{n}}\}$ to construct a classifier $g_{\mathbf{n}}(x)$. The set $D_{\mathbf{n}}$ is called training sample. The spatial version of the classical k-nearest neighbor rule given by

$$g_{\mathbf{n}}(x) = \begin{cases} 0 \text{ if } \sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} w_{\mathbf{ni}} Y_{\mathbf{i}} \leq 1/2 \\ 1 \text{ otherwise,} \end{cases} \tag{1}$$

where $w_{\mathbf{ni}} = w_{\mathbf{ni}}(x; D_{\mathbf{n}})$ is $1/k$ if $X_{\mathbf{i}}$ is one of the k-nearest neighbor of x in $D_{\mathbf{n}}$ and $w_{\mathbf{ni}}$ is zero otherwise with $k = k(\mathbf{n})$ is a sequence of positive integers satisfying

$$k \longrightarrow \infty \quad \text{and} \quad k/\hat{\mathbf{n}} \longrightarrow 0 \ as\ \mathbf{n} \to \infty. \tag{2}$$

Observe that the distance between two observations in $\mathbb{R}^d$ or two sites in $\mathbb{Z}^N$ will be computed by the Euclidean distance. We assume that μ is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^d$, in other words, X has a density f with respect to λ, so that we can avoid messy technicalities necessary to handle distance ties. If we let $\eta_{\mathbf{n}}(x) = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} w_{\mathbf{ni}} Y_{\mathbf{i}}$ be the k-nearest neighbor estimator of $\eta(x)$, (1.1) can be re-written as follows

$$g_{\mathbf{n}}(x) = \begin{cases} 0 \text{ if } \eta_{\mathbf{n}}(x) \leq 1/2 \\ 1 \text{ otherwise.} \end{cases} \tag{3}$$

The best we can expect from $g_{\mathbf{n}}(x)$ is to achieve the Bayes risk. Denote $L_{\mathbf{n}} = L(g_{\mathbf{n}})$ the error rate of $g_{\mathbf{n}}$. The classifier $g_{\mathbf{n}}(x)$ is called consistent if $\mathbb{E}L_{\mathbf{n}} \longrightarrow L^*$ as $\mathbf{n} \to \infty$ and it is strongly consistent if $L_{\mathbf{n}} \longrightarrow L^*$ as $\mathbf{n} \to \infty$ with probability one. In this paper, we investigate both the consistency and strong consistency of $g_{\mathbf{n}}$ under classical conditions.

2 Mixing conditions

Let us first recall the definitions of mixing coefficients α introduced by [14] and β introduced by [13]. Let $\mathcal{A}$ and $\mathcal{C}$ be two sub σ-algebras of $\mathcal{F}$. The α-mixing coefficient between $\mathcal{A}$ and $\mathcal{C}$ is defined by

$$\alpha = \alpha(\mathcal{A}, \mathcal{C}) = \sup_{A\in\mathcal{A}, C\in\mathcal{C}} |\mathbb{P}(A \cap C) - \mathbb{P}(A)\mathbb{P}(C)|$$

and the β-mixing coefficient is defined by

$$\beta = \beta(\mathcal{A}, \mathcal{C}) = \mathbb{E} \sup_{A\in\mathcal{A}} |\mathbb{P}(A|\mathcal{C}) - \mathbb{P}(A)]|.$$

Let $\{Z_{\mathbf{i}}\}_{\mathbf{i}\in\mathbb{Z}^N}$ be a random field on $(\Omega, \mathcal{F}, \mathbb{P})$ and take values in some space $(\Omega', \mathcal{F}')$. For any $E, E' \subset \mathbb{Z}^N$ with finite cardinals, we denote by $\mathcal{B}(E)$ and $\mathcal{B}(E')$ the Borel σ-algebras generated by $\{Z_{\mathbf{i}}\}_{\mathbf{i}\in E}$ and $\{Z_{\mathbf{i}}\}_{\mathbf{i}\in E'}$ respectively. The random field $\{Z_{\mathbf{i}}\}_{\mathbf{i}\in\mathbb{Z}^N}$ is said to be α-mixing (strongly mixing) if

$$\alpha(t) = \sup_{\operatorname{dist}(E,E')\geq t} \alpha\big(\mathcal{B}(E), \mathcal{B}(E')\big) \downarrow 0 \text{ as } t \to \infty,$$

where

$$\operatorname{dist}(E,E') = \inf_{\mathbf{i}\in E,\mathbf{j}\in E'} \|\mathbf{i}-\mathbf{j}\|$$

and $\|.\|$ denotes the Euclidean norm. The above α-mixing condition may be satisfied by many spatial models and examples can be found in [9] and [12]. The random field $\{Z_{\mathbf{i}}\}_{\mathbf{i}\in\mathbb{Z}^N}$ is said to be β-mixing (absolutely regular) if

$$\beta(t) = \sup_{\operatorname{dist}(E,E')\geq t} \beta\big(\mathcal{B}(E),\mathcal{B}(E')\big) \downarrow 0 \text{ as } t\to\infty.$$

The two mixing coefficients α and β are related by the inequality $2\alpha\leq\beta$ (see [10]). Consequently, any β-mixing random field is α-mixing one. Throughout the paper, it will be assumed that the random field $\{(X_{\mathbf{i}},Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ is strongly mixing (absolutely regular) to establish consistency (strong consistency) of the k-nearest neighbor rule. However, the training sample $D_{\mathbf{n}}$ is obtained by observing the feature vector $X_{\mathbf{i}}$ with its label $Y_{\mathbf{i}}$ in each site $\mathbf{i}\in\mathcal{I}_{\mathbf{n}}$. Each one of the above mixing coefficient describes a spatial interdependence between the observations $(X_{\mathbf{i}},Y_{\mathbf{i}})$ based on their locations on the lattice points. According to the above mixing conditions, observations in sites that are close together tend to be more correlated than that are in sites being far apart.

3 Preliminary lemmas and main results

The following lemmas will be needed to establish consistency and strong consistency. The proof of the following lemma is found in [10].

Lemma 1 *Let Z_1 and Z_2 be two $\mathbb{R}$-valued bounded random variables. Then, we have*

$$|\operatorname{cov}(Z_1,Z_2)| \leq 4\|Z_1\|_\infty\|Z_2\|_\infty\alpha(\sigma(Z_1),\sigma(Z_2)),$$

where $\|.\|_\infty$ is the supremum norm and $\sigma(Z_i)$ is the σ-algebra generated by Z_i for $i=1,2$.

Let $\mathcal{A}$ and $\mathcal{C}$ be two sub σ-algebras of $\mathcal{F}$, we denote by $\mathcal{A}\vee\mathcal{C}$ the σ-algebra generated by $\mathcal{A}\cup\mathcal{C}$. The following lemmas will be used to prove strong consistency of the classifier.

Lemma 2 *Let Z be a random variable defined on $(\Omega,\mathcal{F},\mathbb{P})$ and taking values in some Polish space Ω' and $\mathcal{M}$ be a sub σ-algebra of $\mathcal{F}$. Assume that there exists a random variable U uniformly distributed over $[0,1]$, independent of $\sigma(Z)\vee\mathcal{M}$. Then, there exists a random variable Z^* measurable with respect to $\sigma(U)\vee\sigma(Z)\vee\mathcal{M}$, distributed as Z and independent of $\mathcal{M}$, such that*

$$\mathbb{P}(Z\neq Z^*) = \beta(\mathcal{M},\sigma(Z)).$$

For the proof of Lemma 2 (Berbee's lemma), we refer the reader to [1]. Denote $S_{x,r}$ the closed ball centered at x with radius $r>0$.

Lemma 3 *Let*

$$B_a(x') = \{x \in \mathbb{R}^d : \mu\left(S_{x,\|x-x'\|}\right) < a\}.$$

Then, for all $x' \in \mathbb{R}^d$,

$$\mu\left(B_a(x')\right) \leq \gamma_d a$$

with γ_d *is the minimal number of cones centered at the origin of angle* $\pi/6$ *that cover* $\mathbb{R}^d$.

We refer the reader to [6] for the proof of Lemma 3. The number γ_d defined in Lemma 3 exists according to ([6], Lemma 5.5). Now, we state the main results of this paper. In the following theorem, we investigate consistency of the k-nearest neighbor rule.

Theorem 1 *Suppose that* D_n *are observations of* α*-mixing random field such that* $\alpha(t) = O(t^{-\theta})$ *with* $\theta > N$. *Suppose in addition that (1.2) is satisfied and that as* $\mathbf{n} \to \infty$,

$$k/\sqrt{\hat{\mathbf{n}}} \longrightarrow \infty. \tag{4}$$

Then, as $\mathbf{n} \to \infty$,

$$\mathbb{E}L_{\mathbf{n}} \longrightarrow L^*.$$

Theorem 1 extends Stone's consistency theorem (see [11]) to the spatial case when the probability measure μ is absolutely continuous under a slight modification of Stone's condition on the smoothing parameter k. Condition (3.1) is weaker than that used by ([3], Theorem II.3) in the i.i.d. case (see also ([4], theorem 1)). In the following theorem, we investigate strong consistency of the k-nearest neighbor rule.

Theorem 2 *Suppose that* D_n *are observations of strictly stationary* β*-mixing random field such that* $\beta(t) = O(t^{-\theta})$ *with* $\theta > N$ *and that (1.2) and (3.1) are satisfied. Suppose in addition that there is an integer* $p = p(\mathbf{n})$ *with* $p(\mathbf{n}) \in [1, \min_{1 \leq l \leq N} n_l/2]$ *such that as* $\mathbf{n} \to \infty$,

$$\frac{\hat{\mathbf{n}}}{p^N \log \hat{\mathbf{n}}} \longrightarrow \infty \tag{5}$$

and

$$\sum_{\hat{\mathbf{n}} \in (\mathbb{N}^*)^N} k^{-1} \hat{\mathbf{n}} \beta(p) < \infty. \tag{6}$$

Then, as $\mathbf{n} \to \infty$,

$$L_{\mathbf{n}} \longrightarrow L^* \quad \text{with probability one.}$$

Theorem 2 extends the strong consistency of ([6], Theorem 11.1) to the spatial case under some mild additional condition on the smoothing parameter k. Observe that if $\beta(t) = O(t^{-\theta})$, then $\alpha(t) = O(t^{-\theta})$ since $2\alpha(t) \leq \beta(t)$, so that $\alpha(t)$ and $\beta(t)$ tend to zero as $t \to \infty$ with polynomial rate. In addition, if we take for example $p = \hat{\mathbf{n}}^{1/2N}$, (5) and (6) are satisfied for some $\theta > 4N$.

4 Numerical results

In this section, some numerical results are proposed towards some simulations. We consider a two-dimensional space ($N = 2$) with the random field

$$\left\{(X_{(i,j)}, Y_{(i,j)}), (i, j) \in \mathbb{Z}^2\right\}$$

simulated on a rectangular region

$$\mathcal{I}_{(n_1,n_2)} = \{(i, j), 1 \leq i \leq n_1, 1 \leq j \leq n_2\}$$

of $n_1 \times n_2$ sites. Without loss of generality, we take $n_1 = n_2 = n$. We focus on the case where $(X_{(i,j)}, Y_{(i,j)})$ takes values in $\mathbb{R}^2 \times \{0, 1\}$ with $X_{(i,j)} = (X_{1,(i,j)}, X_{2,(i,j)})$ where $X_{1,(i,j)}$ are dependent normal variables with mean 0, variance 0.5 and covariance function $c(u) = 0.5 \exp(-\|u\|)$ for all $u \in \mathbb{R}^2$ with $u \neq 0$, and $X_{2,(i,j)}$ are independent normal variables with mean 0 and variance 0.5. We let $Y_{(i,j)} = 1$ if $\sin(X_{1,(i,j)} - X_{2,(i,j)}) > \sin(X_{1,(i,j)} + X_{2,(i,j)})$ and $Y_{(i,j)} = 0$ otherwise. The R statistical programming environment is used to run simulations. First of all, we give a typical example by using the above scenario for $n = 25$ and we get the following figure.

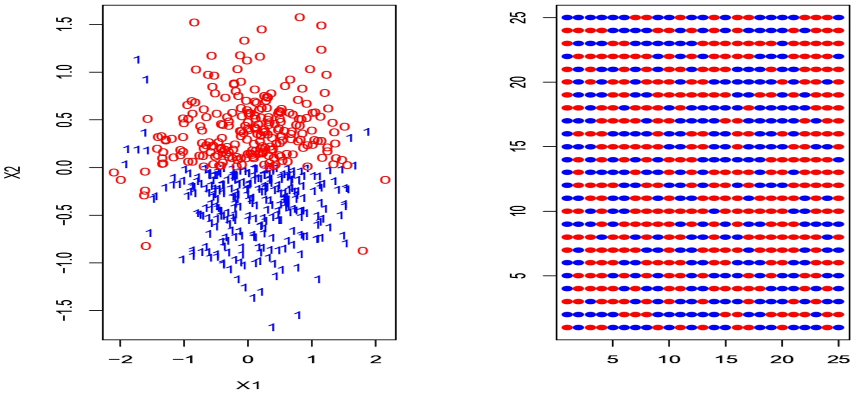

Fig. 1 Labeled feature vectors at left-hand side and labeled sites at right-hand side with red color for the class (0) and blue color for the class (1).

Figure 1 shows the labels of 625 feature vectors $X_{(i,j)}$ with their labeled sites on the region $\mathcal{I}_{(20,20)}$. Now, for each $n \in \{20, 30, 40, 50\}$, we simulate a sample of size n^2 on the rectangular region $\mathcal{I}_{(n,n)}$. Then, each sample is splitted into two sets. The first set contains $n^2 - 100$ elements of the sample for training and the other contains 100 elements of the samples for testing. Figure 2 displays the labeled samples for $n = 20, 30, 40, 50$.

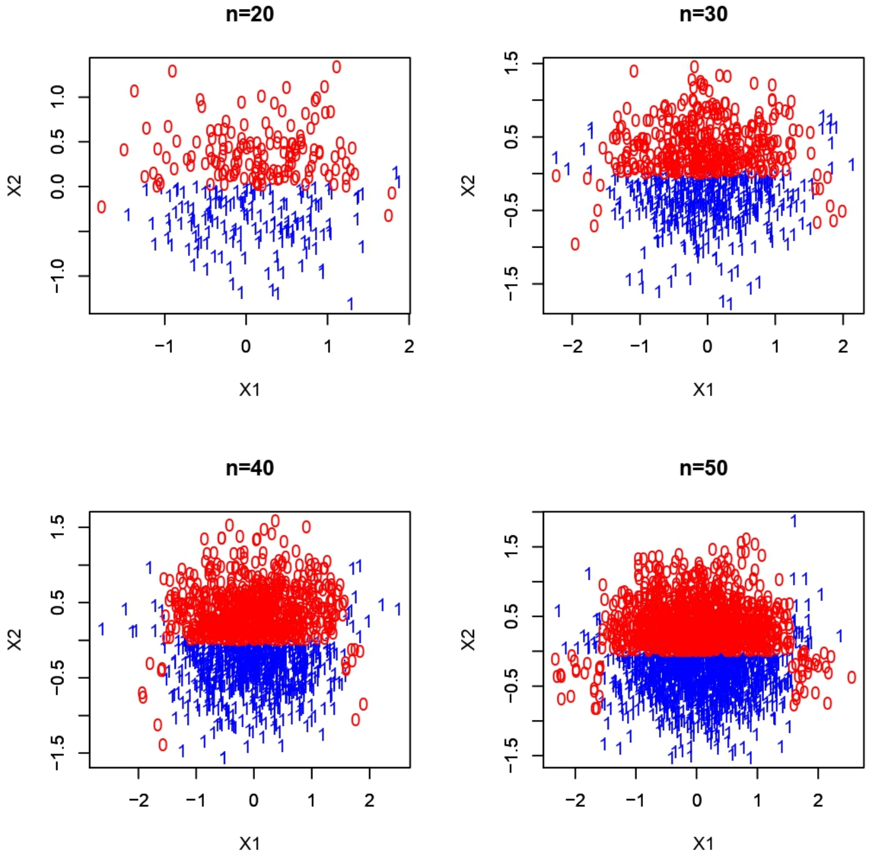

Fig. 2 Four labeled samples corresponding to $n = 20, 30, 40, 50$ with red color for the class (0) and blue color for the class (1).

We apply the cross-validation criterion (CV) to the training samples to choose values of the smoothing parameter k by altering k with various values and choose that corresponding to the lowest $CV(k)$ given by

$$CV(k) = \sum_{\mathbf{i}} (Y_{\mathbf{i}} - g^{-\mathbf{i}}_{(n,n)}(X_{\mathbf{i}}))^2,$$

where $g^{-\mathbf{i}}_{(n,n)}(X_{\mathbf{i}})$ indicates the k-nearest neighbor rule based on leaving out the pair $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ and the summation is taken over all sites of a training sample. It is desirable for k to be odd to make ties less likely. Then, for each n, we estimate the misclassification error rate (ER) using the associated test sample, $i.e$,

$$ER = \frac{1}{100} \sum_{\mathbf{i}} \mathbb{1}_{\{Y_{\mathbf{i}} \neq g_{(n,n)}(X_{\mathbf{i}})\}},$$

where the summation is taken over all sites of a test sample and $\mathbb{1}_A$ denotes the indicator of A. Table 1 includes the optimal chosen values of k together with the corresponding estimated misclassification error rates for one replication of each n. To check the robustness of the proposed classifier, the above simulation is replicated 100 times, and the average error rate (AER) is obtained by averaging the error rates associated with the corresponding 100 test samples of each value of n. We keep the chosen values of k listed in Table 1 for each replication. Finally, we get the following table of average misclassification error rates. Table 2 displays the average error rates corresponding to $n \in \{20, 30, 40, 50\}$. It shows that the AER decreases when the size of training sample increases which make the results of this simulation study in line with the theoretical results.

n	STS	k	ER
20	300	21	0.05
30	800	33	0.03
40	1500	41	0.03
50	2400	51	0.04

Table 1 Misclassification error rates

n	STS	k	AER
20	300	21	0.0495
30	800	33	0.0410
40	1500	41	0.0330
50	2400	51	0.0310

Table 2 Average misclassification error rates

5 Proofs

Define $\rho_{\mathbf{n}} = \rho_{\mathbf{n}}(x)$ as the solution of the equation

$$\frac{k}{\hat{\mathbf{n}}} = \mu(S_{x,\rho_{\mathbf{n}}}). \tag{7}$$

Note that the solution always exists since X has a density by assumption. Also define

$$\widehat{\eta}_{\mathbf{n}}(x) = \frac{1}{k} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} \mathbb{1}_{\{X_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}}.$$

Proof of Theorem 1 By Theorem 2.2 in [6], we have

$$L_{\mathbf{n}} - L^* \leq 2 \int_{\mathbb{R}^d} |\eta(x) - \eta_{\mathbf{n}}(x)| \mu(dx). \tag{8}$$

Hence, it suffices to prove that

$$\mathbb{E}\int_{\mathbb{R}^d}|\eta(x)-\eta_{\mathbf{n}}(x)|\mu(dx)\to 0. \tag{9}$$

But

$$|\eta(x)-\eta_{\mathbf{n}}(x)|\leq|\eta(x)-\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)|+|\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)-\eta_{\mathbf{n}}(x)|. \tag{10}$$

Clearly, by (7), condition (2) implies that $\rho_{\mathbf{n}}\to 0$ as $\mathbf{n}\to\infty$. By Lebesgue's density theorem together with (7), we have as $\mathbf{n}\to\infty$,

$$\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)=\frac{1}{\mu(S_{x,\rho_{\mathbf{n}}})}\int_{S_{x,\rho_{\mathbf{n}}}}\mathbb{E}\left(Y/X=x'\right)\mu(dx')\to\mathbb{E}\left(Y/X=x\right)=\eta(x)$$

for all x mod μ (μ-almost for all $x\in\mathbb{R}^d$). Since $|Y|\leq 1$, the dominated convergence theorem implies that as $\mathbf{n}\to\infty$,

$$\int_{\mathbb{R}^d}|\eta(x)-\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)|\mu(dx)\to 0. \tag{11}$$

Therefore, by (10)-(11), it suffices to prove that as $\mathbf{n}\to\infty$,

$$\mathbb{E}\int_{\mathbb{R}^d}|\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)-\eta_{\mathbf{n}}(x)|\mu(dx)\to 0. \tag{12}$$

We have the following inequality

$$\begin{aligned}&\mathbb{E}\int_{\mathbb{R}^d}|\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)-\eta_{\mathbf{n}}(x)|\mu(dx)\\&\leq\mathbb{E}\int_{\mathbb{R}^d}|\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)-\widehat{\eta}_{\mathbf{n}}(x))|\mu(dx)+\mathbb{E}\int_{\mathbb{R}^d}|\widehat{\eta}_{\mathbf{n}}(x)-\eta_{\mathbf{n}}(x)|\mu(dx).\end{aligned} \tag{13}$$

Thus, we prove that the two terms in the right-hand side of (13) tend to zero as $\mathbf{n}\to\infty$. For the first term, by Cauchy-Schwartz inequality, we get

$$\begin{aligned}&\mathbb{E}\int_{\mathbb{R}^d}|\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)-\widehat{\eta}_{\mathbf{n}}(x)|\mu(dx)\\&\leq\int_{\mathbb{R}^d}\sqrt{\mathbb{E}(\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)-\widehat{\eta}_{\mathbf{n}}(x))^2}\mu(dx)\\&=\int_{\mathbb{R}^d}\sqrt{\operatorname{var}(\widehat{\eta}_{\mathbf{n}}(x))}\mu(dx)\\&\leq\int_{\mathbb{R}^d}\sqrt{\frac{\hat{\mathbf{n}}}{k^2}\operatorname{var}\left(Y\mathbb{1}_{X\in S_{x,\rho_{\mathbf{n}}}}\right)+C_{\mathbf{n}}(x)}\mu(dx)\end{aligned} \tag{14}$$

with

$$C_{\mathbf{n}}(x)=\frac{1}{k^2}\sum_{\mathbf{i}\neq\mathbf{j}}\left|\operatorname{cov}(Y_{\mathbf{i}}\mathbb{1}_{X_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}},Y_{\mathbf{j}}\mathbb{1}_{X_{\mathbf{j}}\in S_{x,\rho_{\mathbf{n}}}})\right|.$$

On the one hand, we have by (7)

$$\frac{\hat{\mathbf{n}}}{k^2}\operatorname{var}\left(Y\mathbb{1}_{X\in S_{x,\rho_{\mathbf{n}}}}\right)\leq\frac{\hat{\mathbf{n}}}{k^2}\mathbb{E}(\mathbb{1}_{X\in S_{x,\rho_{\mathbf{n}}}})=\frac{\hat{\mathbf{n}}}{k^2}\mu(S_{x,\rho_{\mathbf{n}}})=\frac{1}{k}. \tag{15}$$

On the other hand, by Lemma 1, we have

$$\begin{aligned}
C_{\mathbf{n}}(x) &\le \frac{4}{k^2}\sum_{\mathbf{i}\neq\mathbf{j}}\alpha(\|\mathbf{i}-\mathbf{j}\|) \le \frac{4\hat{\mathbf{n}}}{k^2}\sum_{\|\mathbf{i}\|\ge 1}\alpha(\|\mathbf{i}\|)\\
&\le \frac{4\hat{\mathbf{n}}}{k^2}\sum_{i=1}^{\infty} i^{N-1}\alpha(i) \le \frac{C\hat{\mathbf{n}}}{k^2}\sum_{i=1}^{\infty} i^{N-1-\theta} \le \frac{C\hat{\mathbf{n}}}{k^2}\int_{1/2}^{\infty} u^{N-1-\theta}du
\end{aligned}$$

for some generic constant $C > 0$. Therefore

$$C_{\mathbf{n}}(x) \le \frac{C\hat{\mathbf{n}}}{k^2}\int_{1/2}^{\infty} u^{N-1-\theta}du \le \frac{C\hat{\mathbf{n}}}{k^2} \tag{16}$$

since $\int_{1/2}^{\infty} u^{N-1-\theta}du < \infty$ for $\theta > N$. By (4) and (14)-(16) together with the dominated convergence theorem, we get

$$\mathbb{E}\int_{\mathbb{R}^d} |\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x) - \widehat{\eta}_{\mathbf{n}}(x)|\mu(dx) \to 0. \tag{17}$$

It remains to prove that the second term in the right-hand side of (13) tends to zero as as $\mathbf{n} \to \infty$. To do that, let $X_{(k)}(x)$ be the k-nearest neighbor of x and denote $r_{\mathbf{n}} = r_{\mathbf{n}}(x) = \|X_{(k)}(x) - x\|$. Clearly

$$\begin{aligned}
|\widehat{\eta}_{\mathbf{n}}(x) - \eta_{\mathbf{n}}(x)| &= \left|\frac{1}{k}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}}\mathbb{1}_{\{X_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} - \frac{1}{k}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}}\mathbb{1}_{\{X_{\mathbf{i}}\in S_{x,r_{\mathbf{n}}}\}}\right|\\
&\le \frac{1}{k}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\left|\mathbb{1}_{\{X_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} - \mathbb{1}_{\{X_{\mathbf{i}}\in S_{x,r_{\mathbf{n}}}\}}\right|\\
&\le \left|\frac{1}{k}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{1}_{\{X_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} - 1\right| = |\bar{\eta}_{\mathbf{n}}(x) - \mathbb{E}\bar{\eta}_{\mathbf{n}}(x)| \qquad (18)
\end{aligned}$$

with

$$\bar{\eta}_{\mathbf{n}}(x) = \frac{1}{k}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{1}_{\{X_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}}.$$

Hence, we prove that as $\mathbf{n} \to \infty$,

$$\mathbb{E}\int_{\mathbb{R}^d} |\bar{\eta}_{\mathbf{n}}(x) - \mathbb{E}\bar{\eta}_{\mathbf{n}}(x)|\mu(dx) \to 0. \tag{19}$$

Observe that $\bar{\eta}_{\mathbf{n}}(x) = \widehat{\eta}_{\mathbf{n}}(x)$ if we let $Y_{\mathbf{i}} = 1$ for all $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$. Consequently, the proof of (19) is the same as that of (17). Finally, combining (10)-(13) and (17)-(19), we get (9) and the proof is completed. □

Proof of Theorem 2 By (8), the proof is established if we prove that as $\mathbf{n} \to \infty$,

$$\int_{\mathbb{R}^d} |\eta(x) - \eta_{\mathbf{n}}(x)|\mu(dx) \to 0 \text{ with probability one.} \tag{20}$$

By (10)-(11), it suffices to prove that

$$\int_{\mathbb{R}^d} |\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x) - \eta_{\mathbf{n}}(x)|\mu(dx) \to 0 \text{ with probability one.} \tag{21}$$

Since

$$\begin{aligned}\int_{\mathbb{R}^d} |\mathbb{E}\widehat{\eta}_{\mathbf{n}}(x) - \eta_{\mathbf{n}}(x)|\mu(dx) \leq \int_{\mathbb{R}^d} |\widehat{\eta}_{\mathbf{n}}(x) - \mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)|\mu(dx) \\ + \int_{\mathbb{R}^d} |\widehat{\eta}_{\mathbf{n}}(x) - \eta_{\mathbf{n}}(x)|\mu(dx),\end{aligned}$$

the proof of (21) is established if we prove that

$$\int_{\mathbb{R}^d} |\widehat{\eta}_{\mathbf{n}}(x) - \mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)(x)|\mu(dx) \to 0 \text{ with probability one.} \tag{22}$$

and

$$\int_{\mathbb{R}^d} |\widehat{\eta}_{\mathbf{n}}(x) - \eta_{\mathbf{n}}(x))|\mu(dx) \to 0 \text{ with probability one.} \tag{23}$$

We first prove (22). To this aim, we use the blocks decomposition introduced by [7] (see also [16]) which will be useful afterwards. Without loss of generality, suppose for each $l = 1, ..., N$, $n_l = 2pq_l$ where $p = p(\mathbf{n})$ and $q = q_l(\mathbf{n})$ are strictly positive integers with $p(\mathbf{n}) \in [1, \min_{1\leq l\leq N} n_l/2]$ such that (5) and (6). Let

$$J_q = \{\mathbf{j} = (j_1, ..., j_N) \in \mathbb{N}^N : \ 0 \leq j_l \leq q_l - 1, \ \forall l = 1, ..., N\}.$$

We have $\mathrm{card}(J_q) = \prod_{l=1}^N q_l := r$. We define blocks as follow, for each $\mathbf{j} \in J_q$,

$$\begin{aligned}
\mathcal{S}_{\mathbf{j}}^{(1)} &= \{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} : 2j_k p + 1 \leq i_k \leq (2j_k + 1)p, \ k = 1, \dots, N\} \\
\mathcal{S}_{\mathbf{j}}^{(2)} &= \{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} : 2j_k p + 1 \leq i_k \leq (2j_k + 1)p, \ k = 1, \dots, N-1 \\
&\qquad and\ (2j_N + 1)p + 1 \leq i_N \leq 2(j_N + 1)p\} \\
&\qquad \dots \\
\mathcal{S}_{\mathbf{j}}^{(2^N-1)} &= \{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} : (2j_k + 1)p + 1 \leq i_k \leq 2(j_k + 1)p, \ k = 1, \dots, N-1 \\
&\qquad and\ 2j_N p + 1 \leq i_N \leq (2j_N + 1)p\} \\
\mathcal{S}_{\mathbf{j}}^{(2^N)} &= \{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} : (2j_k + 1)p + 1 \leq i_k \leq 2(j_k + 1)p, \ k = 1, \dots, N\}.
\end{aligned}$$

We have

$$\mathcal{I}_{\mathbf{n}} = \bigcup_{i=1}^{2^N} \bigcup_{\mathbf{j}\in J_q} \mathcal{S}_{\mathbf{j}}^{(i)}. \tag{24}$$

One can easily prove that for all $\mathbf{j} \in J_q$, card $\left(\mathcal{S}_{\mathbf{j}}^{(i)}\right) = p^N$ and for all $\mathbf{j} \neq \mathbf{j}'$, dist $\left(\mathcal{S}_{\mathbf{j}}^{(i)}, \mathcal{S}_{\mathbf{j}'}^{(i)}\right) \geq p$. Let $W_{\mathbf{j}}^{(i)} = \left((X_{\mathbf{i}}, Y_{\mathbf{i}}),\ \mathbf{i} \in \mathcal{S}_{\mathbf{j}}^{(i)}\right)$, for each $i = 1, ..., 2^N$ and $\mathbf{j} \in J_q$, and let $\psi : \{1, ..., r\} \to J_q$ be a bijection. We can define a lexicographic order relation $\leq_{\text{lex}}$ on J_q as follows: $\psi(m) \leq_{\text{lex}} \psi(m')$ if $m \leq m'$. For any $\mathbf{j} \in J_q$, we can find $m \in \{1, ..., r\}$ with $\psi(m) = \mathbf{j}$. Now, we use Lemma 2 together with a decomposition in blocks similar to that introduced by [7] (see also [16]) on the family of vectors $\left\{W_{\psi(m)}^{(i)},\ m = 1, ..., r\right\}$ to generate independent copies $\left\{\tilde{W}_{\psi(m)}^{(i)},\ m = 1, ..., r\right\}$ such that: they are mutually independent, for all $m \in \{1, ..., r\}$, $\tilde{W}_{\psi(m)}^{(i)} = \left((\tilde{X}_{\mathbf{i}}, \tilde{Y}_{\mathbf{i}})),\ \mathbf{i} \in \mathcal{S}_{\psi(m)}^{(i)}\right)$ has the same distribution as $W_{\psi(m)}^{(i)} = \left((X_{\mathbf{i}}, Y_{\mathbf{i}}),\ \mathbf{i} \in \mathcal{S}_{\psi(m)}^{(i)}\right)$ and $\mathbb{P}\left(W_{\psi(m)}^{(i)} \neq \tilde{W}_{\psi(m)}^{(i)}\right) \leq \beta(p)$ because dist $\left(\mathcal{S}_{\psi(m)}^{(i)}, \mathcal{S}_{\psi(m')}^{(i)}\right) \geq p$ for any $m \neq m'$. As a consequence, for each $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$, there exist $i = 1, ..., 2^N$ and $m = 1, ..., r$ such that

$$\mathbb{P}\left((X_{\mathbf{i}}, Y_{\mathbf{i}}) \neq (\tilde{X}_{\mathbf{i}}, \tilde{Y}_{\mathbf{i}})\right) \leq \mathbb{P}\left(W_{\psi(m)}^{(i)} \neq \tilde{W}_{\psi(m)}^{(i)}\right) \leq \beta(p) \tag{25}$$

Define

$$\tilde{\eta}_{\mathbf{n}}(x) = \frac{1}{k} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} \tilde{Y}_{\mathbf{i}} \mathbb{1}_{\{\tilde{X}_{\mathbf{i}} \in S_{x, \rho_{\mathbf{n}}}\}} \tag{26}$$

Then, for any $\epsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}\bigg(\int_{\mathbb{R}^d} |\widehat{\eta}_{\mathbf{n}}(x) - \mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)(x)|\mu(dx) > \epsilon\bigg) \\
&\leq \mathbb{P}\bigg(\Big|\int_{\mathbb{R}^d} |\widehat{\eta}_{\mathbf{n}}(x) - \mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)|\mu(dx) - \int_{\mathbb{R}^d} |\tilde{\eta}_{\mathbf{n}}(x) - \mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)|\mu(dx)\Big| > \epsilon/2\bigg) \\
&\qquad + \mathbb{P}\bigg(\int_{\mathbb{R}^d} |\tilde{\eta}_{\mathbf{n}}(x) - \mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)|\mu(dx) > \epsilon/2\bigg) := A_{\mathbf{n}} + B_{\mathbf{n}}.
\end{aligned} \tag{27}$$

We first find an upper bound for $A_{\mathbf{n}}$. We have by Markov's inequality

$$\begin{aligned}
A_{\mathbf{n}} &\leq 2\epsilon^{-1}\mathbb{E}\Big| \int_{\mathbb{R}^d} |\widehat{\eta}_{\mathbf{n}}(x) - \mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)|\mu(dx) - \int_{\mathbb{R}^d} |\tilde{\eta}_{\mathbf{n}}(x) - \mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)|\mu(dx)\Big| \\
&\leq 2\epsilon^{-1}\mathbb{E}\int_{\mathbb{R}^d} \Big| |\widehat{\eta}_{\mathbf{n}}(x) - \mathbb{E}\widehat{\eta}_{\mathbf{n}}(x)| - |\tilde{\eta}_{\mathbf{n}}(x) - \mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)| \Big| \mu(dx) \\
&\leq 2\epsilon^{-1}\mathbb{E}\Big(\int_{\mathbb{R}^d} |\tilde{\eta}_{\mathbf{n}}(x) - \widehat{\eta}_{\mathbf{n}}(x)|\mu(dx) + \mathbb{E}\int_{\mathbb{R}^d} |\tilde{\eta}_{\mathbf{n}}(x) - \widehat{\eta}_{\mathbf{n}}(x)|\mu(dx)\Big) \\
&= 4\epsilon^{-1}\mathbb{E}\int_{\mathbb{R}^d} |\tilde{\eta}_{\mathbf{n}}(x) - \widehat{\eta}_{\mathbf{n}}(x)|\mu(dx) \\
&= 4\epsilon^{-1}\mathbb{E}\int_{\mathbb{R}^d} \left| \frac{1}{k}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \tilde{Y}_{\mathbf{i}} 1\!\!1_{\{\tilde{X}_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} - \frac{1}{k}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} 1\!\!1_{\{X_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} \right| \mu(dx)| \\
&\leq 4\epsilon^{-1}k^{-1}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{E}1\!\!1_{(X_{\mathbf{i}},Y_{\mathbf{i}})\neq(\tilde{X}_{\mathbf{i}},\tilde{Y}_{\mathbf{i}})} \int_{\mathbb{R}^d} \left| \tilde{Y}_{\mathbf{i}} 1\!\!1_{\{\tilde{X}_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} - Y_{\mathbf{i}} 1\!\!1_{\{X_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} \right| \mu(dx) \\
&\leq 8\epsilon^{-1}k^{-1}\hat{\mathbf{n}}\beta(p).
\end{aligned}$$

Consequently, by (6), we get

$$\sum_{\mathbf{n}\in(\mathbb{N}^*)^N} A_{\mathbf{n}} < \infty. \tag{28}$$

Let us now find an upper bound for $B_{\mathbf{n}}$. We have

$$\begin{aligned}
B_{\mathbf{n}} &= \mathbb{P}\Big(\int_{\mathbb{R}^d} |\tilde{\eta}_{\mathbf{n}}(x) - \mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)|\mu(dx) > \epsilon/2 \Big) \\
&= \mathbb{P}\left(\frac{1}{k}\int_{\mathbb{R}^d} \left| \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \tilde{Y}_{\mathbf{i}} 1\!\!1_{\{\tilde{X}_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} - \frac{1}{k}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{E}\tilde{Y}_{\mathbf{i}} 1\!\!1_{\{\tilde{X}_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} \right| \mu(dx) > \epsilon/2 \right)
\end{aligned}$$

Consequently, (24) yields

$$B_{\mathbf{n}} \leq \sum_{i=1}^{2^N} \mathbb{P}\left(\frac{1}{k}\int_{\mathbb{R}^d} \left| \sum_{\mathbf{j}\in J_q} \sum_{\mathbf{i}\in\mathcal{S}_{\mathbf{j}}^{(i)}} \Big(\tilde{Y}_{\mathbf{i}} 1\!\!1_{\{\tilde{X}_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} - \mathbb{E}\tilde{Y}_{\mathbf{i}} 1\!\!1_{\{\tilde{X}_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} \Big) \right| \mu(dx) > \epsilon/2^{N+1} \right)$$

Hence, it suffices to find an upper bound for example for

$$\mathbb{P}\left(\frac{1}{k}\int_{\mathbb{R}^d} \left| \sum_{\mathbf{j}\in J_q} \sum_{\mathbf{i}\in\mathcal{S}_{\mathbf{j}}^{(1)}} \Big(\tilde{Y}_{\mathbf{i}} 1\!\!1_{\{\tilde{X}_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} - \mathbb{E}\tilde{Y}_{\mathbf{i}} 1\!\!1_{\{\tilde{X}_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}} \Big) \right| \mu(dx) > \epsilon/2^{N+1} \right).$$

To do that, we re-consider blocks decomposition and the lexicographic relation defined above. Denote for each $m = 1, ..., r$,

$$\tilde{W}_m := \tilde{W}^{(1)}_{\psi(m)} = \Big((\tilde{X}_{\mathbf{i}}, \tilde{Y}_{\mathbf{i}})),\ \mathbf{i}\in\mathcal{S}^{(1)}_{\psi(m)} \Big).$$

Define

$$F : \left(\left(\mathbb{R}^d \times \{0,1\} \right)^{p^N} \right)^r \to \mathbb{R}$$

such that

$$F(\tilde{W}_1, ..., \tilde{W}_r) = \frac{1}{k} \int_{\mathbb{R}^d} \left| \sum_{m=1}^{r} \sum_{\mathbf{i} \in \mathcal{S}^{(1)}_{\psi(m)}} \left(\tilde{Y}_{\mathbf{i}} \mathbb{1}_{\{\tilde{X}_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}} - \mathbb{E} \tilde{Y}_{\mathbf{i}} \mathbb{1}_{\{\tilde{X}_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}} \right) \right| \mu(dx).$$

With the same method that was used to prove (17), we can easily prove that

$$\mathbb{E} F(\tilde{W}_1, ..., \tilde{W}_r) \to 0.$$

As a consequence, we have for $\hat{\mathbf{n}}$ is enough large, we can write

$$\begin{aligned} &\mathbb{P} \left(\frac{1}{k} \int_{\mathbb{R}^d} \left| \sum_{m=1}^{r} \sum_{\mathbf{i} \in \mathcal{S}^{(1)}_{\psi(m)}} \left(\tilde{Y}_{\mathbf{i}} \mathbb{1}_{\{\tilde{X}_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}} - \mathbb{E} \tilde{Y}_{\mathbf{i}} \mathbb{1}_{\{\tilde{X}_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}} \right) \right| \mu(dx) > \epsilon / 2^{N+1} \right) \\ &\quad \le \mathbb{P} \left(\left| F(\tilde{W}_1, ..., \tilde{W}_r) - \mathbb{E} F(\tilde{W}_1, ..., \tilde{W}_r) > \epsilon / 2^{N+2} \right| \right). \end{aligned} \tag{29}$$

Let us fix the data and denote $\tilde{w}^*_m = \Big((\tilde{x}^*_{\mathbf{i}}, \tilde{y}^*_{\mathbf{i}})),\ \mathbf{i} \in \mathcal{S}^{(1)}_{\psi(m)} \Big)$. Thus, we have

$$\begin{aligned} &|F(\tilde{w}_1, ..., \tilde{w}_m, ..., \tilde{w}_r) - F(\tilde{w}_1, ..., \tilde{w}^*_m, ..., \tilde{w}_r)| \\ &\qquad \le \frac{1}{k} \int_{\mathbb{R}^d} \left| \sum_{\mathbf{i} \in \mathcal{S}^{(1)}_{\psi(m)}} \left(\tilde{y}_{\mathbf{i}} \mathbb{1}_{\{\tilde{x}_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}} - \tilde{y}^*_{\mathbf{i}} \mathbb{1}_{\{\tilde{x}^*_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}} \right) \right| \mu(dx) \\ &\qquad \le \sum_{\mathbf{i} \in \mathcal{S}^{(1)}_{\psi(m)}} \frac{1}{k} \int_{\mathbb{R}^d} \left| \tilde{y}_{\mathbf{i}} \mathbb{1}_{\{\tilde{x}_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}} - \tilde{y}^*_{\mathbf{i}} \mathbb{1}_{\{\tilde{x}^*_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}} \right| \mu(dx). \end{aligned}$$

But $\left| \tilde{y}_{\mathbf{i}} \mathbb{1}_{\{\tilde{x}_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}} - \tilde{y}^*_{\mathbf{i}} \mathbb{1}_{\{\tilde{x}^*_{\mathbf{i}} \in S_{x,\rho_{\mathbf{n}}}\}} \right| \le 2$ and it can be different from zero if and only if $\|\tilde{x}_{\mathbf{i}} - x\| \le \rho_{\mathbf{n}}$ or $\|\tilde{x}^*_{\mathbf{i}} - x\| \le \rho_{\mathbf{n}}$. Observe that by (7), $\|\tilde{x}_{\mathbf{i}} - x\| \le \rho_{\mathbf{n}}$ if and only if $\mu\left(S_{x, \|\tilde{x}_{\mathbf{i}} - x\|}\right) \le k/\hat{\mathbf{n}}$. But the measure of such x's is bounded by $\gamma_d k / \hat{\mathbf{n}}$ by Lemma 3 and card $\left(\mathcal{S}^{(1)}_{\psi(m)}\right) = p^N$. Therefore,

$$\sup_{\tilde{w}_1, ..., \tilde{w}_r, \tilde{w}^*_m} |F(\tilde{w}_1, ..., \tilde{w}_m, ..., \tilde{w}_r) - F(\tilde{w}_1, ..., \tilde{w}^*_m, ..., \tilde{w}_r)| \le \frac{2p^N}{k} \frac{\gamma_d k}{\hat{\mathbf{n}}} = \frac{2p^N \gamma_d}{\hat{\mathbf{n}}}.$$

As a consequence, according to McDiarmid's inequality (see [8]), we have

$$\begin{aligned} \mathbb{P} &\left(\left| F(\tilde{W}_1, ..., \tilde{W}_r) - \mathbb{E} F(\tilde{W}_1, ..., \tilde{W}_r) > \epsilon / 2^{N+2} \right| \right) \\ &\qquad\qquad \le 2 \exp\left(- \frac{\epsilon^2 \hat{\mathbf{n}}^2}{2^{2N+4} r p^{2N} \gamma_d^2} \right) \\ &\qquad\qquad = 2 \exp\left(- \frac{\epsilon^2 \hat{\mathbf{n}}}{2^{2N+3} p^N \gamma_d^2} \right). \end{aligned} \tag{30}$$

Combining (28)-(29) together with (5), we get

$$\sum_{\mathbf{n}\in(\mathbb{N}^*)^N} B_{\mathbf{n}} < \infty. \tag{31}$$

By (27), (28) and (31) together with Borel-Cantelli lemma, we have (22). To complete the proof, it remains to prove (23). As we show above $\bar{\eta}_{\mathbf{n}}(x) = \widehat{\eta}_{\mathbf{n}}(x)$ if we let $Y_{\mathbf{i}} = 1$ for all $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$ with

$$\bar{\eta}_{\mathbf{n}}(x) = \frac{1}{k} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{1}_{\{X_{\mathbf{i}}\in S_{x,\rho_{\mathbf{n}}}\}}.$$

Consequently, if we proceed similarly to (18), we can easily show that the proof of (23) is the same as that of (22) and the proof is completed. □

Acknowledgements

The authors would like to thank the anonymous referees whose valuable comments led to an improved version of the paper.

References

1. H.C.P. Berbee. Random walks with stationary increments and renewal theory. *Math. Cent. Tracts. Amsterdam*, 58, 1979.
2. G. Biau and L. Devroye. *Lectures on the Nearest Neighbor Method.* Springer, Springer Series in the Data Sciences, 2015.
3. D. Bosq and J.P. Lecoutre. *Théorie de l'estimation fonctionnelle.* Economica, Paris, 1987.
4. P. E. Cheng. Strong consistency of nearest neighbor regression function estimators. *Journal of Multivariate Analysis*, 15:63–72, 1984.
5. L. Devroye, L. Györfi, A. Krzyzak, and G. Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 22:1371–1385, 1994.
6. L. Devroye, L. Györfi, and G. Lugosi. *A probabilitic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.
7. P. Doukhan, P. Massart, and E. Rio. Invariance principles for absolutely regular empirical processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 31:393–427, 1995.
8. C. McDiarmid. On the method of bounded differences, in surrveys in combinatorics. *Cambridge University Press, Cambridge*, 794:261–283, 1989.
9. C. C. Neaderhouser. Convergence of block spins defined on random fields. *J. Statist. Phys.*, 22:673–684, 1980.
10. E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendents. Mathématiques et Applications.* Spriner, Berlin, 2000.
11. M. Rosenblatt. A central limit theorem and a strong mixing condition. *The Annals of Statistics*, 5:595–645, 1977.
12. M. Rosenblatt. *Stationary sequences and random fields.* Birkhauser, Boston, 1985.
13. Y. A. Rozanov and V.A. Volkonskii. Some limit theorems for random functions. *I. Teor. Veroyatn. Primen.*, 4:186–207, 1959.
14. C. J. Stone. Consistent nonparametric regression,. *Proc. Nat. Acad. Sci., USA*, 42:43–47, 1956.
15. L.T. Tran and S. Yakowitz. Nearest neighbor estimators for random fields. *Journal of Multivariate Analysis*, 44:23–46, 1993.

16. G. Viennet. Inequalities for absolutely sequence. *Application to density estimation. Probability Theory and Related Fields*, 107:467–492, 1967.
17. A. Younso. On nonparametric classification for weakly dependent functional processes. *ESAIM: Probability and Statistics*, 21:452–466, 2017.
18. A. Younso. On the consistency of a new kernel rule for spatially dependent data. *Statistics & Probability Letters*, 131:64–71, 2017.
19. A. Younso. On the consistency of kernel classification rule for functional random field. *Journal de la Société Française de Statistique*, 159:68–87, 2018.
20. A. Younso. Nonparametric discrimination of areal functional data. *Brazilian Journal of Probability and Statistics*, 34:12–126, 2020.
21. A. Younso, Z. Kanaya, and N. Azhari. Strong consistency of a kernel-based rule for spatially dependent data. *Arab Journal of Mathematical Sciences*, 26:211–225, 2019.
22. X. Zhang, R. Pan, G. Guan, X. Zhu, and H. Wang. Network logistic regression model. *Statistica Sinica*, 30:673–693, 2020.