IEICE TRANS. INF. & SYST., VOL.E94–D, NO.11 NOVEMBER 2011

PAPER

# The Lower Bound for the Nearest Neighbor Estimators with $(p, C)$-Smooth Regression Functions

**Takanori AYANO**[†*a)], ***Student Member***

**SUMMARY** Let $(X, Y)$ be a $\mathbb{R}^d \times \mathbb{R}$-valued random vector. In regression analysis one wants to estimate the regression function $m(x) := \mathbf{E}(Y|X = x)$ from a data set. In this paper we consider the convergence rate of the error for the $k$ nearest neighbor estimators in case that $m$ is $(p, C)$-smooth. It is known that the minimax rate is unachievable by any $k$ nearest neighbor estimator for $p > 1.5$ and $d = 1$. We generalize this result to any $d \geq 1$. Throughout this paper, we assume that the data is independent and identically distributed and as an error criterion we use the expected $L_2$ error.

***key words:*** *regression, nonparametric estimation, nearest neighbor, rate of convergence*

## 1. Introduction

Regression analysis plays an important role in many fields such as pattern recognition, data mining, economics, medicine. For example, in pattern recognition, one wants to estimate the posterior probability of a label given a pattern (cf. [2], pp.44–46, and [9], pp.6–9). In economics, when a bank lends a customer money, it wants to predict the profit from the profile of the customer such as income, age. In medicine, one wants to predict the survival time of a patient with a life-threatening disease from the type of disease, sex, age, therapy, etc. (cf. [9], pp.4–5). Recently, with the development of computers, nonparametric methods are attracting increasing interest. Many estimators are proposed and the performance of them are researched actively from theoretical and numerical aspects, for example, partitioning ([9], [15]), kernel ([9], [16], [26]), $k$-nearest neighbor ([4], [7], [9], [15], [17]), local polynomial ([9], [26]), projection estimators ([26]). In particular, the $k$-NN (nearest neighbor) estimators can be implemented easily and have the good prediction accuracy. So they are used in many practical applications very often. In this paper we analyze the performance of the $k$-NN estimators theoretically. Now we will start to state regression analysis.

Let $(X, Y)$ be a $\mathbb{R}^d \times \mathbb{R}$-valued random vector. In regression analysis, one wants to predict the value of $Y$ after having observed the value of $X$, i.e. to find a measurable function $f$ such that the mean squared error $\mathbf{E}_{XY}(f(X) - Y)^2$ is minimized, where $\mathbf{E}_{XY}$ denotes the expectation with respect to $(X, Y)$. Let $m(x) := \mathbf{E}\{Y|X = x\}$ (regression function), which is the conditional expectation of $Y$ given $X = x$. Then $m(x)$ is the solution of the minimization problem. In fact, one can check for any measurable function $f$,

$$\mathbf{E}_{XY}(f(X)-Y)^2 = \mathbf{E}_{XY}(m(X)-Y)^2 + \mathbf{E}_X(f(X)-m(X))^2 .$$

In statistics, only the data is available, (the distribution of $(X, Y)$ and $m$ are not available), and one needs to estimate the function $m$ from the data $\{(X_i, Y_i)\}_{i=1}^n$, which are independently distributed according to the distribution of $(X, Y)$. We attempt to construct an estimator $m_n$ of $m$ such that the expected $L_2$ error $R(m_n) := \mathbf{E}_{X^nY^n}\mathbf{E}_X(m_n(X) - m(X))^2$ is as small as possible, where $\mathbf{E}_{X^nY^n}$ denotes the expectation with respect to the data. In order to analyze the performance of estimators theoretically, it is very important to evaluate how fast the error $R(m_n)$ converges to zero as the data size $n$ tends to infinity. In this paper we analyze the convergence rate of the $k$-NN estimators in case that $m$ is $(p, C)$-smooth (cf. [9], p.37).

The $k$-NN estimator is defined as follows. Given $x \in \mathbb{R}^d$, we rearrange the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ in the ascending order of the values of $\|X_i - x\|$. As a tie-breaking rule, if $\|X_i - x\| = \|X_j - x\|$ and $i < j$, we declare that $X_i$ is "closer" to $x$ than $X_j$. We write the rearrange sequence by $(X_{1,x}, Y_{1,x}), \ldots, (X_{n,x}, Y_{n,x})$. Notice that $\{(X_{i,x}, Y_{i,x})\}_{i=1}^n$ is expressed by $\{(X_{\pi(i)}, Y_{\pi(i)})\}_{i=1}^n$ using a permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ depending on $x \in \mathbb{R}^d$. Then for $1 \leq k \leq n$, the $k$-NN estimator $m_n$ is defined by

$$m_n(x) = \frac{1}{k}\sum_{i=1}^{k} Y_{i,x} \quad .$$

For the details about the $k$-NN estimators, for example, see Chapter 6 in [9]. Let $\mathcal{K}$ be the class of the $k$-NN estimators such that $k$ may depend only on $n$ (i.e. $k$ does not depend on $x$ and the data, given $n$).

Let $p, C > 0$, and express $p$ by $p = q + r,\ q \in \mathbb{Z}_{\geq 0},\ 0 < r \leq 1$. We say that a function $m : \mathbb{R}^d \to \mathbb{R}$ is $(p, C)$-smooth if for all $q_1, \ldots, q_d \in \mathbb{Z}_{\geq 0}$ with $q = q_1 + \cdots + q_d$, the partial derivatives $\frac{\partial^q m}{\partial x_1^{q_1} \cdots \partial x_d^{q_d}}$ exist and for all $x, z \in \mathbb{R}^d$ the following is satisfied:

$$\left| \frac{\partial^q m}{\partial x_1^{q_1} \cdots \partial x_d^{q_d}}(x) - \frac{\partial^q m}{\partial x_1^{q_1} \cdots \partial x_d^{q_d}}(z) \right| \leq C\|x - z\|^r.$$

For $p, C, \sigma > 0$, let $\mathcal{D}(p, C, \sigma)$ be the class of distributions of $(X, Y)$ such that:

Manuscript received April 27, 2011.

Manuscript revised July 30, 2011.

[†]The author is with the Department of Mathematics, Graduate School of Science, Osaka University, Toyonaka-shi, 560–0043 Japan.

[*]Research Fellow of the Japan Society for the Promotion of Science.

a) E-mail: t-ayano@cr.math.sci.osaka-u.ac.jp

DOI: 10.1587/transinf.E94.D.2244

(I) $X$ is uniformly distributed on $[0, 1]^d$;

(II) $\mathbf{Var}(Y|X = x) \leq \sigma^2, \quad \forall x \in \mathbb{R}^d$; and

(III) $m$ is $(p, C)$-smooth,

where $\mathbf{Var}(Y|X = x)$ denotes the variance of $Y$ given $X = x$.

The lower bound for the class $\mathcal{D}(p, C, \sigma)$ is known (cf. [9], p.38), i.e. there exists $C_1 > 0$ (which does not depend on $n$) such that for any estimator $\{m_n\}_{n\geq 1}$ and any $n$

$$\sup_{(X,Y)\in\mathcal{D}(p,C,\sigma)} R(m_n) \geq C_1 n^{-2p/(2p+d)}. \tag{1}$$

For $0 < p \leq 1$, the minimax rate $n^{-2p/(2p+d)}$ is achieved by a $k$-NN estimator in $\mathcal{K}$ (cf. [9], pp.93,99), i.e. for $0 < p \leq 1$ there exist a $k$-NN estimator $\{m_n\}_{n\geq 1}$ in $\mathcal{K}$ and $C_2 > 0$ (which does not depend on $n$) such that for any $n$

$$\sup_{(X,Y)\in\mathcal{D}(p,C,\sigma)} R(m_n) \leq C_2 n^{-2p/(2p+d)}.$$

In another paper, the author showed that for $1 < p \leq 1.5$ the minimax rate $n^{-2p/(2p+d)}$ is achieved by a $k$-NN estimator in $\mathcal{K}$ (cf. [1]). For $d = 1$ and $p > 1.5$, it is shown that the minimax rate $n^{-2p/(2p+d)}$ is unachievable by any $k$-NN estimator in $\mathcal{K}$ (cf. [9], p.96). In this paper, we generalize the above result to any $d \geq 1$, i.e. for $p > 1.5$ we show that there exists $C_3 > 0$ (which does not depend on $n$) such that for any $k$-NN estimator $\{m_n\}_{n\geq 1}$ in $\mathcal{K}$ and any $n$

$$\sup_{(X,Y)\in\mathcal{D}(p,C,\sigma)} R(m_n) \geq C_3 n^{-3/(3+d)}.$$

Since for $p > 1.5$ we have $n^{-2p/(2p+d)} < n^{-3/(3+d)}$, we find that for $p > 1.5$ the minimax rate $n^{-2p/(2p+d)}$ is unachievable by any $k$-NN estimator in $\mathcal{K}$. The style of the proof is similar to [9], but for $d \geq 2$ the caluculation of the integral in the proof is not easy, because for $d = 1$ we deal with intervals but for $d \geq 2$ we must deal with balls.

Throughout this paper we use the following notations : $\mathbb{R}, \mathbb{R}_{>0}, \mathbb{Z}_{\geq 0}, \mathbb{N}$ are the sets of reals, positive reals, nonnegative integers, and positive integers. For a measurable set $D \subset \mathbb{R}^d$, $\mathrm{vol}(D)$ denotes the Lebesgue measure of $D$. For $x \in \mathbb{R}^d$, $\|x\|$ denotes the Euclidean norm of $x$. For $u, v \in \mathbb{R}^d$, we define $H(u, v) := \{w \in \mathbb{R}^d \mid \|w - u\| \leq \|v - u\|\}$ and $G(u, v) := H(u, v) \cap [0, 1]^d$.

## 2. Related Work

In this section, we overview the related work about consistency and the rate of convergence. For consistency, it was shown in [24] that the nearest neighbor estimators are universally consistent. Since then it was shown that many estimators share this property (cf. [5], [6], [8], [10], [13], [14], [19], [20], [27]–[29]). For the rate of convergence, we know several results as follows:

- [25] proved the lower bound (1);
- for the distributions satisfying (II)(III) with $0 < p \leq 1$ and the partitioning, kernel, and $k$-NN estimators, the minimax rate $n^{-2p/(2p+d)}$ is achievable if $X$ is bounded † (cf. [7], [9], [17], [23]);
- [15], [16] proved the same statement later without assuming that $X$ should be bounded;
- for the partitioning estimators and the class $\mathcal{D}(p, C, \sigma)$ with $p > 1$, [9] proved that the minimax rate $n^{-2p/(2p+d)}$ is unachievable for $d = 1$, but the generalization to any $d \geq 1$ is not difficult;
- for the kernel estimators, [9] proved the minimax rate $n^{-2p/(2p+d)}$ is achievable for $\mathcal{D}(p, C, \sigma)$ with $0 < p \leq 1.5$ and is unachievable for that with $p > 1.5$ and $d = 1$, but the generalization to any $d \geq 1$ is not difficult.

We summarize the above results in Table 1.

**Table 1** The achievability of the minimax rate for the estimators and $\mathcal{D}(p, C, \sigma)$.

| | achievable | unachievable |
|---|---|---|
| partitioning | $0 < p \leq 1$ | $p > 1$ |
| kernel | $0 < p \leq 1.5$ | $p > 1.5$ |
| $k$-NN | $0 < p \leq 1.5$ | $p > 1.5, d = 1$ |

## 3. Main Result

**Theorem** (for $d = 1$, due to [9], p.96)

Let $\sigma > 0$. We consider a distribution such that

(A) $X$ is uniformly distributed on $[0, 1]^d$;

(B) $\mathbf{Var}(Y|X = x) = \sigma^2, \quad \forall x \in \mathbb{R}^d$; and

(C) $m(x) = x^{(1)}$, for $x = (x^{(1)}, \ldots, x^{(d)})$.

Then, there exists $C_3 > 0$ (which does not depend on $n$) such that for any $k$-NN estimator $\{m_n\}_{n\geq 1}$ in $\mathcal{K}$ and any $n$

$$\mathbf{E}_{X^nY^n}\mathbf{E}_X\left(m_n(X) - m(X)\right)^2 \geq C_3 n^{-3/(3+d)}.$$

**Remark 1**

There exists a distribution satisfying (A) (B) (C). In fact, let

$$f(x, y) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - x^{(1)})^2}{2\sigma^2}\right) \cdot \mathbf{1}_{[0,1]^d}(x),$$

where $\mathbf{1}_{[0,1]^d}(x) := \begin{cases} 1 & x \in [0, 1]^d \\ 0 & x \notin [0, 1]^d \end{cases}$. Then, the distribution which has the density function $f(x, y)$ satisfies (A) (B) (C).

**Remark 2**

From (C), $m$ is $(p, C)$-smooth with any $p > 1$ and $C > 0$, thus any distribution satisfying (A) (B) (C) is included in $\mathcal{D}(p, C, \sigma)$ with any $p > 1$ and $C > 0$. On the other hand, for $p > 1.5$, we have $n^{-2p/(2p+d)} < n^{-3/(3+d)}$. Hence, for $\mathcal{D}(p, C, \sigma)$ with $p > 1.5$, the minimax rate $n^{-2p/(2p+d)}$ is unachievable by any $k$-NN estimator in $\mathcal{K}$.

**Proof**. Suppose we are given $X = x, X_1 = x_1, \ldots, X_n = x_n$.

†for the $k$-NN estimators, the condition $d > 2p$ is required as well.

We take the expectation with respect to $Y_1, \ldots, Y_n$. Then the following bias-variance decomposition is derived:

$$\mathbf{E}_{Y^n}(m_n(x) - m(x))^2 = \mathbf{E}_{Y^n}\left(\frac{1}{k}\sum_{i=1}^{k} Y_{i,x} - m(x)\right)^2$$

$$= \mathbf{E}_{Y^n}\left(\frac{1}{k}\sum_{i=1}^{k}(Y_{i,x} - m(x_{i,x}))\right)^2 + \left\{\frac{1}{k}\sum_{i=1}^{k} m(x_{i,x}) - m(x)\right\}^2$$

$$= \frac{\sigma^2}{k} + \left\{\frac{1}{k}\sum_{i=1}^{k} x_{i,x}^{(1)} - x^{(1)}\right\}^2$$

where $x_{i,x} = (x_{i,x}^{(1)}, \ldots, x_{i,x}^{(d)})$ and the last equality is obtained from (B) and (C). We regard $x_1, \ldots, x_n$ as the random variables $X_1, \ldots, X_n$ and take the expectation with respect to $X_1, \ldots, X_n$. By Schwarz's inequality

$$\mathbf{E}_{X^nY^n}(m_n(x) - m(x))^2 = \frac{\sigma^2}{k} + \mathbf{E}_{X^n}\left\{\frac{1}{k}\sum_{i=1}^{k} X_{i,x}^{(1)} - x^{(1)}\right\}^2$$

$$\geq \frac{\sigma^2}{k} + \left\{\frac{1}{k}\,\mathbf{E}_{X^n}\sum_{i=1}^{k} X_{i,x}^{(1)} - x^{(1)}\right\}^2. \tag{2}$$

We evaluate $\mathbf{E}_{X^n}\sum_{i=1}^{k} X_{i,x}^{(1)}$ to obtain the lower bound. For $k = n$, since $\mathbf{E}_{X^n}\sum_{i=1}^{n} X_{i,x}^{(1)} = \mathbf{E}_{X^n}\sum_{i=1}^{n} X_i^{(1)} = (1/2)\cdot n$ and (2), $\mathbf{E}_X\mathbf{E}_{X^nY^n}(m_n(X) - m(X))^2 \geq \frac{1}{12}$. Therefore, in order to prove Theorem, it is enough to consider the case $1 \leq k \leq n-1$.

**Claim 1**
For any $x, x_1, \ldots, x_n \in [0,1]^d$,

$$\sum_{i=1}^{k} x_{i,x}^{(1)} \geq \sum_{i=1}^{k} x_{i,x'}^{(1)} ,$$

where for $x = (x^{(1)}, \ldots, x^{(d)})$ we put $x' = (0, x^{(2)} \ldots, x^{(d)})$. For $d = 1$, since $x' = 0$, Claim 1 is trivial but for $d \geq 2$ it is not easy. See Appendix for proof.

Let $D := \{(x_1, \ldots, x_n) \mid \|x_i - x'\| < \|x_{k+1} - x'\|,\ i = 1, \ldots, k, \|x_j - x'\| > \|x_{k+1} - x'\|, j = k+2, \ldots, n\}$.

We give an example of $(x_1, \ldots, x_n) \in D$ for $d = 2$, $n = 4$ and $k = 1$ (see Fig.1).

**Claim 2**

$$\mathbf{E}_{X^n}\sum_{i=1}^{k} X_{i,x'}^{(1)} = \frac{n\cdots(n-k)}{k!}\sum_{i=1}^{k}\int_D x_i^{(1)}\,dx_1\cdots dx_n ,$$

(see Appendix for proof).

From Claim 1 and Claim 2, we have

$$\frac{1}{k}\cdot\mathbf{E}_{X^n}\sum_{i=1}^{k} X_{i,x}^{(1)} \geq \frac{n\cdots(n-k)}{k\cdot k!}\sum_{i=1}^{k}\int_D x_i^{(1)}\,dx_1\cdots dx_n$$

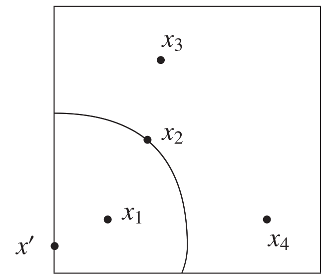



**Fig. 1** The location of $(x_1, x_2, x_3, x_4) \in D$ for $d = 2$, $n = 4$ and $k = 1$.

$$= \frac{n\cdots(n-k)}{k!}\int_{[0,1]^d}\left\{\int_{G(x',x_{k+1})} x_1^{(1)}dx_1\right\}\cdot$$

$$\mathrm{vol}[G(x', x_{k+1})]^{k-1}\{1 - \mathrm{vol}[G(x', x_{k+1})]\}^{n-k-1}\,dx_{k+1} .$$

**Claim 3**
There exists $c_1 > 0$ (depending only on $d$) such that

$$\int_{G(x',x_{k+1})} x_1^{(1)}dx_1 \geq c_1\mathrm{vol}[G(x', x_{k+1})]^{(d+1)/d},$$

(see Appendix for proof).

From Claim 3, we have

$$\frac{1}{k}\cdot\mathbf{E}_{X^n}\sum_{i=1}^{k} X_{i,x}^{(1)} \geq c_1\frac{n\cdots(n-k)}{k!}\cdot$$

$$\int_{[0,1]^d}\mathrm{vol}[G(x', y)]^{k+\frac{1}{d}}\{1 - \mathrm{vol}[G(x', y)]\}^{n-k-1}dy.$$

For $d = 1$, since $x' = 0$ and $\mathrm{vol}[G(x', y)] = y$, it is easy to calculate the above integral directly, but for $d \geq 2$ it is not easy to represent $\mathrm{vol}[G(x', y)]$ by $x'$ and $y$. So it is not easy to calculate the above integral. We get over the problem by the following way. Let $\phi : [0,1]^d \to [0,1]$ and $\psi : [0,1] \to [0,1]$ be the maps defined by $\phi(y) = \mathrm{vol}[G(x', y)]$ and $\psi(u) = u^{k+\frac{1}{d}}(1-u)^{n-k-1}$. Let $\mu$ be the measure on $[0,1]$ induced by $\phi$ and the Lebesgue measure on $[0,1]^d$. Then, by the formula of change of variables (cf. [3], p. 216, Theorem 16.13), we have

$$\int_{[0,1]^d}\psi\circ\phi(y)\,dy = \int_0^1\psi(u)\,\mu(du) .$$

Let $F(u) := \mathrm{vol}[\{y \in [0,1]^d \mid 0 \leq \mathrm{vol}[G(x', y)] \leq u\}]$ for $0 \leq u \leq 1$, then we have $F(u) = u$ and $F'(u) = 1$. Note that $F'(u)$ is the density function of $\mu$.

$$\frac{1}{k}\cdot\mathbf{E}_{X^n}\sum_{i=1}^{k} X_{i,x}^{(1)}$$

$$\geq c_1\frac{n\cdots(n-k)}{k!}\int_0^1 u^{k+\frac{1}{d}}(1-u)^{n-k-1}\mu(du)$$

$$= c_1\frac{n\cdots(n-k)}{k!}\int_0^1 u^{k+\frac{1}{d}}(1-u)^{n-k-1}du .$$

For $\alpha \in \mathbb{R}_{>0}$ and $\beta \in \mathbb{N}$, let

$$B(\alpha,\beta) := \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du \quad \text{(Beta function).}$$

Then the following formula is well-known:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \frac{\Gamma(\alpha)\,(\beta-1)!}{(\alpha+\beta-1)\cdots\alpha\Gamma(\alpha)}$$
$$= \frac{(\beta-1)!}{\alpha\cdots(\alpha+\beta-1)} \quad ,$$

where $\Gamma$ is Gamma function.

On the other hand, by Stirling's formula,

$$\lim_{n\to\infty} \frac{n!}{\left(1+\frac{1}{d}\right)\cdots\left(n+\frac{1}{d}\right)} \cdot n^{\frac{1}{d}}$$
$$= \lim_{n\to\infty} \frac{\Gamma(n+1)\cdot\Gamma\left(1+\frac{1}{d}\right)}{\Gamma\left(n+1+\frac{1}{d}\right)} \cdot n^{\frac{1}{d}}$$
$$= \lim_{n\to\infty} \Gamma\left(1+\frac{1}{d}\right) \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi\left(n+\frac{1}{d}\right)}\left(\frac{n+\frac{1}{d}}{e}\right)^{n+\frac{1}{d}}} \cdot n^{\frac{1}{d}}$$
$$= \Gamma\left(1+\frac{1}{d}\right) .$$

Therefore, there exist $c_2, c_3 > 0$ (depending only on $d$) such that

$$c_2 n^{-\frac{1}{d}} \le \frac{n!}{\left(1+\frac{1}{d}\right)\cdots\left(n+\frac{1}{d}\right)} \le c_3 n^{-\frac{1}{d}} \quad .$$

Therefore,

$$\frac{1}{k}\cdot \mathbf{E}_{X^n} \sum_{i=1}^{k} X_{i,x}^{(1)} \ge c_1 \frac{n\cdots(n-k)}{k!} B\left(k+1+\frac{1}{d},\ n-k\right)$$
$$= c_1 \frac{(k+1)\cdots n}{\left(k+1+\frac{1}{d}\right)\cdots\left(n+\frac{1}{d}\right)}$$
$$= c_1 \frac{n!}{\left(1+\frac{1}{d}\right)\cdots\left(n+\frac{1}{d}\right)} \Big/ \frac{k!}{\left(1+\frac{1}{d}\right)\cdots\left(k+\frac{1}{d}\right)}$$
$$\ge \frac{c_1 c_2}{c_3}\left(\frac{k}{n}\right)^{1/d} \quad .$$

We can take $c_1 \le 1$. From $c_2 \le c_3$, we have $\frac{c_1 c_2}{c_3}\left(\frac{k}{n}\right)^{1/d} \le 1$. Let $c_4 := \frac{c_1 c_2}{c_3}$. From (2), we have

$$\mathbf{E}_X \mathbf{E}_{X^n Y^n} \left(m_n(X) - m(X)\right)^2 \ge \frac{\sigma^2}{k} +$$
$$\int_{0\le x_1 \le c_4\left(\frac{k}{n}\right)^{\frac{1}{d}}, 0\le x_2\le 1,\dots,0\le x_d\le 1} \left\{\frac{1}{k}\,\mathbf{E}_{X^n}\sum_{i=1}^{k} X_{i,x}^{(1)} - x^{(1)}\right\}^2 dx$$
$$\ge \frac{\sigma^2}{k} + \int_{0\le x_1 \le c_4\left(\frac{k}{n}\right)^{\frac{1}{d}}, 0\le x_2\le 1,\dots,0\le x_d\le 1} \left\{c_4\left(\frac{k}{n}\right)^{\frac{1}{d}} - x^{(1)}\right\}^2 dx$$
$$= \frac{\sigma^2}{k} + \frac{c_4^3}{3}\left(\frac{k}{n}\right)^{\frac{3}{d}} \ge C_3 n^{-\frac{3}{d+3}} \quad ,$$

where $C_3$ is a positive constant (which does not depend on $n$). We have got Theorem. □

## 4. Conclusion

In this paper we analyzed the convergence rate of the error for the $k$-NN estimators in $\mathcal{K}$ in the case where the regression function is $(p, C)$-smooth. We showed that no matter how smooth the regression function is, one can not achieve the faster rate than $n^{-3/(3+d)}$ by any $k$-NN estimator in $\mathcal{K}$ (for $d = 1$, due to [9]). Hence we obtained that for $p > 1.5$ the minimax rate $n^{-2p/(2p+d)}$ is unachievable by any $k$-NN estimator in $\mathcal{K}$. On the other hand, in another paper, the author showed for $0 < p \le 1.5$ the minimax rate is achievable by a $k$-NN estimator in $\mathcal{K}$ (for $0 < p \le 1$, due to [9]). As a result, for the classes $\mathcal{D}(p, C, \sigma)$ and $\mathcal{K}$, the upper bound coincides with the lower bound.

## Acknowledgments

The author wishes to thank Prof. Joe Suzuki and the two reviewers for reading the paper carefully and giving a lot of useful advise. This research was supported by the Japan Society for the Promotion of Science.

### References

[1] T. Ayano, Rates of convergence for nearest neighbor estimators with the smoother regression function, arXiv:1102.5633v1 [math.ST] 28 Feb. 2011.
[2] C.M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, 2006.
[3] P. Billingsley, Probability and Measure, Third Edition, Wiley, 1995.
[4] T.M. Cover and P.E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol.13, no.1, pp.21–27, 1967.
[5] L. Devroye, L. Györfi, A. Krzyżak, and G. Lugosi, "On the strong universal consistency of nearest neighbor regression function estimates," Ann. Statist., vol.22, pp.1371–1385, 1994.
[6] W. Greblicki, A. Krzyżak, and M. Pawlak, "Distribution-free pointwise consistency of kernel regression estimate," Ann. Statist., vol.12, pp.1570–1575, 1984.
[7] L. Györfi, "The rate of convergence of $k_n$-NN regression estimates and classification rules," IEEE Trans. Inf. Theory, vol.27, pp.362–364, 1981.
[8] L. Györfi and H. Walk, "On the strong universal consistency of a recursive regression estimate by Pál Révész," Statist. Probab. Lett., vol.31, pp.177–183, 1997.
[9] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer Series in Statistics, Springer, New York, 2002.
[10] M. Kohler, "Universally consistent regression function estimation using hierarchial B-splines," J. Multivariate Anal., vol.67, pp.138–164, 1999.
[11] M. Kohler, "Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression," J. Statist. Plann. Inference, vol.89, pp.1–23, 2000.
[12] M. Kohler and A. Krzyżak, "On the rate of convergence of local averaging plug-in classification rules under a margin condition," IEEE Trans. Inf. Theory, vol.53, no.5, pp.1735–1742, 2007.

[13] M. Kohler and A. Krzyżak, "Nonparametric regression estimation using penalized least squares," IEEE Trans. Inf. Theory, vol.47, pp.3054–3058, 2001.
[14] M. Kohler, "Universal consistency of local polynomial kernel regression estimates," Ann. Inst. Statist. Math., vol.54, pp.879–899, 2002.
[15] M. Kohler, A. Krzyżak, and H. Walk, "Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data," J. Multivariate Anal., vol.97, pp.311–323, 2006.
[16] M. Kohler, A. Krzyżak, H. Walk, "Optimal global rates of convergence for nonparametric regression with unbounded data," J. Statist. Plann. Inference, vol.139, pp.1286–1296, 2009.
[17] S.R. Kulkarni and S.E. Posner, "Rates of convergence of nearest neighbor estimation under arbitrary sampling," IEEE Trans. Inf. Theory, vol.41, pp.1028–1039, 1995.
[18] E. Liitiäinen, F. Corona, and A. Lendasse, "Nearest neighbor distributions and noise variance estimation," ESANN'2007 Proc.- European Symposium on Artificial Neural Networks, pp.67–72, Bruges, Belgium, 2007.
[19] G. Lugosi and K. Zeger, "Nonparametric estimation via empirical risk minimization," IEEE Trans. Inf. Theory, vol.41, pp.677–687, 1995.
[20] A. Nobel, "Histogram regression estimation using data-dependent partitions," Ann. Statist., vol.24, pp.1084–1105, 1996.
[21] M.G. Schimek, Smoothing and Regression, Approaches, Computation, and Application, Wiley Series in Probability and Statistics, 2000.
[22] G. Shakhnarovich, T. Darrell, and P. Indyk, Nearest-Neighbor Methods in Learning and Vision : Theory and Practice, MIT Press, Cambridge, Mass. 2005.
[23] C. Spiegelman and J. Sacks, Consistent window estimation in nonparametric regression, Ann. Statist., vol.8, pp.240–246, 1980.
[24] C.J. Stone, "Consistent nonparametric regression," Ann. Statist., vol.5, pp.595–645, 1977.
[25] C.J. Stone, "Optimal global rates of convergence for nonparametric regression," Ann. Statist., vol.10, pp.1040–1053, 1982.
[26] A.B. Tsybakov, Introduction to Nonparametric Estimation, Springer Series in Statistics, 2009.
[27] H. Walk, "Almost sure convergence properties of Nadaraya-Watson regression estimates," Modeling Uncertainty:An Examination of its Theory, Methods and Applications, ed. M. Dror, P. L'Ecuyer, F. Szidarovszky, pp.201–223, Kluwer Academic Publishers, Dordrecht, 2002, 2002.
[28] H. Walk, "Strong universal consistency of smooth kernel regression estimates," Ann. Inst. Statist. Math., vol.57, pp.665–685, 2005.
[29] H. Walk, "A universal strong law of large numbers for conditional expectations via nearest neighbors," J. Multivariate. Anal., vol.99, pp.1035–1050, 2008.

## Appendix A: Proof of Claim 1

Suppose $\sum_{i=1}^k x_{i,x}^{(1)} < \sum_{i=1}^k x_{i,x'}^{(1)}$. Let

$$\{x_{i_1}, \ldots, x_{i_t}\} = \{x_{i,x}\}_{i=1}^k - \left(\{x_{i,x}\}_{i=1}^k \cap \{x_{i,x'}\}_{i=1}^k\right)$$

and

$$\{x_{j_1}, \ldots, x_{j_t}\} = \{x_{i,x'}\}_{i=1}^k - \left(\{x_{i,x}\}_{i=1}^k \cap \{x_{i,x'}\}_{i=1}^k\right) .$$

Without loss of generality, we assume $x_{i_1}^{(1)} \le \cdots \le x_{i_t}^{(1)}$ and $x_{j_1}^{(1)} \le \cdots \le x_{j_t}^{(1)}$. Then, from $\sum_{i=1}^k x_{i,x}^{(1)} < \sum_{i=1}^k x_{i,x'}^{(1)}$, we get $x_{i_1}^{(1)} < x_{j_t}^{(1)}$. From $\|x_{i_1} - x\| \le \|x_{j_t} - x\|$, we have $2\sum_{s=1}^d (x_{j_t}^{(s)} - x_{i_1}^{(s)})x^{(s)} \le \sum_{s=1}^d \left((x_{j_t}^{(s)})^2 - (x_{i_1}^{(s)})^2\right)$. For $x^{(1)} = 0$, Claim 1 is trivial, so we assume $x^{(1)} > 0$. From $x_{i_1}^{(1)} < x_{j_t}^{(1)}$, we have $2\sum_{s=2}^d (x_{j_t}^{(s)} - x_{i_1}^{(s)})x^{(s)} < \sum_{s=1}^d \left((x_{j_t}^{(s)})^2 - (x_{i_1}^{(s)})^2\right)$. Therefore, $\|x_{i_1} - x'\| < \|x_{j_t} - x'\|$. This is contradiction. Hence we get Claim 1. □

## Appendix B: Proof of Claim 2

Let $h \in N := \{1, \ldots, n\}$ and $I, J \subset N\backslash\{h\}$ such that $\sharp I = k, I \cap J = \{\}$, and $I \cup J = N\backslash\{h\}$, where $\sharp$ denotes the number of the elements. Let $D(I, J, h) := \{(x_1, \ldots, x_n) \mid \|x_i - x'\| < \|x_h - x'\|, i \in I, \|x_j - x'\| > \|x_h - x'\|, j \in J\}$. First, we show $\mathrm{vol}\left\{[0,1]^{dn}\backslash \bigcup_{I,J,h} D(I,J,h)\right\} = 0$. Let $A_{ij} = \{(x_1, \ldots, x_n) \in [0,1]^{dn} \mid \|x_i - x'\| = \|x_j - x'\|\}$. Then we have

$$\left\{[0,1]^{dn}\backslash \bigcup_{I,J,h} D(I,J,h)\right\} \subset \bigcup_{1\le i,j\le n,\ i\ne j} A_{ij} .$$

Let $A = \{(u, v) \in [0,1]^d \times [0,1]^d \mid \|u - x'\| = \|v - x'\|\}$, then we have $\mathrm{vol}(A_{ij}) = \mathrm{vol}(A)$ for any $i, j$ with $i \ne j$. Given $u^{(2)}, \ldots, u^{(d)}$ and $v$, let $D(u^{(2)}, \ldots, u^{(d)}, v) = \{u^{(1)} \in [0,1] \mid \|u - x'\| = \|v - x'\|\}$. Since $\sharp D(u^{(2)}, \ldots, u^{(d)}, v) \le 2$, we have $\mathrm{vol}\left\{D(u^{(2)}, \ldots, u^{(d)}, v)\right\} = 0$. Hence, by Fubini's theorem,

$$\begin{aligned}\mathrm{vol}(A) &= \int_A 1 du dv \\ &= \int_{[0,1]^{2d-1}} \left\{\int_{D(u^{(2)},\ldots,u^{(d)},v)} 1\ du^{(1)}\right\} du^{(2)} \cdots du^{(d)} dv = 0.\end{aligned}$$

Hence,

$$\mathrm{vol}\left\{[0,1]^{dn}\backslash \bigcup_{I,J,h} D(I,J,h)\right\} \le \sum_{1\le i,j\le n,\ i\ne j} \mathrm{vol}(A_{ij}) = 0$$

Next, we show $D(I, J, h) \cap D(I', J', h') = \{\}$ for $(I, J, h) \ne (I', J', h')$. Assume $D(I, J, h) \bigcap D(I', J', h') \ne \{\}$. Let $(y_1, \ldots, y_n) \in D(I, J, h) \bigcap D(I', J', h')$. First, suppose $h \ne h'$. Since $(y_1, \ldots, y_n) \in D(I, J, h)$, $y_h$ is the $(k+1)$-th nearest element to $x'$ in $\{y_1, \ldots, y_n\}$. On the other hand, since $(y_1, \ldots, y_n) \in D(I', J', h')$, $y_{h'}$ is the $(k+1)$-th nearest element to $x'$ in $\{y_1, \ldots, y_n\}$, and $y_h$ is not. This is the contradiction. Next, suppose $h = h'$. There exists $s$ such that $s \in I$ and $s \in J'$. Since $(y_1, \ldots, y_n) \in D(I, J, h)$, we have $\|y_s - x'\| < \|y_h - x'\|$. On the other hand, since $(y_1, \ldots, y_n) \in D(I', J', h')$, we have $\|y_s - x'\| > \|y_h - x'\|$. This is the contradiction. Hence we obtain the assertion. From the above two results, we have

$$\mathbf{E}_{X^n} \sum_{i=1}^k X_{i,x'}^{(1)} = \sum_{I,J,h} \int_{D(I,J,h)} \sum_{i\in I} x_i^{(1)} dx_1 \cdots dx_n .$$

Finally, we show that for each $(I, J, h)$ the above integral has the same value. For $(I, J, h)$ and $(I', J', h')$, let $I = \{i_1, \ldots, i_k\}, J = \{j_{k+2}, \ldots, j_n\}, I' = \{i'_1, \ldots, i'_k\}$, and $J' = \{j'_{k+2}, \ldots, j'_n\}$ (for $k = n-1$, $J$ and $J'$ are empty). Let $\eta$ be the permutation of $\{1, \ldots, n\}$ defined by $\eta(i'_s) = i_s$,

$\eta(h') = h$ and $\eta(j'_t) = j_t$ for $1 \le s \le k$ and $k+2 \le t \le n$. Let $\Phi : [0,1]^{dn} \to [0,1]^{dn}$ be the bijection defined by $\Phi((x_1,\dots,x_n)) = (x_{\eta(1)},\dots,x_{\eta(n)})$. Let $g : [0,1]^{dn} \to \mathbb{R}$ be the function defined by $g(x_1,\dots,x_n) = \sum_{i\in I'} x_i^{(1)}$. Then by the formula of change of variables we have

$$\begin{aligned}&\int_{D(I',J',h')} \sum_{i\in I'} x_i^{(1)} dx_1\cdots dx_n\\ &= \int_{\Phi^{-1}(D(I',J',h'))} g\circ\Phi\,(x_1,\dots,x_n)\,|\det(J_\Phi)|\,dx_1\cdots dx_n\\ &= \int_{D(I,J,h)} \sum_{i\in I} x_i^{(1)} dx_1\cdots dx_n,\end{aligned}$$

where $J_\Phi$ is the Jacobian matrix of $\Phi$. Hence we obtain the assertion. Since the number of $(I,J,h)$ is ${}_nC_k\cdot(n-k)$, we get Claim 2.

□

## Appendix C: Proof of Claim 3

Let $e_1 := \int_{\|y\|\le 1,\ y^{(1)}\ge 0} y^{(1)}dy$, and $e_2 := \int_{\|y\|\le 1} dy$, then for any $R \ge 0$,

$$\int_{\|y\|\le R,\ y^{(1)}\ge 0} y^{(1)}dy = e_1R^{d+1} \text{ and } \int_{\|y\|\le R} dy = e_2R^d.$$

Suppose $\|x'-x_{k+1}\| \le 1/2$. Let $I := \{i \mid 0 \le x'^{(i)} \le 1/2,\ 1 \le i \le d\}$ and $M := \{y \mid \|y-x'\| \le \|x'-x_{k+1}\|,\ y^{(i)} \ge x'^{(i)},\ i \in I,\ y^{(j)} \le x'^{(j)},\ j \notin I\}$. We give a figure of $M$ for $d=2$ and $0 \le x'^{(2)} \le \frac{1}{2}$ (see Fig.A·1). We show $M \subset G(x',x_{k+1})$. Let $y \in M$. Since $\|x'-x_{k+1}\| \le 1/2$, we have $\|y-x'\| \le 1/2$. Hence, for any $1 \le i \le d$, we have $|y^{(i)}-x'^{(i)}| \le 1/2$. For $i \in I$, since $y^{(i)} \ge x'^{(i)}$ and $0 \le x'^{(i)} \le 1/2$, we have $0 \le y^{(i)} \le 1$. For $i \notin I$, since $y^{(i)} \le x'^{(i)}$ and $1/2 < x'^{(i)} \le 1$, we have $0 \le y^{(i)} \le 1$. Hence we have $y \in [0,1]^d$, and we obtain the assertion.

$$\begin{aligned}&\int_{G(x',x_{k+1})} x_1^{(1)}dx_1 \ge \int_M y^{(1)}dy\\ &= \frac{1}{2^{d-1}}\int_{\|y-x'\|\le\|x'-x_{k+1}\|,\ y^{(1)}\ge 0} y^{(1)}dy\\ &= \frac{e_1}{2^{d-1}}\|x'-x_{k+1}\|^{d+1}\\ &= \frac{e_1}{2^{d-1}e_2^{(d+1)/d}}(e_2\|x'-x_{k+1}\|^d)^{(d+1)/d}\\ &\ge \frac{e_1}{2^{d-1}e_2^{(d+1)/d}}\mathrm{vol}[G(x',x_{k+1})]^{(d+1)/d}\quad .\end{aligned}$$

Suppose $\|x'-x_{k+1}\| > 1/2$. Let $z \in \mathbb{R}^d$ such that $\|x'-z\| = 1/2$. By applying the proof of $\|x'-x_{k+1}\| \le 1/2$ for $z$, we have

$$\int_{G(x',z)} x_1^{(1)}dx_1 \ge \frac{e_1}{2^{d-1}}\|x'-z\|^{d+1} = \frac{e_1}{2^{d-1}2^{d+1}}\quad .$$

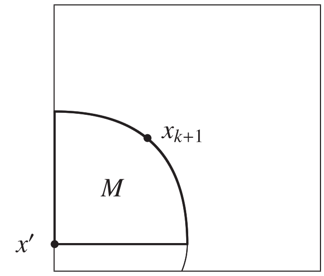



**Fig. A·1** The domain of $M$ for $d=2$ and $0 \le x'^{(2)} \le 1/2$.

Hence, we have

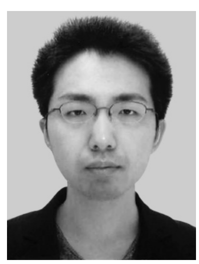
$$\int_{G(x',x_{k+1})} x_1^{(1)}dx_1 \ge \int_{G(x',z)} x_1^{(1)}dx_1 \ge \frac{e_1}{2^{d-1}2^{d+1}}\quad .$$

From $\|x'-x_{k+1}\| \le \sqrt{d}$, we have $\mathrm{vol}[G(x',x_{k+1})] \le e_2d^{d/2}$. Therefore,

$$\begin{aligned}&\int_{G(x',x_{k+1})} x_1^{(1)}dx_1\\ &\ge \frac{e_1}{2^{2d}e_2^{(d+1)/d}d^{(d+1)/2}}\mathrm{vol}[G(x',x_{k+1})]^{(d+1)/d}.\end{aligned}$$

We get Claim 3.

□



**Takanori Ayano** was born in 1985. He received his B.E. degree from Department of Mathematics, Kyoto University in 2008, and M.E. degree from Department of Mathematics, Osaka University in 2010. Currently he is in second grade of the doctoral course in Department of Mathematics, Osaka University, and Research Fellow of the Japan Society for the Promotion of Science. His research interests include learning theory, pattern recognition and statistics. He is a member of MSJ and JSS.