On histogram-based regression and classification with incomplete data

Eric Han[1] **and Majid Mojirsheibani**[2]
Department of Mathematics, California State University Northridge, CA, 91330, USA

Abstract

We consider the problem of nonparametric regression with possibly incomplete covariate vectors. The proposed estimators, which are based on histogram methods, are fully nonparametric and straightforward to implement. The presence of incomplete covariates is handled by an inverse weighting method, where the weights are estimates of the conditional probabilities of having incomplete covariate vectors. We also derive various exponential bounds on the L_1 norms of our estimators, which can be used to establish strong consistency results for the corresponding, closely related, problem of nonparametric classification with missing covariates. As the main focus and application of our results, we consider the problem of pattern recognition and statistical classification in the presence of incomplete covariates and propose histogram classifiers that are asymptotically optimal.

Keywords Local averaging, incomplete data, strong consistency, classification.

Acknowledgments This work is supported by the National Science Foundation Grant DMS-1916161 of Majid Mojirsheibani.

1 Introduction

In this article we consider the problem of estimating a regression function where the main interest and focus is to apply our results to classification problems when the covariates vectors in the data are not necessarily fully observable. Our proposed approach is based on local averaging techniques and, in particular, involves histogram estimators. Since the early 90's, there has been a growing interest in developing methods to tackle the presence of incomplete data in estimation and inference. Although the great majority of the existing

[1]Email: chieh-wei.han.96@my.csun.edu
[2]Corresponding author: Email: majid.mojirsheibani@csun.edu.

literature deals mainly with missingness in response variables, there have also been several results dealing with missing covariate components (which is the setup of this paper). These include Chen et al. (2016) who proposed an estimating equation method for logistic partially linear models with missing covariates, Liu and Yuan (2016) who considered the estimation of conditional quantiles with some covariates missing at random, and the results of Lukusa, et al. (2016) on Poisson regression. Bravo (2015) considered the estimation of a general class of semi-parametric models where the nonparametric component of the model is computed iteratively using local linear estimation. Sinha et al. (2014) proposed semi-parametric estimators for the parameters in a parametric regression model with missing covariates, and Hu et al. (2014) considered a two-stage multiple imputation approach for nonparametric estimation in quantile regression. Guo et al. (2014) considered the estimation of a semi-parametric multi-index model using a weighted estimating equation approach. Lee, et al. (2012) considered logistic regression models with missing covariates and outcome. Efromovich (2012) dealt with adaptive orthogonal series estimators when the regression function belongs to a Sobolev class. Wu and Wu (2007) studied generalized linear mixed models with missing covariates. Liang et al. (2004) proposed estimators in partially linear models with missing covariates, whereas Chen (2004) considered consistent maximum likelihood estimation of the parameters of a regression function. Earlier results along these lines include Robins, et al. (1994) as well as Lipsitz and Ibrahim (1996).

Virtually all of the results obtained by these authors are based on the assumption that the data are missing at random, which is also used in this paper; this assumption will be formally defined and addressed in the next section. Our results in this paper are fully nonparametric in that the form of the underlying regression function is completely unknown. Our contributions may be summarized as follows. In Section 2 we propose a histogram estimator of the regression function that takes into account the fact that some of the covariates are not fully observable. We also derive exponential bounds on the L_1 norms of the proposed estimators; however, our results readily extend to general L_p norms. These findings yield various convergence results (and the strong consistency) of the proposed estimators, but more importantly, they can be used to perform statistical classification (nonparametrically) with missing covariates. In fact, in Section 3 we consider the problem of classification and pattern recognition with incomplete covariates and construct histogram classifiers that are asymptotically optimal. To assess the finite-sample performance of our proposed estimators

and classifiers, we provide some numerical work in Section 4. All proofs are deferred to Section 5.

2 Main results

2.1 Histogram estimates of a regression function

Let $(\mathbf{X}, Y)$ be a $\mathbb{R}^s \times \mathbb{R}$-valued random vector, where $s \geq 1$, and consider the problem of estimating the regression function $m(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$ based on a random sample (the data) $\mathbb{D}_n = \{(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)\}$, where the $(\mathbf{X}_i, Y_i)$'s are independently and identically distributed (i.i.d) random vectors with the same distribution as $(\mathbf{X}, Y)$. Let $\mathcal{P}_n = \{A_{n1}, A_{n2}, \ldots\}$ be a partition of $\mathbb{R}^s$ into cubes of length $b_n > 0$, i.e., sets of the form $\bigtimes_{i=1}^{s} \big[k_i b_n, (k_i+1)b_n\big)$, where k_i's are integers. For every $\mathbf{x} \in \mathbb{R}^s$, let $A_n(\mathbf{x})$ denote the unique cell of $\mathcal{P}_n$ that contains the point $\mathbf{x}$. Cubic histogram estimates work by taking the average of those Y_j's whose corresponding $\mathbf{X}_j$'s fall in the cell $A_n(\mathbf{x})$, and thus they are local averaging estimators. More precisely, when the data are fully observable, the histogram estimator of the regression function $m(\mathbf{x})$ is defined by

$$m_n(\mathbf{x}) = \frac{\sum_{j=1}^{n} Y_j \, I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}{\sum_{j=1}^{n} I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}, \tag{1}$$

with the convention that $0/0 = 0$. As for the performance of the estimator in (1), let μ be the probability measure of $\mathbf{X}$. Then, by a classical result of Devroye and Györfi (1983), one has the strong consistency property (in L_2) that $\lim_{n\to\infty} \int (m_n(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) =^{\text{a.s.}} 0$, under the *shrinking cell* condition

$$b_n \to 0, \text{ as } n \to \infty, \text{ with } nb_n^s \to \infty, \tag{2}$$

and Y is bounded. Györfi (1991) also considers a slightly revised version of (1) which is strongly consistent without any boundedness assumption on the square integrable response variable; see Györfi et al. (2002, Ch. 23) for further detail. In passing we note that the histogram estimator in (1) is a *local averaging* estimator in the sense that it is of the form $m_n(\mathbf{x}) = \sum_{j=1}^{n} W_{n,j}(\mathbf{x}) \cdot Y_j$ with weights $W_{n,j}(\mathbf{x}) = I\{\mathbf{X}_j \in A_n(\mathbf{x})\} / \sum_{j=1}^{n} I\{\mathbf{X}_j \in A_n(\mathbf{x})\}$. Such estimators are quite popular in nonparametric estimation and also include the nearest neighbor as well as kernel estimators.

2.2 Missing covariates

In this section we consider the case where some components of the covariate vector $\mathbf{X}$ may be unavailable (missing). More precisely, for $j = 1, \ldots, n$, let $\mathbf{X}'_j = (\mathbf{U}'_j, \mathbf{V}'_j) \in \mathbb{R}^{d+p}$, where $\mathbf{U}_j \in \mathbb{R}^d$, $d \geq 1$, is always observable, but $\mathbf{V}_j \in \mathbb{R}^p$, $p \geq 1$, may be missing. Clearly, the estimator in (1) is no longer available because some of the $\mathbf{V}_j$'s may be missing. In order to revise (1) accordingly, we start by defining the independent Bernoulli random variables $\delta_1, \ldots, \delta_n$, where $\delta_j = 1$ if $\mathbf{V}_j$ is not missing, and $\delta_j = 0$ otherwise. Then the data may be represented by

$$\mathbb{D}_n = \{(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)\} = \{(\mathbf{U}_1, \mathbf{V}_1, Y_1, \delta_1), \cdots, (\mathbf{U}_n, \mathbf{V}_n, Y_n, \delta_n)\}.$$

We also need to take into account the missing probability mechanism (i.e., the selection property), which is the quantity $P\{\delta = 1|\mathbf{X}, Y\} = E(\delta|\mathbf{X}, Y)$. If the missing probability mechanism satisfies $P\left\{\delta = 1\middle|\mathbf{X}, Y\right\} = P\{\delta = 1\} = E(\delta)$ then we say $\mathbf{V}$ is *Missing Completely at Random* (MCAR). However, in practice, the MCAR assumption is rather unrealistic and restrictive. A more widely used assumption in the literature is the *Missingness at Random* (MAR) assumption, which amounts to

$$P\left\{\delta = 1\middle|\mathbf{X}, Y\right\} = P\left\{\delta = 1\middle|\mathbf{U}, Y\right\}, \text{ where } \mathbf{X}' = (\mathbf{U}', \mathbf{V}') \in \mathbb{R}^{d+p}, \tag{3}$$

i.e., the probability that $\mathbf{V}$ is missing does not depend on $\mathbf{V}$ itself. For a detailed account of these and other missing patterns one can refer, for example, to Little and Rubin (2002). It is straightforward to see that when the missing probability satisfies the MCAR assumption, one can just use the complete cases to estimate $m(\mathbf{x})$, where a complete case refers to a fully observable $\mathbf{X}_j$ (i.e., when $\delta_j = 1$). In other words, in this case the correct estimator is given by

$$\widetilde{m}_n(\mathbf{x}) = \frac{\sum_{j=1}^n \delta_j Y_j I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}{\sum_{j=1}^n \delta_j I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}. \tag{4}$$

To appreciate this, let $\widetilde{m}_{1,n}(\mathbf{x}) = \sum_{j=1}^n \delta_j I\{\mathbf{X}_j \in A_n(\mathbf{x})\} / \sum_{j=1}^n I\{\mathbf{X}_j \in A_n(\mathbf{x})\}$ and $\widetilde{m}_{2,n}(\mathbf{x}) = \sum_{j=1}^n \delta_j Y_j I\{\mathbf{X}_j \in A_n(\mathbf{x})\} / \sum_{j=1}^n I\{\mathbf{X}_j \in A_n(\mathbf{x})\}$, with the convention that $0/0 = 0$. Then the simple estimator in (4) can be written as the ratio, $\widetilde{m}_{2,n}(\mathbf{x})/\widetilde{m}_{1,n}(\mathbf{x})$. Now, since $\widetilde{m}_{2,n}(\mathbf{x})$ and $\widetilde{m}_{1,n}(\mathbf{x})$ are just the histogram estimators of $E(\delta Y|\mathbf{X} = \mathbf{x})$ and $E(\delta|\mathbf{X} = \mathbf{x})$, respectively, and since $E(\delta Y|\mathbf{X})/E(\delta|\mathbf{X}) = m(\mathbf{X})$ holds under the MCAR assumption, the ratio in (4) is indeed a correct estimator of $m(\mathbf{x})$. In fact more is true: it is a simple exercise to show that in this case, the simple estimator in (4) is strongly consistent in the sense that

$\lim_{n\to\infty}\int(\widetilde{m}_n(\mathbf{x})-m(\mathbf{x}))^2\mu(d\mathbf{x})=^{\text{a.s.}}0$, under the same conditions that render (1) consistent. Of course, in general, (4) is not the correct estimator of $m(\mathbf{x})$ because the unrealistic MCAR assumption may not hold in practice. In the next section we will focus on estimators that relax this assumption.

2.3 The proposed histogram estimator

In this section we propose revised versions of (4) that take into account the MAR assumption. One common approach to handle the presence of the missing cases is by weighting the complete cases by the inverse of the probability that $\mathbf{V}$ is missing, i.e., $\pi(\mathbf{U},Y):=P\{\delta=1|\mathbf{U},Y\}$ (or its estimator, if the function π is unknown). This approach, which is originally due to Horvitz and Thompson (1952), has been used in the literature on the analysis of incomplete data extensively. See, for example, Lukusa, et al. (2016) and Robins, et al. (1994). To motivate our approach, first consider the simple but unrealistic case where the missing probability function $\pi(\mathbf{U},Y)=P\{\delta=1|\mathbf{U},Y\}$ is completely known (as a function of $\mathbf{U}$ and Y). In this case, our proposed estimator of $m(\mathbf{x})$ is

$$\overline{m}_n(\mathbf{x})=\frac{\sum_{j=1}^n\frac{\delta_jY_j}{\pi(\mathbf{U}_j,Y_j)}I\left\{\mathbf{X}_j\in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n\frac{\delta_j}{\pi(\mathbf{U}_j,Y_j)}I\left\{\mathbf{X}_j\in A_n(\mathbf{x})\right\}} \tag{5}$$

Observe that if we define

$$\overline{m}_{1,n}(\mathbf{x}):=\frac{\sum_{j=1}^n\frac{\delta_j}{\pi(\mathbf{U}_j,Y_j)}I\left\{\mathbf{X}_j\in A_n(\mathbf{x})\right\}}{\sum_{j=1}^nI\left\{\mathbf{X}_j\in A_n(\mathbf{x})\right\}}\ ,\quad \overline{m}_{2,n}(\mathbf{x}):=\frac{\sum_{j=1}^n\frac{\delta_jY_j}{\pi(\mathbf{U}_j,Y_j)}I\left\{\mathbf{X}_j\in A_n(\mathbf{x})\right\}}{\sum_{j=1}^nI\left\{\mathbf{X}_j\in A_n(\mathbf{x})\right\}}, \tag{6}$$

then (5) can be written as $\overline{m}_n(\mathbf{x})=\overline{m}_{2,n}(\mathbf{x})/\,\overline{m}_{1,n}(\mathbf{x})$. But $\overline{m}_{2,n}$ and $\overline{m}_{1,n}$ are the histogram estimators of $E[\delta Y/\pi(\mathbf{U},Y)|\mathbf{X}]$ and $E[\delta/\pi(\mathbf{U},Y)|\mathbf{X}]$, respectively. Furthermore, under the MAR assumption, $E[\delta Y/\pi(\mathbf{U},Y)|\mathbf{X}]=E(Y|\mathbf{X})$ and $E[\delta/\pi(\mathbf{U},Y)|\mathbf{X}]=E(1|\mathbf{X})=1$. Therefore, the estimator $\overline{m}_n(\mathbf{x})$ in (5) can be viewed as the histogram estimator of the regression function $E(Y|\mathbf{X}=\mathbf{x})=:m(\mathbf{x})$.

In practice, the regression estimator $\overline{m}_n(\mathbf{x})$ is not available because the function $\pi(\mathbf{u},y)=P\left\{\delta=1\middle|\mathbf{U}=\mathbf{u},Y=y\right\}$ that appears in (5) is almost always unknown and has to be estimated. In what follows, we consider two estimators of $\pi(\mathbf{u},y)$; the first one is based on kernel regression, whereas the second approach is based on the least-squares method.

2.3.1 A local averaging estimator of the selection probability $\pi(\cdot\,,\cdot)$

Let $\mathbf{Z}^{'} = (\mathbf{U}^{'}, Y')$ and consider the kernel regression estimator of $\pi(\mathbf{U}_j, Y_j) = E(\delta_j|\mathbf{U}_j, Y_j)$, given by

$$\widehat{\pi}(\mathbf{U}_j, Y_j) = \widehat{\pi}(\mathbf{Z}_j) = \frac{\sum_{k=1,\neq j}^{n} \delta_j \mathcal{H}(\frac{\mathbf{Z}_k - \mathbf{Z}_j}{h_n})}{\sum_{k=1,\neq j}^{n} \mathcal{H}(\frac{\mathbf{Z}_k - \mathbf{Z}_j}{h_n})} \tag{7}$$

with the convention $0/0 = 0$, where the function $\mathcal{H} : \mathbb{R}^{d+1} \to \mathbb{R}_+$ is the kernel used with the smoothing parameter h_n ($h_n \to 0$, as $n \to \infty$). Here the choice of the kernel is at the discretion of the practitioner. If Y is a discrete random variable taking values in a set $\mathcal{Y} = \{y_1, y_2, \dots\}$, we consider the following kernel-type estimator of π,

$$\widehat{\pi}(\mathbf{U}_j, Y_j) = \frac{\sum_{k=1,\neq j}^{n} \delta_k I\{Y_k = Y_j\} \mathcal{K}(\frac{\mathbf{U}_k - \mathbf{U}_j}{h_n})}{\sum_{k=1,\neq j}^{n} I\{Y_k = Y_j\} \mathcal{K}(\frac{\mathbf{U}_k - \mathbf{U}_j}{h_n})}\,, \tag{8}$$

where $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}_+$ is the kernel with the smoothing parameter h_n. Now consider the following revised version of (5):

$$\widehat{m}_n(\mathbf{x}) = \frac{\sum_{j=1}^{n} \frac{\delta_j}{\widehat{\pi}(\mathbf{U}_j, Y_j)} Y_j I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^{n} \frac{\delta_j}{\widehat{\pi}(\mathbf{U}_j, Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}\,, \tag{9}$$

where $\widehat{\pi}(\mathbf{U}_j, Y_j)$ can be taken to be either (7) or (8), depending on whether Y has a continuous or a discrete distribution. To assess the performance of $\widehat{m}_n(\mathbf{x})$ in (9), we first need to state a number of conditions:

(A1) The *shrinking cell* condition (2) holds, where $s = d + p$.

(A2) $\pi_{\min} := \inf_{\mathbf{u},y} \pi(\mathbf{u}, y) > 0$, where $\pi(\mathbf{u}, y) = P\{\delta = 1|\mathbf{U} = \mathbf{u}, Y = y\}$.

(A3) The kernel $\mathcal{K}$ is a probability density function and satisfies $\int |w_j| \mathcal{K}(\mathbf{w}) d\mathbf{w} < \infty\,, j = 1\,, \cdots\,, d$, and $||\mathcal{K}||_\infty < \infty$. Furthermore, the smoothing parameter h_n satisfies $h_n \to 0$ and $n h_n^d \to \infty$, as $n \to \infty$.

(A4) The random vector $\mathbf{U}$ has a compactly supported probability density function $f(\mathbf{u}) = \sum_{y \in \mathcal{Y}} p_y f_y(\mathbf{u})$ and is bounded away from zero on its support, where $p_y = P(Y = y)$, and $f_y(\mathbf{u})$ is the conditional density of $\mathbf{U}$ given $Y = y$. Furthermore, f is uniformly bounded on its support and its first-order partial derivatives are bounded on the interior of the support.

(A5) The partial derivatives $\frac{\partial}{\partial u_j} \pi(\mathbf{u}, y), j = 1\,, \cdots\,, d$ exist and are bounded on the compact support of f, uniformly in $\mathbf{u}$.

Here, condition (A2) essentially states that $\mathbf{V}$ can be observed with a non-zero probability for all $\mathbf{u}$ and y. Condition (A3) is not restrictive since the choice of the kernel is at our discretion, whereas condition (A4) is often imposed in nonparametric regression in order to avoid having unstable estimates in the tails of the pdf f of $\mathbf{U}$. Condition (A5) is technical. The following result gives bounds on the performance of the estimator $\widehat{m}_n$ in (9) with $\widehat{\pi}(\mathbf{U}_j, Y_j)$ estimated via (8).

Theorem 1 *Let $\widehat{m}_n$ be the estimator defined in (9) with $\widehat{\pi}(\mathbf{U}_j, Y_j)$ given by (8). Suppose that conditions (A1)–(A5) hold and that Y is a bounded random variable. Then for every $\epsilon > 0$, there is an $n_0 > 0$ such that for all $n > n_0$,*

$$P\left\{\int \left|\widehat{m}_n(\mathbf{x}) - m(\mathbf{x})\right|\mu(d\mathbf{x}) > \epsilon\right\} \leq 6\exp(-c_1 n\epsilon^2) + 16n\exp(-c_2 n h_n^d) + 8n\exp(-c_3 n h_n^d \epsilon^2)$$

where c_1, c_2, and c_3 are positive constants not depending on n or ϵ.

Remark 1 *It is straightforward to note that Theorem 1 continues to hold for general L_p norms ($1 \leq p < \infty$) of $\widehat{m}_n(\mathbf{x})$ with the a bound of the form*

$$\begin{aligned} P\left\{\int \left|\widehat{m}_n(\mathbf{x}) - m(\mathbf{x})\right|^p \mu(d\mathbf{x}) > \epsilon\right\} \quad &\leq \quad 6\exp(-c_1 k_1^2 n\epsilon^2) + 16n\exp(-c_3 n h_n^d) \\ &\quad + 8n\exp(-c_3 k_1^2 n h_n^d \epsilon^2)\,, \end{aligned}$$

where $k_1 = 1/(2L)^{p-1}$ and the constants c_1, c_2, and c_3 are as in Theorem 1. Therefore, in view of the Borel-Cantelli lemma,

$$\int |\widehat{m}_n(\mathbf{x}) - m(\mathbf{x})|^p \mu(d\mathbf{x}) \longrightarrow^{a.s.} 0 \quad \textit{whenever } (n h_n^d)^{-1} \log n \to 0.$$

If Y is a continuous random variable then we use (7) for the term $\widehat{\pi}(\mathbf{U}_j, Y_j)$ in the definition of the estimator $\widehat{m}_n(\mathbf{x})$ in (9). In this case the conclusion of Theorem 1 continues to hold with d replaced by $d+1$ and different constants c_1, c_2, and c_3 provided that conditions (A3), (A4), and (A5) are slightly revised as follows:

(A3′) The kernel $\mathcal{H}$ in (7) is a probability density function satisfying $\int |w_j|\mathcal{H}(\mathbf{w})d\mathbf{w} < \infty$, $j = 1\,, \cdots, d+1\,$, and $||\mathcal{H}||_\infty < \infty$. Furthermore, $h_n \to 0$ and $n h_n^{d+1} \to \infty$, as $n \to \infty$.

(A4′) The random vector $\mathbf{Z}' = (\mathbf{U}', Y)$ has a compactly supported probability density function, $f(\mathbf{z})$, which is bounded away from zero on its support. Furthermore, f and its first order partial derivatives are uniformly bounded on the support of f.

(A5′) The partial derivatives $\frac{\partial}{\partial z_j}\pi(\mathbf{z})$, $j = 1, \cdots, (d+1)$, exist and are bounded on the compact support of f, uniformly in $\mathbf{z}$.

Remark 2 *When* $\mathbf{U}$ *is high-dimensional, one has to find ways to counter the curse of dimensionality from which a kernel estimator can suffer in the sense of having slower rates of convergence. Here, PCA appears to be a popular dimension reduction technique for classification with high-dimensional covariates.*

2.3.2 The least-squares estimator of the selection probability $\pi(\cdot\,,\cdot)$

Our second approach to estimate the selection probability $\pi(\mathbf{U}, Y) := P\{\delta = 1|\mathbf{U}, Y\} = E(\delta|\mathbf{U}, Y)$ uses the least-squares method. More specifically, suppose that the function π belongs to a given known class $\mathscr{P}$ of functions of the form $\pi : \mathbb{R}^d \times \mathbb{R} \to [\pi_{\min}, 1]$, where $\pi_{\min} = \inf_{\mathbf{u},y} \pi(\mathbf{u}, y)$, as before. Then the least-squares estimator of π is given by

$$\widehat{\pi}_{\text{LS}} = \underset{\pi\in\mathscr{P}}{\operatorname{argmin}} \frac{1}{n}\sum_{j=1}^{n} \left(\delta_j - \pi(\mathbf{U}_j, Y_j)\right)^2 , \tag{10}$$

with the corresponding least-squares based estimator of $m(\mathbf{x})$ given by

$$\widehat{m}_{\text{LS}}(\mathbf{x}) = \frac{\sum_{j=1}^{n} \frac{\delta_j Y_j}{\widehat{\pi}_{\text{LS}}(\mathbf{U}_j, Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^{n} \frac{\delta_j}{\widehat{\pi}_{\text{LS}}(\mathbf{U}_j, Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} . \tag{11}$$

In order to study the performance of $\widehat{m}_{\text{LS}}(\mathbf{x})$ in (11), we need the following standard notation and terminology from the empirical process theory (see, for example, Pollard (1984, p. 25), or Györfi, et al (2002, p. 135)). For fix points $(\mathbf{u}_1, y_1), \cdots, (\mathbf{u}_n, y_n)$ in $\mathbb{R}^d \times \mathbb{R}$, let $\mathcal{N}_1\left(\epsilon, \mathscr{P}, (\mathbf{u}_j, y_j)_{j=1}^n\right)$ be the ϵ-covering number of the class $\mathscr{P}$ with respect to the empirical L_1 norm on the points $(\mathbf{u}_1, y_1), \cdots, (\mathbf{u}_n, y_n)$, i.e., $\mathcal{N}_1\left(\epsilon, \mathscr{P}, (\mathbf{u}_j, y_j)_{j=1}^n\right)$ is the cardinality of the smallest subclass of functions $\left\{\pi_1, \cdots, \pi_N : \mathbb{R}^d \times \mathbb{R} \to [\pi_{\min}, 1]\right\}$ with the property that for every $\pi \in \mathscr{P}$ and every $\epsilon > 0$, one has $\min_{1\le k\le N} \frac{1}{n}\sum_{j=1}^{n} |\pi(\mathbf{u}_j, y_j) - \pi_k(\mathbf{u}_j, y_j)| < \epsilon$. Then, with this notation, we have the following result on the performance of $\widehat{m}_{\text{LS}}$

Theorem 2 *Let* $\widehat{m}_{LS}$ *be as in (11). Suppose that conditions (A1) and (A2) hold and* Y *is bounded. Then, for every* $\epsilon > 0$*, there is an* $n_0 > 0$ *such that for all* $n > n_0$*,*

$$\begin{aligned} & P\left\{\int |\widehat{m}_{LS}(\mathbf{x}) - m(\mathbf{x})|\, \mu(d\mathbf{x}) > \epsilon\right\} \\ & \quad \le\ 6\exp(-C_1 n\epsilon^2) + 16E\left[\mathcal{N}_1\left(a_1\epsilon, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n\right)\right]\exp(-C_6 n\epsilon^2) \\ & \qquad + 16E\left[\mathcal{N}_1\left(a_2\,\epsilon^2, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n\right)\right]\exp(-C_7 n\epsilon^2)\,, \end{aligned}$$

where a_1*,* a_2*,* C_1, C_6*, and* C_7 *are positive constants not depending on* n *or* ϵ*.*

Remark 3 *Theorem 2 can be used to establish the strong consistency of* $\widehat{m}_{LS}$*: Let* $C_\epsilon = \min\left(a_1\epsilon\,,\, a_2\epsilon^2\right)$*. If* $n^{-1}\log\left(E\left[\mathcal{N}_1\left(C_\epsilon, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n\right)\right]\right) \to 0$*, for all* $\epsilon > 0$*, then the bound in Theorem 2 together with the Borel-Cantelli lemma yield the almost-sure convergence result* $\int |\widehat{m}_{LS}(\mathbf{x}) - m(\mathbf{x})|\mu(d\mathbf{x}) \to^{a.s.} 0$*. In fact, as in Remark 1, it is straightforward to see that the above results can be readily extended to general* L_p *norms* $\int |\widehat{m}_{LS}(\mathbf{x}) - m(\mathbf{x})|^p \mu(d\mathbf{x})$ *of the estimator in (11) for all* $1 \leq p < \infty$*.*

3 Applications to problems in pattern recognition and classification

In this section we consider an application of the results developed in the previous section to the problem of pattern recognition and statistical classification. More specifically, let $(\mathbf{X}, Y)$ be an $\mathbb{R}^s \times \{1, \ldots, M\}$-valued random pair. The problem of statistical classification involves the prediction of the class variable, Y, based on the covariate vector $\mathbf{X}$. In practice, one wants to find a classifier, i.e., a function of the form $\Psi : \mathbb{R}^s \longrightarrow \{1, \ldots, M\}$, for which the probability of misclassification, $L(\Psi) := P\{\Psi(\mathbf{X}) \neq Y\}$, is as small as possible. To present the optimal classifier, let $P_k(\mathbf{x}) := P\{Y = k \mid \mathbf{X} = \mathbf{x}\}$ be the class conditional probability corresponding to class $k \in \{1, 2, \ldots, M\}$. The classifier with the lowest misclassification probability is given by the function $\Psi_B(\mathbf{x})$ which assigns $\mathbf{x}$ to class $k \in \{1, 2, \ldots, M\}$ if $\max_{1 \leq i \leq M} P_i(\mathbf{x}) = P_k(\mathbf{x})$. More specifically, $\Psi_B(\mathbf{x})$ satisfies $P_{\Psi_B(\mathbf{x})}(\mathbf{x}) = \max_i P_i(\mathbf{x})$; see, for example, Devroye and Györfi (1985, p. 253). The theoretically best classifier Ψ_B is almost always unknown because it depends on the underlying distribution of $(\mathbf{X}, Y)$ which is unknown and, therefore, one has to use the data to construct a classifier.

Given a random sample $\mathbb{D}_n = \left\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\right\}$, one tries to construct a sample based classifier Ψ_n in such a way that its misclasfication error, $L_n(\Psi_n) = P\left\{\Psi_n(\mathbf{X}) \neq Y \mid D_n\right\}$, is in some sense as small as possible. Let $L(\Psi_B) = P\left\{\Psi_B(\mathbf{X}) \neq Y\right\}$. The classifier Ψ_n is said to be strongly consistent if $L_n(\Psi_n) \to^{\text{a.s.}} L(\Psi_B)$. If the convergence holds in probability, Ψ_n is said to be weakly consistent. We also note that, by the dominated convergence theorem, if $P\{\Psi_n(\mathbf{X}) \neq Y | D_n\} \to^{\text{a.s.}} L(\Psi_B)$ then $P\left\{\Psi_n(\mathbf{X}) \neq Y\right\} \to L(\Psi_B)$. To estimate the Bayes classifier Ψ_B, we consider a plug-in estimator that works by replacing each conditional probability $P_k(\mathbf{x}) := P\{Y = k \mid \mathbf{X} = \mathbf{x}\}$ by an estimator (function of the data) $P_{k,n}(\mathbf{x})$. The

resulting classifier, Ψ_n, is defined by

$$P_{\Psi_n(\mathbf{x}),n}(\mathbf{x}) = \max_{1\le k\le M} P_{k,n}(\mathbf{x}), \tag{12}$$

i.e., $\Psi_n(\mathbf{x})$ assigns $\mathbf{x}$ to class $k \in \{1, 2, \ldots, M\}$ if $\max_{1\le i\le M} P_{i,n}(\mathbf{x}) = P_{k,n}(\mathbf{x})$. To study the performance of $\Psi_n(\mathbf{x})$ defined via (12), we first state the following standard result (see, for example, Devroye and Györfi, (1985, p. 254)): $0 \le L_n(\Psi_n) - L(\Psi_B) \le \sum_{i=1}^{M} \int |P_i(\mathbf{x}) - P_{i,n}(\mathbf{x})|\mu(d\mathbf{x})$. Therefore, the plug-in estimator of Ψ_B is strongly consistent whenever $\int |P_i(\mathbf{x}) - P_{i,n}(\mathbf{x})|\mu(d\mathbf{x}) \to^{\text{a.s.}} 0$, for each $i = 1, \ldots, M$. Here, $P_{i,n}(\mathbf{x})$ is just an estimator of the regression function $P_i(\mathbf{x}) = E\big[I\{Y = i\} \mid \mathbf{X} = \mathbf{x}\big]$, except that in our case there are missing covariates in $\mathbf{X}$. More specifically, let $\mathbf{X}' = (\mathbf{U}', \mathbf{V}')$, where $\mathbf{U} \in \mathbb{R}^p$ is always observable, but $\mathbf{V} \in \mathbb{R}^d$ may be missing. We can represent the data by $\mathbb{D}_n = \{(\mathbf{U}_1, \mathbf{V}_1, Y_1, \delta_1), \ldots, (\mathbf{U}_n, \mathbf{V}_n, Y_n, \delta_n)\}$, where $\delta_j = 0$ if $\mathbf{V}_j$ is missing, and $\delta_j = 1$ otherwise. Therefore, based on our earlier results, we can use the following histogram estimator of $P_k(\mathbf{x}), k = 1, \ldots, M$,

$$\widehat{P}_{k,n}(\mathbf{x}) = \frac{\sum_{j=1}^{n} \dfrac{\delta_j \cdot I\{Y_j = k\}}{\widehat{\pi}(\mathbf{U}_j, Y_j)} I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}{\sum_{j=1}^{n} \dfrac{\delta_j}{\widehat{\pi}(\mathbf{U}_j, Y_j)} I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}, \tag{13}$$

where one can take $\widehat{\pi}$ to be either the kernel-type estimator in (8), or, if it is known that $\pi \in \mathscr{P}$ (where $\mathscr{P}$ is a known class), then one may use the least-square estimator given by (10). Now, in view of (12), we propose the classifier $\widehat{\Psi}_n$ which is defined via

$$\widehat{P}_{\widehat{\Psi}_n(\mathbf{x}),n}(\mathbf{x}) = \max_{1\le k\le M} \widehat{P}_{k,n}(\mathbf{x}), \tag{14}$$

i.e., $\widehat{\Psi}_n(\mathbf{x})$ assigns $\mathbf{x}$ to class $k \in \{1, 2, \ldots, M\}$ if $\max_{1\le i\le M} \widehat{P}_{i,n}(\mathbf{x}) = \widehat{P}_{k,n}(\mathbf{x})$. As for the asymptotic performance of this classifier, we have the following strong consistency results. The first result corresponds to the case where π is estimated by the kernel estimator in (8).

Theorem 3 *Let $\widehat{\Psi}_n$ be the histogram classifier defined via (14) in conjunction with (13) and (8). If $(nh_n^d)^{-1} \log n \to 0$, as $n \to \infty$, then under conditions (A1)–(A5) one has $L_n(\widehat{\Psi}_n) \longrightarrow^{a.s.} L(\Psi_B)$, where $L_n(\widehat{\Psi}_n) = P\{\widehat{\Psi}_n(\mathbf{X}) \neq Y | \mathbb{D}_n\}$.*

If the selection π is estimated by the least-squares estimator (10), we have the following corresponding strong consistency result.

Theorem 4 *Let $\widehat{\Psi}_n$ be the histogram classifier defined via (14) in conjunction with (13) and (10). If $\forall c > 0$, $n^{-1} \log \left(E \left[\mathcal{N}_1 \left(c, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n \right) \right] \right) \to 0$, as $n \to \infty$, then under conditions (A1) and (A2) one has $L_n(\widehat{\Psi}_n) \longrightarrow^{a.s.} L(\Psi_B)$, where $L_n(\widehat{\Psi}_n) = P\{\widehat{\Psi}_n(\mathbf{X}) \neq Y | \mathbb{D}_n\}$.*

In passing we note that regarding the optimal bandwidths for Theorem 3, it is well-understood that for kernel-based classifiers the optimal bandwidth that minimizes quantities such as the MISE or ISE is not necessarily optimal in classification (in the sense of minimizing the misclassification error); see Devroye et al. (1996; Sec. 25.9). In fact, an interesting counter-example is given in Theorem 25.9 of the cited monograph, where it is shown that the optimal bandwidth based on the MISE yields a rather poor misclassification error. As argued in Chapter 25 of the cited monograph, the optimal bandwidth h_{opt} is the one that minimizes the error $L_n(\widehat{\Psi}_n)$, which is unfortunately always unknown, as is the overall error, $E[L_n(\widehat{\Psi}_n)]$; see Devroye et al. (1996; Sec. 25.1). Additionally, Hall and Kang (2005) noted that for kernel-based classification with univariate distributions and just two classes, the optimal bandwidth can be different for each class and its asymptotic magnitude can vary from terms of order $O(n^{-1/5})$ to $O(n^{-1/9})$ depending on the conditions imposed on the relationship between higher order derivatives of the marginal densities. Furthermore, their results show that in general there are no closed form expression for any one of the bandwidths. These difficulties are further compounded by the fact that finding a data-dependent bandwidth $\hat{h}_{\text{opt}}$ which is in some sense close to h_{opt} does not necessarily imply the closeness of the corresponding misclassification errors. Since, in classification, consistency is often the minimum requirement for any classifier, $\hat{h}_{\text{opt}}$ must be chosen in such a way that the resulting classifier will be consistent (either weakly or strongly); see Devroye et al. (1996; p. 424). To that end, a number of methods have been proposed in the literature for finding data-dependent bandwidths that yield the minimum requirements; see Devroye et al. (1996; Ch. 25).

4 Numerical results

Here, we carry out some numerical studies in order to assess the performance of the following estimators in both classification and regression setups: (i) the estimator $\widehat{m}_n(\mathbf{x})$ defined via (9), (ii) the estimator $\widehat{m}_{\text{LS}}(\mathbf{x})$ defined by (11), and (iii) the complete case estimator $\widetilde{m}_n(\mathbf{x})$ in (4) that discards all of the incomplete covariates. Our examples show that the proposed estimators can perform well in the sense of having lower error rates.

Example (A) *[Simulated data.]*
In what follows, we consider three different models to generate our data from. These are of the form

$$Y = m_k(\mathbf{X}) + \epsilon, \quad k = 1, 2, 3, \text{ where } \epsilon \sim N(0, 0.5), \text{ and } \epsilon \text{ is independent of } \mathbf{X}.$$

Here $\mathbf{X} = (X_1, X_2)'$ in Model (I) and $\mathbf{X} = (X_1, X_2, X_3, X_4)'$ for Models (II) and (III), where

$$\begin{aligned}
m_1(\mathbf{X}) &= X_1X_2 + X_2^2 && \textit{Model (I)}\\
m_2(\mathbf{X}) &= -\sin(2X_1) + X_2^2 + X_3 - \exp(-X_4) && \textit{Model (II)}\\
m_3(\mathbf{X}) &= X_1 + (2X_2 - 1)^2 + \frac{\sin(2\pi X_3)}{2 - \sin(2\pi X_3)} + \sin(2\pi X_4) + 2\cos(2\pi X_4) && \textit{Model (III)}\\
&\quad + 3\sin^2(2\pi X_4) + 4\cos^2(2\pi X_4).
\end{aligned}$$

In all the above models, $\mathbf{X}$ has a multivariate normal distribution with mean zero and a covariance matrix whose (ij)-th component is equal to $2^{-|i-j|}$, $i, j \geq 1$. Next a sample of sizes n=150 was drawn from each of these models. In passing we also note that models 2 and 3 are similar to those of Meier et al. (2009), where as Model (I) is essentially a toy example. As for the missing covariates, in Model (I) the variable X_2 is allowed to be missing at random according to the logistic selection probability

$$\pi(X_1, Y) := P\{\delta = 1|X_1, Y\} = \exp(a_0 + a_1X_1 + a_2Y)/[1 + \exp(a_0 + a_1X_1 + a_2Y)],$$

with $(a_0, a_1, a_2) = (1, 0.2, -0.5)$. However, in models II and III, both X_3 and X_4 are allowed to be missing at random according to the logistic selection probability

$$\pi(X_1, X_2, Y) = \exp(b_0 + b_1X_1 + b_2X_2 + b_3Y)/[1 + \exp(b_0 + b_1X_1 + b_2X_2 + b_3Y)],$$

where (b_0, b_1, b_2, b_3) is equal to $(0.1, -0.2, 1, 0.2)$ in Model (II) and it is $(0.8, 0.2, 0.2, -0.1)$ in Model (III). Our choice of the numerical values of the coefficients (a_0, a_1, a_2) and (b_0, b_1, b_2, b_3) yield 50% missing data (approximately) for each case. Next, the cross-validation approach of Racine and Li (2004), which is available from the 'R' package "np" (see Racine and Hayfield 2008), was employed to compute the kernel estimator of the selection probabilities in (7). Similarly, to find the least squares estimators of the parameters of the logistic selection probabilities, we employed the nonlinear least squares package in 'R' called "nls2". To construct the histogram regression estimators $\widehat{m}_n(\mathbf{x})$, $\widehat{m}_{\text{LS}}(\mathbf{x})$, and $\widetilde{m}_n(\mathbf{x})$, the leave-one-out cross-validation was used to select the cube length, b_n (see (2)), from the equally-spaced grid

$\{0.05, 0.10, \ldots, 0.95, 1.00\}$ that minimized the empirical mean-squared error. Our initial pilot study shows that increasing the upper limit of the grid from 1.00 to values as large as 2 or even 3 does not change the results. Finally, the empirical L_p errors ($p = 1, 2$) were computed for each method. The entire numerical work above was repeated 300 times, each time using a sample of size n=150. Rows 1, 2, and 3 of Table 1 summarize the average empirical L_p errors ($p = 1, 2$), over 300 runs, along with their standard errors (in parentheses) for the case of logistic missingness mechanism.

Table 1 goes here.

Rows 1, 2, and 3 of Table 1 show that the estimators $\widehat{m}_n(\mathbf{x})$ and $\widehat{m}_{\text{LS}}(\mathbf{x})$ both have the ability to outperform the estimator $\widetilde{m}_n(\mathbf{x})$ which uses the complete cases only. Additionally, a comparison of rows 2 and 3 shows that $\widehat{m}_{\text{LS}}(\mathbf{x})$ that uses the least-squares estimator of the selection probabilities has lower error rates than $\widehat{m}_n(\mathbf{x})$ which uses kernel regression to estimate the selection probabilities. This is to be expected because we are assuming that we know the exact functional form of the underlying selection probability (which is logistic here).

In addition to the logistic selection probabilities discussed above, we have also considered some highly nonlinear trigonometric functions. More specifically, in Model (I), once again we allowed X_2 to be missing at random. However, instead of logistic, we consider the nonlinear selection probability

$$\pi(X_1, Y) := P\{\delta = 1 | X_1, Y\} = |\cos(\exp(0.6Y) - 0.1 \sin(-2X_1Y + Y^2))|.$$

Similarly, in models (II) and (III), once again X_3 and X_4 may be missing at random; the MAR selection probability is taken to be

$$\pi(X_1, X_2, Y) := P\{\delta = 1 | X_1, X_2, Y\} = 0.8\,|\cos(X_1 + X_2 - Y - 2\sin(X_1X_2Y))|.$$

These selection probabilities yield approximately 50% missing data in each of the 3 models. The corresponding results based on 300 Monte Carlo runs appear in rows 4, 5, and 6 of Table 1. These rows show that the estimator $\widehat{m}_n(\mathbf{x})$ continues to have lower error rates than the complete case estimator $\widetilde{m}_n(\mathbf{x})$. However, the estimator $\widehat{m}_{\text{LS}}(\mathbf{x})$ fails to outperform $\widetilde{m}_n(\mathbf{x})$ because of the obvious fact that the selection probabilities are no longer logistic (they are trigonometric) and thus $\widehat{m}_{\text{LS}}(\mathbf{x})$ is not even consistent. The reason for including this

comparison here is that, in practice, many practitioners tend to assume the popular logistic selection probability when, in fact, it does not hold. Rows numbered 7, 8, and 9 in Table 1 correspond to the case where the covariates are missing completely at random (MCAR) with

$$\pi(\mathbf{x}, y) := P\{\delta = 1 | \mathbf{X} = \mathbf{x}, Y = y\} = P\{\delta = 1\} = 0.5$$

for all three models. As rows 7, 8, and 9 show, although $\widehat{m}_n(\mathbf{x})$ can typically perform better than $\widehat{m}_{\text{LS}}(\mathbf{x})$, none of the estimators is uniformly better than the other ones. This is not surprising because, under the MCAR assumption, even the estimator $\widetilde{m}_n(\mathbf{x})$ is strongly convergent.

Example (B) *[Real dataset: Pima Indian Diabetes and classification.]*
This data set involves 768 patients, 268 of whom have "tested positively for diabetes", which are labeled as being in class 1, and the remaining 500 patients are in class 0. There are also eight numeric-valued covariates measured on each patient. A full description of this data set can be found at the University of California Irvine, repository of machine learning databases. A close examination of this data set shows that many of the variables are reported to be zero, some of which may be viewed as missing. Here, we focus on one dominant missing pattern where the variables Triceps skin fold thickness and 2-Hour serum insulin are, jointly, reported to be zero for 227 patients. Here, we consider the classification of a patients diabetes status, i.e., class 0 or class 1, based on the available covariates. The proposed classifier used is of the form $\widehat{\Psi}_n(\mathbf{x})$, defined via (14) and (13), where the estimated selection probability $\widehat{\pi}(\cdot, \cdot)$ in (13) can be either (8) or (10). More specifically, depending of whether (8) or (10) is used to estimate $\pi(\cdot, \cdot)$, the proposed histogram classifier will be denoted by $\widehat{\Psi}_n(\mathbf{x})$ and $\widehat{\Psi}_{\text{LS}}(\mathbf{x})$, respectively. The complete case classifier will be denoted by $\widetilde{\Psi}_n(\mathbf{x})$. To estimate the misclassification error of various classifiers, two different procedures are employed: (i) the *resubstitution* method, also called the *apparent error rate*, i.e., the approach based on the error committed on the data itself, and (ii) the leave-one-out cross-validation method. The results are summarized in Table 2.

Table 2 goes here.

Table 2 shows that the resubstitution-based estimates of the error rates are slightly lower than those based on cross-validation. This is not surprising since the resubstitution estimates

tend to be optimistically biased (they uses the same data set that was employed to construct the classifiers).

5 Proofs

To prove our results, we first state a number of lemmas.

Lemma 1 *[Pollard (1984).] Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be iid $\mathbb{R}^d$-valued random vectors. Let $\mathcal{F}$ be a class of measurable functions $g : \mathbb{R}^d \to [0, B]$, $B < \infty$. Then, for any $n \geq 1$ and any $\epsilon > 0$,*

$$P\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{Z}_j) - E[f(\mathbf{Z})]\right| > \epsilon\right\} \leq 8E\left[\mathcal{N}_1\big(\epsilon/8, \mathcal{F}, (\mathbf{Z}_j)_{j=1}^n\big)\right] e^{-n\epsilon^2/(128B^2)}.$$

For more on Lemma 1 and its proof one may refer, for example, to Pollard (1984, p. 25) or Györfi et al. (2002, p. 136).

Lemma 2 *Let $\mathcal{P}_n = \{A_{n1}, A_{n2}, \ldots\}$ be a cubic partition of $\mathbb{R}^d$. Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be n iid $\mathbb{R}^d \times [-L, L]$-valued random vectors where $0 < L < \infty$. Let $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ be the regression function and put*

$$m_n^*(\mathbf{x}) = \frac{\sum_{j=1}^n Y_j I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))}, \tag{15}$$

where $A_n(\mathbf{x})$ denotes the unique cell of the partition that contains the point $\mathbf{x}$, and where $\mu(A_n(\mathbf{x})) = P\left\{\mathbf{X} \in A_n(\mathbf{x})\right\}$. Then, under conditions **(A1)** *and* **(A2)**, *for every $\epsilon > 0$, there is a $n_0 > 0$ such that for all $n > n_0$*

$$P\left\{\int \left|m(\mathbf{x}) - m_n^*(\mathbf{x})\right|\mu(d\mathbf{x}) > \epsilon\right\} \leq \exp(-n\epsilon^2/(32L^2)).$$

The proof of this lemma can be found, for example, in Györfi et al. (2002, p. 463).

Lemma 3 *Let the iid random pairs $(\mathbf{U}_j, Y_j)$, $j = 1, \ldots, n$ and the kernel $\mathcal{K}$ (with the smoothing parameter h_n) be as in (8). Define $S(\mathbf{U}_j, Y_j) := f(\mathbf{U}_j)P\{Y = Y_j|Y_j\}\pi(\mathbf{U}_j, Y_j)$ and $\widehat{S}(\mathbf{U}_j, Y_j) := (n-1)^{-1}h_n^{-d}\sum_{k=1,\neq j}^n \delta_k I\{Y_k = Y_j\}\mathcal{K}\left((\mathbf{U}_k - \mathbf{U}_j)/h_n\right)$. Then, under the conditions of Theorem 1, $|S(\mathbf{U}_j, Y_j) - E[\widehat{S}(\mathbf{U}_j, Y_j)|\mathbf{U}_j, Y_j]| \overset{a.s.}{\leq} Ch_n$, where the constant $C > 0$ does not depend on n.*

The proof of Lemma 3 is similar to that of Mojirsheibani (2012, Lemma 3) and will not be given here.

Lemma 4 *Let $\overline{m}_n$ be as in (5) and suppose that conditions (A1) and (A2) hold. If Y is bounded then, for every $\epsilon > 0$, there is an $n_0 > 0$ such that for all $n > n_0$,*

$$\begin{aligned} P\left\{\int \left|\overline{m}_n(\mathbf{x}) - m(\mathbf{x})\right|\mu(d\mathbf{x}) > \epsilon\right\} &\leq P\left\{L\left|\overline{m}_{1,n}(\mathbf{X}) - 1\right| > \frac{\epsilon}{2}\right\} + P\left\{\left|\overline{m}_{2,n}(\mathbf{X}) - E(Y|\mathbf{X})\right| > \frac{\epsilon}{2}\right\} \\ &\leq 4\exp(-n\epsilon^2\pi_{\min}^2/512L^2)\,, \end{aligned}$$

where $\overline{m}_{1,n}$ and $\overline{m}_{2,n}$ are as in (6) and μ is the probability measure of $\mathbf{X}$.

PROOF OF LEMMA 4

Let $\overline{m}_{1,n}$ and $\overline{m}_{2,n}$ be as in (6). Also, let $L < \infty$ be the upper bound on $|Y|$ and observe that

$$\begin{aligned} |\overline{m}_n(\mathbf{X}) - m(\mathbf{X})| &= \left|\frac{\overline{m}_{2,n}(\mathbf{X})}{\overline{m}_{1,n}(\mathbf{X})} - \frac{m(\mathbf{X})}{1}\right| = \left|(\overline{m}_{2,n}(\mathbf{X}) - m(\mathbf{X})) - \frac{\overline{m}_{2,n}(\mathbf{X})}{\overline{m}_{1,n}(\mathbf{X})}(\overline{m}_{1,n}(\mathbf{X}) - 1)\right| \\ &\overset{\text{a.s.}}{\leq} L\left|\overline{m}_{1,n}(\mathbf{X}) - 1\right| + \left|\overline{m}_{2,n}(\mathbf{X}) - E(Y|\mathbf{X})\right| \\ &\overset{\text{a.s.}}{=} L\left|\overline{m}_{1,n}(\mathbf{X}) - E\left(\frac{\delta}{\pi(\mathbf{U},Y)}\middle|\mathbf{X}\right)\right| + \left|\overline{m}_{2,n}(\mathbf{X}) - E\left(\frac{\delta Y}{\pi(\mathbf{U},Y)}\middle|\mathbf{X}\right)\right|, \end{aligned}$$

where we have used the fact that $\left|\frac{\overline{m}_{2,n}(\mathbf{X})}{\overline{m}_{1,n}(\mathbf{X})}\right| \leq^{\text{a.s.}} L$. Therefore,

$$\begin{aligned} \int \left|\overline{m}_n(\mathbf{x}) - m(\mathbf{x})\right|\mu(d\mathbf{x}) &\leq \int \left|\overline{m}_{2,n}(\mathbf{x}) - E\left(\frac{\delta Y}{\pi(\mathbf{U},Y)}\middle|\mathbf{X} = \mathbf{x}\right)\right| \mu(d\mathbf{x}) \\ &\quad + L\int \left|\overline{m}_{1,n}(\mathbf{x}) - E\left(\frac{\delta}{\pi(\mathbf{U},Y)}\middle|\mathbf{X} = \mathbf{x}\right)\right| \mu(d\mathbf{x}) \\ &:= \mathbf{I}_{1,n} + \mathbf{II}_{1,n} \end{aligned} \tag{16}$$

To deal with the term $\mathbf{I}_{1,n}$, first define

$$m_{2,n}^*(\mathbf{x}) = \sum_{j=1}^{n} \frac{\delta_j Y_j}{\pi(\mathbf{U}_j, Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\} / n\mu(A_n(\mathbf{x}))$$

and note that

$$\mathbf{I}_{1,n} \leq \int \left|\overline{m}_{2,n}(\mathbf{x}) - m_{2,n}^*(\mathbf{x})\right| \mu(d\mathbf{x}) + \int \left|m_{2,n}^*(\mathbf{x}) - E\left(\frac{\delta Y}{\pi(\mathbf{U},Y)}|\mathbf{X} = \mathbf{x}\right)\right| \mu(d\mathbf{x}). \tag{17}$$

But by Lemma 2, for every $\epsilon > 0$, and n large enough,

$$P\left\{\int \left|m_{2,n}^*(\mathbf{x}) - E\left(\frac{\delta Y}{\pi(\mathbf{U},Y)}|\mathbf{X} = \mathbf{x}\right)\right| \mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} \leq \exp(-n(\epsilon\pi_{\min})^2/512L^2)\,. \tag{18}$$

As for the term $\int \left|\overline{m}_{2,n}(\mathbf{x}) - m_{2,n}^*(\mathbf{x})\right| \mu(d\mathbf{x})$ in (17), first note that

$$\left|\overline{m}_{2,n}(\mathbf{x}) - m_{2,n}^*(\mathbf{x})\right|$$

$$= \left| \frac{\sum_{j=1}^n \frac{\delta_j Y_j}{\pi(\mathbf{U}_j,Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} - \frac{\sum_{j=1}^n \frac{\delta_j Y_j}{\pi(\mathbf{U}_j,Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))} \right|$$

$$\leq L \sum_{j=1}^n \frac{\delta_j}{\pi(\mathbf{U}_j,Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\} \left| \frac{1}{\sum_{j=1}^n I\{\mathbf{X}_j \in A_n(\mathbf{x})\}} - \frac{1}{n\mu(A_n(\mathbf{x}))} \right|$$

$$\leq \frac{L}{\pi_{\min}} \left| 1 - \frac{\sum_{j=1}^n I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}{n\mu(A_n(\mathbf{x}))} \right| . \tag{19}$$

But the term $\sum_{j=1}^n I\{\mathbf{X}_j \in A_n(\mathbf{x})\}/n\mu(A_n(\mathbf{x}))$ that appears in (19) is just a special case of $m_n^*(\mathbf{x})$ given by (15) corresponding to the situation where Y_j's and Y are all equal to 1 (with probability one). Therefore, once again by Lemma 2, for every $\epsilon > 0$ (and n large enough),

$$\begin{aligned} P\left\{ \int \left|\overline{m}_{2,n}(\mathbf{x}) - m_{2,n}^*(\mathbf{x})\right| \mu(d\mathbf{x}) > \frac{\epsilon}{4} \right\} &\leq P\left\{ \frac{L}{\pi_{\min}} \int \left| 1 - \frac{\sum_{j=1}^n I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}{n\mu(A_n(\mathbf{x}))} \right| \mu(d\mathbf{x}) > \frac{\epsilon}{4} \right\} \\ &\leq \exp\left(-n(\epsilon\pi_{min})^2/521L^2\right) . \end{aligned}$$

Now this last bound together with (18) imply that

$$P\left\{\mathbf{I}_{1,n} > \epsilon/2\right\} \leq 2\exp\left(-n(\epsilon\pi_{min})^2/512L^2\right) . \tag{20}$$

Next, to deal with the term $\mathbf{II}_{1,n}$ in (16), define

$$m_{1,n}^*(\mathbf{x}) = \sum_{j=1}^n \frac{\delta_j}{\pi(\mathbf{U}_j,Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\} / n\mu(A_n(\mathbf{x}))$$

and observe that

$$\mathbf{II}_{1,n} \leq L \int \left|\overline{m}_{1,n}(\mathbf{x}) - m_{1,n}^*(\mathbf{x})\right| \mu(d\mathbf{x}) + L \int \left| m_{1,n}^*(\mathbf{x}) - E\left(\frac{\delta}{\pi(\mathbf{U},Y)} | \mathbf{X} = \mathbf{x} \right) \right| \mu(d\mathbf{x}). \tag{21}$$

Now, by Lemma 2, for every $\epsilon > 0$, and n large enough,

$$P\left\{ L \int \left| m_{1,n}^*(\mathbf{x}) - E\left(\frac{\delta}{\pi(\mathbf{U},Y)} | \mathbf{X} = \mathbf{x} \right) \right| \mu(d\mathbf{x}) > \frac{\epsilon}{4} \right\} \leq \exp(-n(\epsilon\pi_{\min})^2/512L^2) . \tag{22}$$

To deal with the term $L \int \left|\overline{m}_{1,n}(\mathbf{x}) - m_{1,n}^*(\mathbf{x})\right| \mu(d\mathbf{x})$ in (21), first note that

$$L\left|\overline{m}_{1,n}(\mathbf{x}) - m_{1,n}^*(\mathbf{x})\right|$$

$$= L \left| \frac{\sum_{j=1}^n \frac{\delta_j}{\pi(\mathbf{U}_j,Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} - \frac{\sum_{j=1}^n \frac{\delta_j}{\pi(\mathbf{U}_j,Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))} \right|$$

$$\leq L \sum_{j=1}^n \frac{\delta_j}{\pi(\mathbf{U}_j,Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\} \left| \frac{1}{\sum_{j=1}^n I\{\mathbf{X}_j \in A_n(\mathbf{x})\}} - \frac{1}{n\mu(A_n(\mathbf{x}))} \right|$$

$$\leq \frac{L}{\pi_{\min}} \left| 1 - \frac{\sum_{j=1}^n I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}{n\mu(A_n(\mathbf{x}))} \right| .$$

Thus, by Lemma 2, one has (for n large enough)

$$\begin{aligned} & P\left\{L\int\left|\overline{m}_{1,n}(\mathbf{x})-m^*_{1,n}(\mathbf{x})\right|\mu(d\mathbf{x})>\frac{\epsilon}{4}\right\} \\ & \leq P\left\{\frac{L}{\pi_{\min}}\int\left|1-\frac{\sum_{j=1}^n I\{\mathbf{X}_j\in A_n(\mathbf{x})\}}{n\mu(A_n(\mathbf{x}))}\right|\mu(d\mathbf{x})>\frac{\epsilon}{4}\right\} \\ & \leq \exp\left(-n(\epsilon\pi_{min})^2/512L^2\right)\ . \end{aligned} \tag{23}$$

Therefore, in view of (22) and (23), one finds $P\left\{\mathbf{II}_{1,n}>\epsilon/2\right\}\leq 2\exp\left(-n(\epsilon\pi_{min})^2/(512L^2)\right)$. This fact together with (20) yields

$$\begin{aligned} P\left\{\int\left|\overline{m}_n(\mathbf{x})-m(\mathbf{x})\right|\mu(d\mathbf{x})>\epsilon\right\} &\leq P\left\{\mathbf{I}_{1,n}>\frac{\epsilon}{2}\right\}+P\left\{\mathbf{II}_{1,n}>\frac{\epsilon}{2}\right\} \\ &\leq 4\exp\left(-n(\epsilon\pi_{min})^2/512L^2\right), \end{aligned}$$

which completes the proof of Lemma 4.

□

PROOF OF THEOREM 1

Consider the case where $\widehat{\pi}(\mathbf{U}_j,Y_j)$ is taken to be the estimator in (8). Let

$$\widehat{m}_{1,n}(\mathbf{x})=\frac{\sum_{j=1}^n\frac{\delta_j}{\widehat{\pi}(\mathbf{U}_j,Y_j)}I\left\{\mathbf{X}_i\in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j\in A_n(\mathbf{x})\right\}} \quad \text{and} \quad \widehat{m}_{2,n}(\mathbf{x})=\frac{\sum_{j=1}^n\frac{\delta_j Y_j}{\widehat{\pi}(\mathbf{U}_j,Y_j)}I\left\{\mathbf{X}_i\in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j\in A_n(\mathbf{x})\right\}}.$$

Then $\widehat{m}_n(\mathbf{x})$ in (9) can be written as $\widehat{m}_n(\mathbf{x})=\frac{\widehat{m}_{2,n}(\mathbf{x})}{\widehat{m}_{1,n}(\mathbf{x})}$. Furthermoe, it is straightforward to see that

$$\begin{aligned} |\widehat{m}_n(\mathbf{x})-m(\mathbf{x})| &= \left|\frac{\widehat{m}_{2,n}(\mathbf{x})}{\widehat{m}_{1,n}(\mathbf{x})}-\frac{m(\mathbf{x})}{1}\right| = \left|-\frac{\widehat{m}_{2,n}(\mathbf{x})}{\widehat{m}_{1,n}(\mathbf{x})}(\widehat{m}_{1,n}(\mathbf{x})-1)+(|\widehat{m}_n(\mathbf{x})-m(\mathbf{x}))\right| \\ &\leq L\left|\widehat{m}_{1,n}(\mathbf{x})-1\right|+\left|\widehat{m}_{2,n}(\mathbf{x})-E(Y|\mathbf{X}=\mathbf{x})\right| \\ &\leq L\left|\widehat{m}_{1,n}(\mathbf{x})-\overline{m}_{1,n}(\mathbf{x})\right|+\left|\widehat{m}_{2,n}(\mathbf{x})-\overline{m}_{2,n}(\mathbf{x})\right|+L\left|\overline{m}_{1,n}(\mathbf{x})-1\right| \\ &\quad +\left|\overline{m}_{2,n}(\mathbf{x})-E(Y\left|\mathbf{X}=\mathbf{x}\right)\right| \\ &= \Delta_1(\mathbf{x})+\Delta_2(\mathbf{x})+\Delta_3(\mathbf{x})+\Delta_4(\mathbf{x}) \end{aligned} \tag{24}$$

where $\overline{m}_{1,n}(\mathbf{x})$ and $\overline{m}_{2,n}(\mathbf{x})$ are as in (6). But by Lemma 4, one finds

$$P\left\{\int\Delta_3(\mathbf{x})\mu(d\mathbf{x})>\frac{\epsilon}{4}\right\}+P\left\{\int\Delta_4(\mathbf{x})\mu(d\mathbf{x})>\frac{\epsilon}{4}\right\}\leq 4\exp\left(-n(\epsilon\pi_{\min})^2/2048L^2\right)\ .$$

To deal with the term $\triangle_2(\mathbf{x})$ in (24), first observe that

$$\begin{aligned}
\Delta_2(\mathbf{x}) &= \left| \frac{\sum_{j=1}^n \frac{\delta_j}{\widehat{\pi}(\mathbf{U}_j,Y_j)} Y_j I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} - \frac{\sum_{j=1}^n \frac{\delta_j}{\pi(\mathbf{U}_j,Y_j)} Y_j I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} \right| \\
&\leq \left| \frac{\sum_{j=1}^n \left(\frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)} - \frac{1}{\pi(\mathbf{U}_j,Y_j)} \right) \delta_j Y_j I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} \right| \\
&\leq \left| \frac{\sum_{j=1}^n \left(\frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)} - \frac{1}{\pi(\mathbf{U}_j,Y_j)} \right) \delta_j Y_j I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} \right. \\
&\qquad \left. - \frac{\sum_{j=1}^n \left(\frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)} - \frac{1}{\pi(\mathbf{U}_j,Y_j)} \right) \delta_j Y_j I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))} \right| \\
&\qquad + \left| \frac{\sum_{j=1}^n \left(\frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)} - \frac{1}{\pi(\mathbf{U}_j,Y_j)} \right) \delta_j Y_j I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))} \right| \\
&:= \mathbf{I}_{2,n}(\mathbf{x}) + \mathbf{II}_{2,n}(\mathbf{x}). \qquad (25)
\end{aligned}$$

However, we can write

$$\begin{aligned}
\mathbf{I}_{2,n}(\mathbf{x}) &= \left| \sum_{j=1}^n \left(\frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)} - \frac{1}{\pi(\mathbf{U}_j,Y_j)} \right) \delta_j Y_j I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\} \right. \\
&\qquad \left. \times \left(\frac{1}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} - \frac{1}{n\mu(A_n(\mathbf{x}))} \right) \right| \\
&\leq L \max_{1\leq j\leq n} \left| \frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)} - \frac{1}{\pi(\mathbf{U}_j,Y_j)} \right| \times \frac{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} \left| 1 - \frac{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))} \right| \\
&:= \mathbf{I}_{3,n}(\mathbf{x})\,.
\end{aligned}$$

Furthermore, the term $\mathbf{II}_{2,n}(\mathbf{x})$ in (25) satisfies

$$\mathbf{II}_{2,n}(\mathbf{x}) \leq \frac{L}{n} \sum_{j=1}^n \left| \frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)} - \frac{1}{\pi(\mathbf{U}_j,Y_j)} \right| \times \frac{I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\mu(A_n(\mathbf{x}))} := \mathbf{II}_{3,n}(\mathbf{x})\,.$$

Hence, for every $\epsilon > 0$ we have:

$$\begin{aligned}
P\left\{ \int \Delta_2(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4} \right\} &\leq P\left\{ \int \mathbf{I}_{3,n}(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{8} \right\} + P\left\{ \int \mathbf{II}_{3,n}(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{8} \right\} \\
&:= \mathbf{I}_{4,n} + \mathbf{II}_{4,n}\,. \qquad (26)
\end{aligned}$$

To deal with the term $\mathbf{I}_{4,n}$ in (26), we start by writing

$$\begin{aligned}
\mathbf{I}_{4n} &\leq P\Bigg\{ \left[\max_{1\leq j\leq n} \left| \frac{1}{\widehat{\pi}(\mathbf{U}_j, Y_j)} - \frac{1}{\pi(\mathbf{U}_j, Y_j)} \right| \int \left| 1 - \frac{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))} \right| \mu(d\mathbf{x}) > \frac{\epsilon}{8L} \right] \\
&\qquad \bigcap_{j=1}^n \left[\widehat{\pi}(\mathbf{U}_j, Y_j) \geq \frac{\pi_{\min}}{2} \right] \Bigg\} + P\left\{ \bigcup_{j=1}^n \left[\widehat{\pi}(\mathbf{U}_j, Y_j) < \frac{\pi_{\min}}{2} \right] \right\} \\
&\leq P\left\{ \frac{2}{\pi_{\min}} \int \left| 1 - \frac{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))} \right| \mu(d\mathbf{x}) > \frac{\epsilon}{8L} \right\} + \sum_{j=1}^n P\left\{ \widehat{\pi}(\mathbf{U}_j, Y_j) < \frac{\pi_{\min}}{2} \right\} \\
&:= \mathbf{I}_{5,n} + \mathbf{II}_{5,n} \,. \qquad (27)
\end{aligned}$$

But by Lemma 2, and for n large enough,

$$\mathbf{I}_{5,n} = P\left\{ \int \left| 1 - \frac{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))} \right| \mu(d\mathbf{x}) > \frac{\epsilon\pi_{\min}}{16L} \right\} \leq \exp(-n(\epsilon\pi_{\min})^2/8192L^2)$$

As for the term $\mathbf{II}_{5,n}$ in (27), let f be the density of $\mathbf{U}$ and put $\mathcal{R}(\mathbf{U}_j, Y_j) = f(\mathbf{U}_j)P\{Y = Y_j|Y_j\}$. Also take $\widehat{\mathcal{R}}(\mathbf{U}_j, Y_j) = \frac{1}{n-1}\frac{1}{h_n^d}\sum_{k=1\neq j}^n I\{Y_k = Y_j\}\mathcal{K}(\frac{\mathbf{U}_k - \mathbf{U}_j}{h_n})$. Furthermore, let $\widehat{S}(\mathbf{U}_j, Y_j)$ and $S(\mathbf{U}_j, Y_j)$ be as in Lemma 3, and observe that,

$$\begin{aligned}
|\widehat{\pi}(\mathbf{U}_j, Y_j) - \pi(\mathbf{U}_j, Y_j)| &= \left| \frac{\widehat{S}(\mathbf{U}_j, Y_j)}{\widehat{\mathcal{R}}(\mathbf{U}_j, Y_j)} - \frac{S(\mathbf{U}_j, Y_j)}{\mathcal{R}(\mathbf{U}_j, Y_j)} \right| \\
&= \left| \frac{-\widehat{S}(\mathbf{U}_j, Y_j)/\widehat{\mathcal{R}}(\mathbf{U}_j, Y_j)}{\mathcal{R}(\mathbf{U}_j, Y_j)} \left(\widehat{\mathcal{R}}(\mathbf{U}_j, Y_j) - \mathcal{R}(\mathbf{U}_j, Y_j) \right) + \frac{\widehat{S}(\mathbf{U}_j, Y_j) - S(\mathbf{U}_j, Y_j)}{\mathcal{R}(\mathbf{U}_j, Y_j)} \right| \\
&\leq \left| \frac{\widehat{\mathcal{R}}(\mathbf{U}_j, Y_j) - \mathcal{R}(\mathbf{U}_j, Y_j)}{\mathcal{R}(\mathbf{U}_j, Y_j)} \right| + \left| \frac{\widehat{S}(\mathbf{U}_j, Y_j) - S(\mathbf{U}_j, Y_j)}{\mathcal{R}(\mathbf{U}_j, Y_j)} \right| \,. \qquad (28)
\end{aligned}$$

We also note that $\mathcal{R}(\mathbf{U}_j, Y_j) \geq \min_{y\in Y} P\{Y = y\} f_{\min}$, where $f_{\min} := \inf_{\mathbf{u}} f(\mathbf{u}) > 0$ by condition **(A6)**. Now, put $p := \min_{y\in Y} P\{Y = y\}$ and observe that

$$\begin{aligned}
P\left\{ \widehat{\pi}(\mathbf{U}_j, Y_j) < \frac{\pi_{\min}}{2} \right\} &\leq P\left\{ |\widehat{\pi}(\mathbf{U}_j, Y_j) - \pi(\mathbf{U}_j, Y_j)| > \frac{\pi_{\min}}{2} \right\} \\
&\leq P\left\{ \left| \widehat{S}(\mathbf{U}_j, Y_j) - S(\mathbf{U}_j, Y_j) \right| > \frac{p\pi_{\min}}{4} f_{\min} \right\} \\
&\qquad + P\left\{ \left| \widehat{\mathcal{R}}(\mathbf{U}_j, Y_j) - \mathcal{R}(\mathbf{U}_j, Y_j) \right| > \frac{p\pi_{\min}}{4} f_{\min} \right\} \\
&:= \mathbf{I}_{6,j,n} + \mathbf{II}_{6,j,n} \,. \qquad (29)
\end{aligned}$$

But for n large enough, Lemma 3 implies that

$$\begin{aligned}
\mathbf{I}_{6,j,n} &\leq P\left\{\left|\widehat{S}(\mathbf{U}_j,Y_j)-E\left[\widehat{S}(\mathbf{U}_j,Y_j)\middle|\mathbf{U}_j,Y_j\right]\right|+\left|E\left[\widehat{S}(\mathbf{U}_j,Y_j)\middle|\mathbf{U}_j,Y_j\right]-S(\mathbf{U}_j,Y_j)\right|>\frac{p\pi_{\min}}{4}f_{\min}\right\}\\
&\leq P\left\{\left|\widehat{S}(\mathbf{U}_j,Y_j)-E\left[\widehat{S}(\mathbf{U}_j,Y_j)\middle|\mathbf{U}_j,Y_j\right]\right|>\frac{p\pi_{\min}}{8}f_{\min}\right\}\\
&= E\left[P\left\{\left|\widehat{S}(\mathbf{U}_j,Y_j)-E\left[\widehat{S}(\mathbf{U}_j,Y_j)\middle|\mathbf{U}_j,Y_j\right]\right|>\frac{p\pi_{\min}}{8}f_{\min}\middle|\mathbf{U}_j,Y_j\right\}\right]\\
&= E\left[P\left\{\frac{1}{n-1}\left|\sum_{k=1,\neq j}^{n}\mathcal{T}_k(\mathbf{U}_j,Y_j)\right|>\frac{p\pi_{\min}}{8}f_{\min}\middle|\mathbf{U}_j,Y_j\right\}\right],
\end{aligned}$$

where

$$\mathcal{T}_k(\mathbf{U}_j,Y_j)=\frac{1}{h_n^d}\left[\delta_k I\{Y_k=Y_j\}\mathcal{K}\left(\frac{\mathbf{U}_k-\mathbf{U}_j}{h_n}\right)-E\left(\delta_k I\{Y_k=Y_j\}\mathcal{K}\left(\frac{\mathbf{U}_k-\mathbf{U}_j}{h_n}\right)\middle|\mathbf{U}_j,Y_j\right)\right].$$

However, conditional on $(\mathbf{U}_j,Y_j)$, the terms $\mathcal{T}_k(\mathbf{U}_j,Y_j)$, $k=1,\cdots,n,\ k\neq j$, are independent, zero mean random variables, bounded by $-h_n^{-d}\,||\mathcal{K}||_\infty$ and $h_n^{-d}\,||\mathcal{K}||_\infty$. Also, we note that $\mathrm{Var}(\mathcal{T}_k(\mathbf{U}_j,Y_j)|\mathbf{U}_j,Y_j)=E(\mathcal{T}_k^2(\mathbf{U}_j,Y_j)\big|\mathbf{U}_j,Y_j)\leq||\mathcal{K}||_\infty\,||f||_\infty h_n^{-d}$. Therefore if we let $D_1=p\pi_{\min}f_{\min}/8$ then by Bernestein's inequality,

$$P\left\{\frac{1}{n-1}\left|\sum_{k=1,\neq j}^{n}\mathcal{T}_k(\mathbf{U}_j,Y_j)\right|>D_1\middle|\mathbf{U}_j,Y_j\right\}\leq 2\exp\left\{\frac{-(n-1)h_n^dD_1^2}{2\,||\mathcal{K}||_\infty\,||f||_\infty+D_1}\right\},\tag{30}$$

which implies that

$$\mathbf{I}_{6,j,n}\leq 2\exp\left\{\frac{-(n-1)h_n^dD_1^2}{2\,||\mathcal{K}||_\infty\,||f||_\infty+D_1}\right\}.\tag{31}$$

Since $\widehat{\mathcal{R}}(\mathbf{U}_j,Y_j)$ and $\mathcal{R}(\mathbf{U}_j,Y_j)$ are just the special cases of $\widehat{S}(\mathbf{U}_j,Y_j)$ and $S(\mathbf{U}_j,Y_j)$ in Lemma 3 with $\delta=1$, one finds that for n large enough,

$$\mathbf{II}_{6,j,n}\leq 2\exp\left\{\frac{-(n-1)h_n^dD_1^2}{2\,||\mathcal{K}||_\infty\,||f||_\infty+D_1}\right\}.\tag{32}$$

Now, let $D_2=D_1^2/[2\,||\mathcal{K}||_\infty\,||f||_\infty+D_1]$ and observe that in view of (19) and (21) $\mathbf{II}_{5,n}\leq\sum_{j=1}^n(\mathbf{I}_{6,j,n}+\mathbf{II}_{6,j,n})\leq 4n\exp\{-D_2nh_n^d\}$. Therefore, the term $\mathbf{I}_{4,n}$ in (26) can be bounded according to

$$\mathbf{I}_{4,n}\leq\mathbf{I}_{5,n}+\mathbf{II}_{5,n}\leq\exp\left\{-n(\epsilon\pi_{\min})^2/8192L^2\right\}+4n\exp\left\{-D_2nh_n^d\right\}.\tag{33}$$

As for the term $\mathbf{II}_{4,n}$ in (26), it can be written as

$$\begin{aligned}
\mathbf{II}_{4,n} &= P\left\{\frac{L}{n}\sum_{j=1}^{n}\left|\frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)}-\frac{1}{\pi(\mathbf{U}_j,Y_j)}\right|\int\frac{I\{\mathbf{X}_j\in A_n(\mathbf{x})\}}{\mu(A_n(\mathbf{x}))}\mu(d\mathbf{x})>\frac{\epsilon}{8}\right\}\\
&= P\left\{\frac{L}{n}\sum_{j=1}^{n}\left|\frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)}-\frac{1}{\pi(\mathbf{U}_j,Y_j)}\right|>\frac{\epsilon}{8}\right\},
\end{aligned}$$

where we have used the fact that $\int \frac{I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}{\mu(A_n(\mathbf{x}))}\mu(d\mathbf{x}) \stackrel{\text{a.s.}}{=} 1$ (see for example, Györfi et al. (2002, p. 462)). Consequently,

$$\begin{aligned}
\mathbf{II}_{4,n} &= P\left\{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)}\left|1-\frac{\widehat{\pi}(\mathbf{U}_j,Y_j)}{\pi(\mathbf{U}_j,Y_j)}\right| > \frac{\epsilon}{8L}\right\} \\
&\leq P\left\{\left[\frac{1}{n}\sum_{j=1}^{n}\frac{1}{\widehat{\pi}(\mathbf{U}_j,Y_j)}\left|\frac{\widehat{\pi}(\mathbf{U}_j,Y_j)}{\pi(\mathbf{U}_j,Y_j)}-1\right| > \frac{\epsilon}{8L}\right]\bigcap\bigcap_{j=1}^{n}\left\{\widehat{\pi}(\mathbf{U}_j,Y_j)\geq\frac{\pi_{\min}}{2}\right\}\right\} \\
&\quad + P\left\{\bigcup_{j=1}^{n}\left\{\widehat{\pi}(\mathbf{U}_j,Y_j)<\frac{\pi_{\min}}{2}\right\}\right\} \\
&\leq \sum_{j=1}^{n}P\left\{\frac{2}{\pi_{\min}^2}\left|\widehat{\pi}(\mathbf{U},Y_j)-\pi(\mathbf{U}_j,Y_j)\right|\geq\frac{\epsilon}{8L}\right\}+\sum_{j=1}^{n}P\left\{\widehat{\pi}(\mathbf{U}_j,Y_j)<\frac{\pi_{\min}}{2}\right\} \\
&:= \mathbf{I}_{7,n}+\mathbf{II}_{7,n}\,. \qquad (34)
\end{aligned}$$

Since $\mathbf{II}_{7,n} = \mathbf{II}_{5,n}$, one has $\mathbf{II}_{7,n} \leq 4n\exp\{-D_2 n h_n^d\}$. To handle the term $\mathbf{I}_{7,n}$, we note that since

$$|\widehat{\pi}(\mathbf{U},Y_j)-\pi(\mathbf{U}_j,Y_j)| \leq \left|\frac{\widehat{\mathcal{R}}(\mathbf{U}_j,Y_j)-\mathcal{R}(\mathbf{U}_j,Y_j)}{\mathcal{R}(\mathbf{U}_j,Y_j)}\right| + \left|\frac{\widehat{S}(\mathbf{U}_j,Y_j)-S(\mathbf{U}_j,Y_j)}{\mathcal{R}(\mathbf{U}_j,Y_j)}\right|,$$

one finds

$$\begin{aligned}
P\left\{|\widehat{\pi}(\mathbf{U}_j,Y_j)-\pi(\mathbf{U}_j,Y_j)|>\frac{\pi_{\min}^2\epsilon}{16L}\right\} &\leq P\left\{\left|\widehat{S}(\mathbf{U}_j,Y_j)-S(\mathbf{U}_j,Y_j)\right|>\frac{pf_{\min}\pi_{\min}^2\epsilon}{32L}\right\} \\
&\quad + P\left\{\left|\widehat{\mathcal{R}}(\mathbf{U}_j,Y_j)-\mathcal{R}(\mathbf{U}_j,Y_j)\right|>\frac{pf_{\min}\pi_{\min}^2\epsilon}{32L}\right\} \\
&:= \mathbf{I}_{8,j,n}+\mathbf{II}_{8,j,n} \qquad (35)
\end{aligned}$$

Now Lemma 3 and the arguments that lead to (30) yield

$$\mathbf{I}_{8,j,n} \leq 2\exp\left\{\frac{-(n-1)h_n^d D_3^2\epsilon^2}{2\,||\mathcal{K}||_\infty\,||f||_\infty + D_3\epsilon}\right\},$$

where $D_3 = pf_{\min}\pi_{\min}^2/(32L)$. Since $|\widehat{\pi}(\mathbf{U}_j,Y_j)-\pi(\mathbf{U}_j,Y_j)| \leq 2$, one only needs to consider the case where $\epsilon < 32L/\pi_{\min}^2$ in the probability statement on the far left side of (35). Therefore, for n large enough, one has

$$\mathbf{I}_{8,j,n} \leq 2\exp\left\{\frac{-(n-1)h_n^d D_3^2\epsilon^2}{2\,||\mathcal{K}||_\infty\,||f||_\infty + (D_3 32L/\pi_{\min}^2)}\right\}.$$

One can also show that for n larger enough,

$$\mathbf{II}_{8,j,n} \leq 2\exp\left\{\frac{-(n-1)h_n^d D_3^2\epsilon^2}{2\,||\mathcal{K}||_\infty\,||f||_\infty + (D_3 32L/\pi_{\min}^2)}\right\}.$$

Therefore, $\mathbf{I}_{7,n} \leq \sum_{j=1}^{n}(\mathbf{I}_{8,j,n} + \mathbf{II}_{8,j,n}) \leq 4n\exp\{-nD_4 4_n^d \epsilon^2\}$, where we can take $D_4 = D_3^2/[2\,||\mathcal{K}||_\infty\,||f||_\infty + (D_3 32L/\pi_{\min}^2)]$. Putting all of the above together, one has

$$P\left\{\int \Delta_2(\mathbf{x})\mu(d\mathbf{x})) > \frac{\epsilon}{4}\right\} \leq \mathbf{I}_{4,n} + \mathbf{II}_{4,n} \leq (\mathbf{I}_{5,n} + \mathbf{II}_{5,n}) + (\mathbf{I}_{7,n} + \mathbf{II}_{7,n})$$
$$\leq \exp\{-nD_0\epsilon^2\} + 8n\exp\{-nD_2 h_n^d\} + 4n\exp\{-nD_4 h_n^d \epsilon^2\}, \qquad (36)$$

where $D_0 = \pi_{\min}^2/(8192L^2)$, and D_2 and D_4 are as stated before. Similarly, for n large enough, we find

$$P\left\{\int \Delta_1(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} \leq \exp\{-nD_0\epsilon^2\} + 8n\exp\{-nD_2 h_n^d\} + 4n\exp\{-nD_4 h_n^d \epsilon^2\}\,.$$

Therefore,

$$\begin{aligned}
&P\left\{\int |\widehat{m}_n(\mathbf{x}) - m(\mathbf{x})|\,\mu(d\mathbf{x}) > \epsilon\right\} \\
&\leq P\left\{\int \Delta_1(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} + P\left\{\int \Delta_2(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} \\
&\quad + P\left\{\int \Delta_3(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} + P\left\{\int \Delta_4(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} \\
&\leq 4\exp\{-C_1 n\epsilon^2\} + 2\exp\{-D_0 n\epsilon^2\} + 16n\exp\{-D_2 n h_n^d\} + 8n\exp\{-D_4 n h_n^d \epsilon^2\} \\
&\leq 6\exp\{-\min(C_1, D_0)n\epsilon^2\} + 16n\exp\{-D_2 n h_n^d\} + 8n\exp\{-D_4 n h_n^d \epsilon^2\},
\end{aligned}$$

where $C_1 = \pi_{\min}^2/(2048L^2)$. This completes the proof of Theorem 1.

□

PROOF OF THEOREM 2

Define

$$\widehat{m}_{\text{L1}}(\mathbf{x}) = \frac{\sum_{j=1}^{n} \frac{\delta_j}{\widehat{\pi}_{\text{LS}}(\mathbf{U}_j, Y_j)} I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}{\sum_{j=1}^{n} I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}\,.$$

and

$$\widehat{m}_{\text{L2}}(\mathbf{x}) = \frac{\sum_{j=1}^{n} \frac{\delta_j Y_j}{\widehat{\pi}_{\text{LS}}(\mathbf{U}_j, Y_j)} I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}{\sum_{j=1}^{n} I\{\mathbf{X}_j \in A_n(\mathbf{x})\}}\,.$$

Then, as in the proof of Theorem 1, we can write

$$\begin{aligned}
|\widehat{m}_{\text{LS}}(\mathbf{x}) - m(\mathbf{x})| &\leq L\,|\widehat{m}_{\text{L1}}(\mathbf{x}) - \overline{m}_{1,n}(\mathbf{x})| + L\,|\overline{m}_{1,n}(\mathbf{x}) - 1| \\
&\quad + |\widehat{m}_{\text{L2}}(\mathbf{x}) - \overline{m}_{2,n}(\mathbf{x})| + |\overline{m}_{2,n}(\mathbf{x}) - E(Y|\mathbf{X} = \mathbf{x})| \\
&:= \Delta_5(\mathbf{x}) + \Delta_6(\mathbf{x}) + \Delta_7(\mathbf{x}) + \Delta_8(\mathbf{x})
\end{aligned}$$

where $\overline{m}_{1,n}(\mathbf{x})$ and $\overline{m}_{2,n}(\mathbf{x})$ are as in (6). Now by Lemma 4, we have

$$P\left\{\int \Delta_6(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} + P\left\{\int \Delta_8(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} \leq 4\exp(-n(\epsilon\pi_{\min})^2/2048L^2)\,.$$

Next, we note that

$$\begin{aligned}
\Delta_7(\mathbf{x}) &= \left|\frac{\sum_{j=1}^n \frac{\delta_j Y_j}{\widehat{\pi}_{\rm LS}(\mathbf{U}_j,Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} - \frac{\sum_{j=1}^n \frac{\delta_j Y_j}{\pi(\mathbf{U}_j,Y_j)} I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}\right| \\
&\leq \left|\frac{\sum_{j=1}^n \frac{\delta_j Y_j}{\widehat{\pi}_{\rm LS}(\mathbf{U}_j,Y_j)} \left[\frac{\widehat{\pi}_{\rm LS}(\mathbf{U}_j,Y_j)}{\pi(\mathbf{U}_j,Y_j)} - 1\right] I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}\right. \\
&\qquad \left. - \frac{\sum_{j=1}^n \frac{\delta_j Y_j}{\widehat{\pi}_{\rm LS}(\mathbf{U}_j,Y_j)} \left[\frac{\widehat{\pi}_{\rm LS}(\mathbf{U}_j,Y_j)}{\pi(\mathbf{U}_j,Y_j)} - 1\right] I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))}\right| \\
&\qquad + \left|\frac{\sum_{j=1}^n \frac{\delta_j Y_j}{\widehat{\pi}_{\rm LS}(\mathbf{U}_j,Y_j)} \left[\frac{\widehat{\pi}_{\rm LS}(\mathbf{U}_j,Y_j)}{\pi(\mathbf{U}_j,Y_j)} - 1\right] I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(A_n(\mathbf{x}))}\right| \\
&:= \mathbf{I}_{9,n}(\mathbf{x}) + \mathbf{II}_{9,n}(\mathbf{x})\,. \qquad (37)
\end{aligned}$$

Since $\left|[\widehat{\pi}_{\rm LS}(\mathbf{U}_j,Y_j)/\pi(\mathbf{U}_j,Y_j)] - 1\right|/\widehat{\pi}_{\rm LS}(\mathbf{U}_j,Y_j) \leq 2/\pi_{\min}$, one finds that

$$\begin{aligned}
\mathbf{I}_{9,n}(\mathbf{x}) &\leq \left|\frac{2L}{\pi_{\min}}\left[\frac{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}} - \frac{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(An(\mathbf{x}))}\right]\right| \\
&= \left|\frac{2L}{\pi_{\min}}\left[1 - \frac{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(An(\mathbf{x}))}\right]\right|\,. \qquad (38)
\end{aligned}$$

Furthermore, with $\mathbf{I}_{9,n}(\mathbf{x})$ and $\mathbf{II}_{9,n}(\mathbf{x})$ as in (37), one has

$$\begin{aligned}
P\left\{\int \Delta_7(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} &\leq P\left\{\int \mathbf{I}_{9,n}(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{8}\right\} + P\left\{\int \mathbf{II}_{9,n}(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{8}\right\} \\
&:= \mathbf{I}_{10,n} + \mathbf{II}_{10,n}\,. \qquad (39)
\end{aligned}$$

Now, by (38) and Lemma 2, for every $\epsilon > 0$ and n larger enough,

$$\mathbf{I}_{10,n} \leq P\left\{\frac{2L}{\pi_{\min}}\int \left|1 - \frac{\sum_{j=1}^n I\left\{\mathbf{X}_j \in A_n(\mathbf{x})\right\}}{n\mu(An(\mathbf{x}))}\right| \mu(d\mathbf{x}) > \frac{\epsilon}{8}\right\} \leq \exp(-n(\epsilon\pi_{\min})^2/(8192L^2))\,.$$

To deal with the term $\mathbf{II}_{10,n}$ in (39), first observe that,

$$\begin{aligned}\int \mathbf{II}_{9,n}(\mathbf{x})\mu(d\mathbf{x}) &= \frac{L}{n}\sum_{j=1}^{n}\frac{1}{\widehat{\pi}_{\text{LS}}(\mathbf{U}_j,Y_j)}\left|\frac{\widehat{\pi}_{\text{LS}}(\mathbf{U}_j,Y_j)}{\pi(\mathbf{U}_j,Y_j)}-1\right|\int\frac{I\{\mathbf{X}_j\in A_n(\mathbf{x})\}}{\mu(A_n(\mathbf{x}))}\mu(d\mathbf{x}) \\ &\overset{\text{a.s.}}{=} \frac{L}{n}\sum_{j=1}^{n}\frac{1}{\widehat{\pi}_{\text{LS}}(\mathbf{U}_j,Y_j)}\left|\frac{\widehat{\pi}_{\text{LS}}(\mathbf{U}_j,Y_j)}{\pi(\mathbf{U}_j,Y_j)}-1\right| \\ &\leq \frac{L}{n\pi_{\min}^2}\sum_{j=1}^{n}\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_j,Y_j)-\pi(\mathbf{U},Y_j)\right|,\end{aligned}$$

where, once again, we have used the fact that $\int\frac{I\{\mathbf{X}_j\in A_n(\mathbf{x})\}}{\mu(A_n(\mathbf{x}))}\mu(d\mathbf{x})\overset{\text{a.s.}}{=}1$. Therefore,

$$\begin{aligned}\mathbf{II}_{10,n} &\leq P\left\{\frac{L}{n\pi_{\min}^2}\sum_{j=1}^{n}\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_j,Y_j)-\pi(\mathbf{U}_j,Y_j)\right|>\frac{\epsilon}{8}\right\} \\ &= P\Bigg\{\frac{1}{n}\sum_{j=1}^{n}\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_j,Y_j)-\pi(\mathbf{U}_j,Y_j)\right|-E\left(\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_1,Y_1)-\pi(\mathbf{U}_1,Y_1)\right|\Big|\mathbb{D}_n\right) \\ &\qquad +E\left(\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_1,Y_1)-\pi(\mathbf{U}_1,Y_1)\right|\Big|\mathbb{D}_n\right)>\frac{\pi_{\min}^2\epsilon}{8L}\Bigg\} \\ &\leq P\left\{\sup_{\pi'\in\mathscr{P}}\left|\frac{1}{n}\sum_{j=1}^{n}\left|\pi'(\mathbf{U}_j,Y_j)-\pi(\mathbf{U}_j,Y_j)\right|-E\left(\left|\pi'(\mathbf{U}_1,Y_1)-\pi(\mathbf{U}_1,Y_1)\right|\right)\right|>\frac{\pi_{\min}^2\epsilon}{16L}\right\} \\ &\qquad +P\left\{E\left(\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_1,Y_1)-\pi(\mathbf{U}_1,Y_1)\right|\Big|\mathbb{D}_n\right)>\frac{\pi_{\min}^2\epsilon}{16L}\right\} \\ &:= \mathbf{I}_{11,n}+\mathbf{II}_{11,n}.\end{aligned}$$

Consequently, by Lemma 1, one finds

$$\mathbf{I}_{11,n}\leq 8E\left[\mathcal{N}_1\left(\frac{\pi_{\min}^2\epsilon}{128L},\mathscr{P},(\mathbf{U}_j,Y_j)_{j=1}^n\right)\right]\exp(-C_6n\epsilon^2),$$

where $C_6=\pi_{\min}^4/(2^{15}L^2)$. To deal with the term $\mathbf{II}_{11,n}$, put $S_n(\pi)=n^{-1}\sum_{j=1}^n(\pi(\mathbf{U}_j,Y_j)-\delta_j)^2$, and note that by the Cauchy-Schwarz inequality

$$\begin{aligned}\mathbf{II}_{11,n} &\leq P\left\{E\left[\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_1,Y_1)-\pi(\mathbf{U}_1,Y_1)\right|^2\Big|\mathbb{D}_n\right]>\frac{\pi_{\min}^4\epsilon^2}{256L^2}\right\} &(40)\\ &= P\left\{E\left[\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_1,Y_1)-\delta_1\right|^2\Big|\mathbb{D}_n\right]-E\left|\pi(\mathbf{U}_1,Y_1)-\delta_1\right|^2>\frac{\pi_{\min}^4\epsilon^2}{256L^2}\right\} \\ &\leq P\left\{2\sup_{\pi'\in\mathscr{P}}\left|S_n(\pi')-E\left|\pi'(\mathbf{U}_1,Y_1)-\delta_1\right|^2\right|>\frac{\pi_{\min}^4\epsilon^2}{256L^2}\right\} &(41)\end{aligned}$$

Here, (41) follows because

$$
\begin{aligned}
&E\left[\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_1, Y_1) - \delta_1\right|^2 \middle| \mathbb{D}_n\right] - E\left|\pi(\mathbf{U}_1, Y_1) - \delta_1\right|^2 \\
&= E\left[\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_1, Y_1) - \delta_1\right|^2 \middle| \mathbb{D}_n\right] - \inf_{\pi' \in \mathscr{P}} E\left|\pi'(\mathbf{U}_1, Y_1) - \delta_1\right|^2 \\
&= \sup_{\pi' \in \mathscr{P}} \left\{ E\left[\left|\widehat{\pi}_{\text{LS}}(\mathbf{U}_1, Y_1) - \delta_1\right|^2 \middle| \mathbb{D}_n\right] \right. \\
&\quad \left. - S_n(\widehat{\pi}_{\text{LS}}) + S_n(\widehat{\pi}_{\text{LS}}) - S_n(\pi') + S_n(\pi') - E\left|\pi'(\mathbf{U}_1, Y_1) - \delta_1\right|^2 \right\} \\
&\leq 2 \sup_{\pi' \in \mathscr{P}} \left| S_n(\pi') - E\left|\pi'(\mathbf{U}_1, Y_1) - \delta_1\right|^2 \right| ,
\end{aligned}
$$

where we have used the fact that $S_n(\widehat{\pi}_{\text{LS}}) - S_n(\pi') \leq 0$ by the definition of $\widehat{\pi}_{\text{LS}}$. Therefore, by Lemma 1,

$$
\mathbf{II}_{11,n} \leq 8E\left[\mathcal{N}_1\left(\frac{\pi_{\min}^4 \epsilon^2}{4096L}, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n\right)\right] \exp(-C_7 n\epsilon^4) , \tag{42}
$$

where $C_7 = \pi_{\min}^8/(2^{25}L^4)$. Hence, one has

$$
\begin{aligned}
P\left\{\int \Delta_7(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} &\leq \mathbf{I}_{10,n} + \mathbf{II}_{10,n} \\
&\leq \exp(-C_5 n\epsilon^2) + \mathbf{I}_{11,n} + \mathbf{II}_{11,n} \\
&\leq \exp(-C_5 n\epsilon^2) + 8E\left[\mathcal{N}_1\left(\frac{\pi_{\min}^2 \epsilon}{128L}, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n\right)\right] \exp(-C_6 n\epsilon^2) \\
&\quad + 8E\left[\mathcal{N}_1\left(\frac{\pi_{\min}^4 \epsilon^2}{4096L}, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n\right)\right] \exp(-C_7 n\epsilon^4)
\end{aligned}
$$

where $C_5 = \pi_{\min}^2/(2^{12}L^2)$, $C_6 = \pi_{\min}^4/(2^{15}L^2)$, and $C_7 = \pi_{\min}^8/(2^{25}L^4)$. Furthermore, using arguments similar to those in the derivation of (37) - (42), one can show that

$$
\begin{aligned}
P\left\{\int \Delta_5(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} &\leq \exp(-C_5 n\epsilon^2) + 8E\left[\mathcal{N}_1\left(\frac{\epsilon\pi_{\min}^2}{128L}, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n\right)\right] \exp(-C_6 n\epsilon^2) \\
&\quad + 8E\left[\mathcal{N}_1\left(\frac{\epsilon^2\pi_{\min}^4}{4096L}, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n\right)\right] \exp(-C_7 n\epsilon^4)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&P\left\{\int |\widehat{m}_{\text{LS}}(\mathbf{x}) - m(\mathbf{x})|\,\mu(d\mathbf{x}) > \epsilon\right\} \\
&\quad \le P\left\{\int \Delta_5(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} + P\left\{\int \Delta_6(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} \\
&\qquad\quad + P\left\{\int \Delta_7(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} + P\left\{\int \Delta_8(\mathbf{x})\mu(d\mathbf{x}) > \frac{\epsilon}{4}\right\} \\
&\quad \le 4\exp(-C_1 n\epsilon^2) + 2\exp(-C_5 n\epsilon^2) + 16E\left[\mathcal{N}_1\left(\frac{\epsilon\pi_{\min}^2}{128L}, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n\right)\right]\exp(-C_6 n\epsilon^2) \\
&\qquad\quad + 16E\left[\mathcal{N}_1\left(\frac{\epsilon^2\pi_{\min}^4}{4096L}, \mathscr{P}, (\mathbf{U}_j, Y_j)_{j=1}^n\right)\right]\exp(-C_7 n\epsilon^4)\,,
\end{aligned}
$$

where $C_1 = \pi_{\min}^2/(2^{11}L^2)$, and C_5, C_6, and C_7 are as above.

□

PROOF OF THEOREM 3

Standard arguments (such as those in the proof of Theorem 1 of Devroye and Györfi (1985, p. 254)) can be used to show that for every $\epsilon > 0$,

$$
\begin{aligned}
P\left\{L_n(\widehat{\Psi}_n) - L(\Psi_B) > \epsilon\right\} &\le P\left\{\sum_{i=1}^M \int \left|P_i(\mathbf{x}) - \widehat{P}_{i,n}(\mathbf{x})\right|\mu(d\mathbf{x}) > \varepsilon\right\} \\
&\le \sum_{i=1}^M P\left\{\int \left|P_i(\mathbf{x}) - \widehat{P}_{i,n}(\mathbf{x})\right|\mu(d\mathbf{x}) > \frac{\epsilon}{M}\right\}.
\end{aligned}
$$

Now, the conditions of the theorem together with the Borel-Cantelli lemma yield the strong consistency $L_n(\widehat{\Psi}_n) \to^{\text{a.s.}} L(\Psi_B)$.

□

PROOF OF THEOREM 4

The proof is similar to that for Theorem 3.

□

Conflict of interest statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] Bravo, F. (2015). Semiparametric estimation with missing covariates. Journal of Multivariate Analysis, 139, 329-346.

[2] Chen, H. Y. (2004). Nonparametric and semiparametric models for missing covariates in parametric regression. Journal of the American Statistical Association, 99, 1176-1189.

[3] Chen, Q., Paik, M.C., Kim, M., Wang, C. (2016). Using link-preserving imputation for logistic partially linear models with missing covariates. Computational Statistics & Data Analysis, 101, 174-185.

[4] Devroye L., Györfi L., and Lugosi G. (1996). A probabilistic theory of pattern recognition. Springer, New York.

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

[6] Devroye, L. and Györfi, L. (1983). Distribution-Free Exponential Bound on the L_1 Error of partitioning Estimates of a Regression Function. In Proceeding of the Fourth Pannonian Symposium on Mathematical Statistics, F. Konecny, J. Mogyorodi, W. Wertz, editors. Pages 67-76, Akademiai Kiado, Budapest, Hungary.

[7] Efromovich, S. (2012). Nonparametric regression with predictors missing at random. Journal of the American Statistical Association, 106, 306-319.

[8] Guo, X., Xu, W., Zhu, L. (2014). Multi-index regression models with missing covariates at random. Journal of Multivariate Analysis, 123, 345-363.

[9] Györfi, L. (1991). Universal consistencies of a regression estimate for unbounded regression functions. In *Nonparametric Functional Estimation and Related Topics,* Roussa, G., editor, pages 329-338. NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

[10] Györfi, L. Kohler, M. Krzyżak, A. Walk, H. (2002). A Distribution-Free Theory of Nonparametric Regression, Springer, New York.

[11] Horvitz, D .G and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, 663-685.

[12] Hu, Y., Zhu, Q., Tian, M. (2014). An efficient technique of multiple imputation in nonparametric quantile regression. Journal of Mathematics and Statistics, 10, 30-44.

[13] Lee, S.M., Li, C.S., Hsieh, S.H., and Huang, L.H. (2012). Semiparametric estimation of logistic regression model with missing covariates and outcome. Metrika 75, 621-653.

[14] Liang, H. , Wang, S. , Robins, J. , Carroll, R. (2004). Estimation in partially linear models with missing covariates. Journal of the American Statistical Association, 99, 357-367.

[15] Lipsitz, S. R. and Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. Biometrika, 83, 916-922.

[16] Little, R.J.A. and Rubin, D. B. (2002). Statistical Analysis with Missing Data, John Wiley, New Jersey.

[17] Liu, T. and Yuan, X. (2016). Weighted quantile regression with missing covariates using empirical likelihood. Statistics, 50, 89-113.

[18] Lukusa, T.M., Lee, S.M., and Li, C.S. (2016). Semiparametric estimation of a zero-inflated Poisson regression model with missing covariates. Metrika, 79, 457-483.

[19] Meier, L., van de Geer, S., and Bühlmann, P. (2009) High-dimensional additive modeling. Annals of Statistics, 37, 3779-3821.

[20] Mojirsheibani, M. (2012). Some results on classifier selection with missing covariates. Metrika, 75, 521-539.

[21] Pollard, D. (1984). Convergence of Stochastic Processes, Springer-Verlag, New York.

[22] Racine, J. and Hayfield, T. (2008). Nonparametric econometrics: the np package. J. Stat Softw. 27, 1-32.

[23] Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. J Econom. 119, 99-130.

[24] Robins, J. M. Rotnitzky, A. Zhao, L. (1994). Estimation of Regression Coefficients when some Regresoors are not Always Observed, Journal of the American Statistical Association, 89, 846-866.

[25] Sinha, S., Saha, K. K., Wang, S. (2014). Semiparametric approach for non-monotone missing covariates in a parametric regression model. Biometrics, 70, 299-311.

[26] Wu, K. and Wu, L. (2007). Generalized linear mixed models with informative dropouts and missing covariates. Metrika 66, 1-18.

Table 1: Empirical errors of histogram regression estimators under various missing probability mechanism: The results in rows 1, 2, and 3 correspond to logistic selection probabilities, those in rows 4, 5, and 6 correspond to trigonometric selection probabilities, and those in rows 7, 8, and 9 correspond to MCAR selection probabilities.

		Empirical L_1 errors			Empirical L_2 errors		
	Estimator	Model (I)	Model (II)	Model (III)	Model (I)	Model (II)	Model (III)
1	$\widetilde{m}_n(\mathbf{x})$	0.465	0.387	1.082	0.352	0.300	1.932
		(0.0063)	(0.0098)	(0.0219)	(0.0073)	(0.0120)	(0.0493)
2	$\widehat{m}_n(\mathbf{x})$	0.450	0.375	1.076	0.343	0.286	1.900
		(0.0071)	(0.0111)	(0.0244)	(0.0092)	(0.0129)	(0.0594)
3	$\widehat{m}_{\text{LS}}(\mathbf{x})$	0.384	0.274	0.978	0.295	0.221	1.793
		(0.0086)	(0.0119)	(0.0260)	(0.0128)	(0.0166)	(0.0685)
4	$\widetilde{m}_n(\mathbf{x})$	0.536	0.341	1.083	0.456	0.270	1.916
		(0.0078)	(0.0101)	(0.0203)	(0.0086)	(0.0142)	(0.0529)
5	$\widehat{m}_n(\mathbf{x})$	0.457	0.275	1.003	0.353	0.261	1.822
		(0.0093)	(0.0102)	(0.0226)	(0.0183)	(0.0146)	(0.0638)
6	$\widehat{m}_{\text{LS}}(\mathbf{x})$	0.552	0.374	1.113	0.467	0.290	1.968
		(0.0079)	(0.0097)	(0.0200)	(0.0097)	(0.0135)	(0.0504)
7	$\widetilde{m}_n(\mathbf{x})$	0.491	0.347	1.098	0.381	0.297	2.098
		(0.0072)	(0.0105)	(0.0211)	(0.0082)	(0.0121)	(0.0531)
8	$\widehat{m}_n(\mathbf{x})$	0.484	0.325	1.133	0.381	0.291	2.100
		(0.0079)	(0.0112)	(0.0217)	(0.0098)	(0.0127)	(0.0704)
9	$\widehat{m}_{\text{LS}}(\mathbf{x})$	0.493	0.339	1.137	0.396	0.293	2.105
		(0.0071)	(0.0098)	(0.0199)	(0.0084)	(0.0123)	(0.0532)

Table 2: Misclassification errors of various classifiers based on both resubstitution and leave-one-out cross-validation methods.

Method	$\widetilde{\Psi}_n(\mathbf{x})$	$\widehat{\Psi}_n(\mathbf{x})$	$\widehat{\Psi}_{\text{LS}}(\mathbf{x})$
Resubstitution	0.231	0.198	0.192
Cross-validation	0.245	0.207	0.198